Impact Assessment Framework: SME Finance

OCTOBER 2012

THE WORLD BANK

GPFI
Global Partnership
for Financial Inclusion

Impact Assessment Framework: SME Finance

October 2012

Prepared by the World Bank on behalf of the G20 Global Partnership for Financial Inclusion (GPFI) SME Finance Sub-Group.

Available online at: www.gpfi.org, www.worldbank.org/financialinclusion

Acknowledgments
This report was commissioned by the GPFI SME Finance Sub-Group, and prepared by Claudia Ruiz (lead author) and Inessa Love, under the guidance of Douglas Pearce.

The report benefited from input and review by: Giorgio Albareto (Banca d'Italia), Aidan Coville (World Bank), Elizabeth Davidson (World Bank), Susanne Dorasil (BMZ), Felipe Alexander Dunsch (World Bank), Aurora Ferrari (World Bank), Matthew Gamser (IFC), Randall Kempner (Aspen Network of Development Entrepreneurs), Leora Klapper (World Bank), Miriam Koreen (OECD), David McKenzie (World Bank), and Riccardo Settimo (Banca d'Italia).

The report was completed under the leadership of the co-chairs of the G20 SME Finance Sub-Group: Susanne Dorasil (Federal Ministry for Economic Cooperation and Development, Germany), Aysen Kulakoglu (Treasury of Turkey), and Jonathan Rose (U.S. Treasury).

1818 H Street NW
Washington DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

This work is a product of the staff of The World Bank with external contributions. The findings, interpretations, and conclusions expressed in this work do not necessarily reflect the views of The World Bank, its Board of Executive Directors, or the governments they represent.

The World Bank does not guarantee the accuracy of the data included in this work. The boundaries, colors, denominations, and other information shown on any map in this work do not imply any judgment on the part of The World Bank concerning the legal status of any territory or the endorsement or acceptance of such boundaries.

Table of Contents

Executive Summary

Small and medium enterprises (SMEs) are a policy priority for many countries, given their significance in terms of employment and economic activity. Many new policies, legal reforms, programs, and funds from both the public sector and donors focus on access to financial services and investment for SMEs. It is therefore important to assess and understand the impacts of these interventions to support SME finance so that they can be designed and implemented to most effectively meet their goals in a particular market or country.

Impact evaluation is an empirical assessment of whether a program or policy has achieved desired objectives. Impact evaluations help policy makers to quantify the effects of different policies, design the most effective interventions (that is, programs, policies, and regulations), improve targeting, refine policies to better fit objectives, optimize the scarce use of resources, and understand the underlying mechanisms. Tracking the impact of a policy, regulation, or program during its implementation (*real-time impact evaluation*) allows modifications to be made that can ensure the intended results are achieved.

Surprisingly, the cost of more rigorous impact evaluations is not much higher than the cost of minimal-standard monitoring. The most expensive part of both monitoring and assessing impact is collecting new data. If data are available, then the difference in cost between two methods is not substantial. For instance, in cases where administrative data can be used, the budget to design and implement an impact evaluation is significantly reduced.

This *Impact Assessment Framework* discusses the importance of rigorous impact evaluation and provides an overview of the relevance, application, strengths, and limitations of impact evaluation techniques. Relevant operational information regarding budget and timing issues is also present in the Framework.

The Framework covers experimental and non-experimental approaches that can be used to evaluate a broad context of SME policies and programs and provides examples of actual impact evaluations for each of the components of the *SME Finance Policy Guide* (GFPI 2011).

The *experimental* approaches discussed in this Framework include basic randomized control trials, oversubscription, randomized phase-in, and encouragement design. All of these approaches rely on a randomization device that allows the evaluator to isolate the impact of a policy:

- *Basic RCTs* refer to classic random assignments that take a baseline survey and randomly select some SMEs to receive the intervention. This approach can prove useful for interventions that are not implemented at the national level, such as local/regional interventions.
- In the *oversubscription design*, a subset of firms is randomly assigned to receive a program from the set of eligible firms that apply to it. This approach is useful to evaluate interventions in which the lack of funds necessitates limiting the number of firms that can participate in the intervention.
- The *randomized phase-in* approach randomizes the timing or sequence in which a project is rolled out. As its name suggests, this methodology is well suited to evaluate policies that are implemented at different stages.

- In the *encouragement* design, certain firms are randomly promoted (for instance, through financial incentives or marketing campaigns) to participate in the program, although the program is available to the rest of the population. This approach can be used to evaluate policies implemented at the national level and that were not rolled out differently.

The experimental approach allows for credible identification of the intervention impact and can be used to plan impact evaluations of different types in advance. However, experimental evaluations need to be set up before the policy or program is put in place, and their findings might not hold in different contexts (an issue commonly referred to as external validity).

The non-experimental methodologies covered in this Framework include difference-in-differences, instrumental variables, regression discontinuity, and propensity score matching. Unlike experiments, non-experimental evaluations do not include an exogenous device planned in advance to isolate the impact of a policy. Thus, these methods rely on identifying a control group and then using statistical techniques to ensure the impact estimate is properly measured. These approaches are commonly used to evaluate policies when an evaluation was not planned in advance.

- The *difference-in-difference* approach uses a comparable group of firms that was not exposed to the policy of interest as its control group. The approach then compares the outcomes over time of SMEs exposed to the policy relative to other firms from the control group. As long as a control group can be identified, this approach could be used to evaluate a variety of policies, including national-level interventions targeting SMEs and interventions at the regional level, among others.
- The *instrumental variables* approach relies on instruments to isolate the impact of a policy. Instruments are strong predictors of participation in the intervention but should not be associated with the outcome variable for reasons other than participation in the intervention. For instance, if a lending project took place in a municipality with a particular political party ruling, then the presence of this political party would strongly predict SME exposure to the lending project. But any change in SME outcomes should be due to the project itself and not through other channels associated to the political party in charge.
- *Regression discontinuity* is used to evaluate interventions in which a defined cutoff determines eligibility (such as policies provided to certain SMEs with less than a specific number of workers in the year before the intervention). By comparing the outcomes of firms that just passed the cutoff with firms that just missed the cutoff, evaluators can measure the intervention's effect.
- The *propensity score matching* (PSM) methodology can be used to evaluate an SME intervention in which the institutional arrangements that defined selection into the project are observed and known, and a control group is not maintained. A control group can be made up of firms not participating in the program, and the impact of the intervention determined by comparing the evolution of outcomes over time between the two sets of firms.

While the lack of a randomization device makes it more challenging for non-experimental methodologies to isolate the impact of an intervention, when done properly, these approaches provide robust estimates of the effect of interventions.

The Framework also discusses minimal standard monitoring, which consists of monitoring outcomes over time for the subjects receiving the intervention. The main difference with other impact evaluation methods is that a minimal standard monitoring does not follow a control group to identify the effect of a policy, which makes the results less rigorous and credible.

This Framework provides insights and criteria on the basis of which a suitable approach can be selected to evaluate an SME Finance policy, regulation, or program, including:

- Basic RCTs are well suited to evaluate SME interventions that have a clear distinction between those who participate in the program and those who do not (for example, public programs providing financial training to SMEs).
- Approaches that randomize the rollout of the implementation through randomized phase-in or encouragement design can be more suitable to evaluate interventions where the distinction of who participates is not clear, such as broad SME finance policies or regulatory reforms.
- To evaluate policies such as bank lending to SMEs, where institutions follow certain criteria to select eligible firms, both oversubscription and regression discontinuity might be suitable approaches. Oversubscription is particularly relevant when there are limited resources or implementation capacity and demand for a program or service exceed supply.
- Where the evaluation takes place after the policy has been already implemented, the evaluation approach is mainly determined by the characteristics of the intervention and how it was implemented. For instance, the difference-in-difference approach might be well suited to evaluate policies aimed at improving opportunities for female-led SMEs (since the evaluator can compare the evolution of female-led relative to male-led SMEs) or SME interventions that were rolled out sequentially across regions for political or logistical reasons, such as financial infrastructure projects.
- Alternatively, policies with a cutoff that determined who was eligible for the intervention are well suited for the regression discontinuity approach, such as a factoring project for SMEs employing fewer than 50 workers at the time of registration.

The Framework offers the following overall guidance:

- To isolate a policy's effect, it is important to conduct a rigorous impact evaluation instead of relying on before–after comparisons, which tend to generate flawed results.
- Impact evaluations planned ahead of the intervention offer more evaluation method options than evaluations conducted after the program or policy has been rolled out. Thus, it pays to plan evaluations before the intervention has started.
- There is no "one size fits all" approach to impact evaluation, and the most appropriate approach to evaluate an intervention will depend on the operational characteristics of the policy being evaluated.
- Rigorous impact evaluations can be complemented by qualitative assessments to provide a better understanding on the functioning, limitations, and strengths of the evaluated policy.
- Data collection is typically the most costly component of an evaluation. Evaluations that rely on existing data and ongoing or already-planned surveys can save on this cost component.
- Real-time impact evaluation during implementation allows modifications to be made to help ensure that the intended impacts are achieved. Rigorous impact evaluation assessments can improve the design, implementation, and impact of policies, regulations, and programs to support SME finance.

I. Introduction and Overview

SMEs play a key role in economic development and make an important contribution to employment. Financial access is critical for SME growth and development, and the availability of external finance is positively associated with productivity and growth. However, access to financial services remains a key constraint to SME growth and development, especially in emerging economies (GFPI 2011).

Policy makers and regulators have a wide menu of tools at their disposal to support increased access to financial services, as demonstrated in the comprehensive GPFI *SME Finance Policy Guide* (2011). Financial access for SMEs can be expanded by promoting a favorable legal and regulatory environment, complemented by a sound financial infrastructure and targeted public interventions. It is important to assess the impacts of various policies in order to prioritize, tailor, and sequence reforms to be most effective in addressing constraints to financial access in a particular market or country.

Impact evaluations assess whether a program or policy has achieved the desired objectives. These evaluations are usually systematic empirical studies, most often using actual data and statistical methods to measure outcomes and quantify the impact of the program or policy. Impact evaluations are a key ingredient for policy analysis and for understanding what works—that is, what are the most effective policies to achieve desired objectives, such as alleviating poverty, increasing access to finance, or enhancing growth and development in certain contexts. Thus, it is important to include impact evaluation in the design of policy and legal reforms and interventions.

This Framework was prepared as a resource for regulators and policy makers to provide an overview of methodologies used to evaluate the impact of various SME finance policies, interventions, and regulations. The Framework provides a comprehensive set of impact evaluation techniques; their key assumptions, strengths, and limitations; and examples of their implementation in SME finance policy contexts.[1] The techniques described in this Framework can be applied to *real-time* impact assessment that feeds back into policy implementation. Operational aspects of impact evaluation, such as budget and timing issues, are also discussed in the Framework. As detailed in the Framework, the impact evaluation approaches can then be selected to suit different policy contexts and priorities.

The first part of the Framework introduces the various impact evaluation approaches, discusses budget and time considerations for planning an evaluation, presents an outline of all necessary steps in the impact evaluation process, and maps evaluation approaches to different types of SME finance policies. The role of qualitative assessments as a complement of impact evaluation is also discussed. The second part of the Framework addresses in more detail the different methods. Section V describes the experimental approach. Section VI covers non-experimental methodologies, which range from difference-in-difference and instrumental variables to regression discontinuity and propensity score matching. Section VII describes the minimal standard monitoring, discusses its advantages and disadvantages, and contrasts this method to more rigorous impact evaluation techniques. Appendices 1 and 2 present technical considerations regarding

[1] The intention of the Framework is to provide an overview of impact evaluation methods and how they can be applied, rather than to present an exhaustive survey of all existing or ongoing evaluations.

estimation approaches. Appendix 3 compiles some examples of impact evaluations of SME finance interventions. Finally, Appendix 4 summarizes the key assumptions, strengths, and limitations of each evaluation approach examined in the Framework.

Several recent surveys on the topic of impact evaluation are relevant for this paper. McKenzie (2010) offers a survey of impact evaluations in a broader area of finance and private sector development. He makes a strong case for impact evaluations in the financial private development area. This paper complements his work, as it offers a systematic review of various evaluation methods relevant to SME finance, with pros and cons of each method and examples from their applications in SME finance policies. Gertler et al. (2011) offer a comprehensive impact evaluation guideline with detailed information on operational and technical issues. Bauchet et al. (2011) provide an excellent survey of randomized evaluations of microfinance. Winters, Salazar, and Maffioli (2010) provide a thorough survey of impact evaluations of agricultural projects. While the objective of this framework is also to review different impact evaluation approaches, our focus is on SME finance policies.

II. Why Are Impact Evaluations Relevant for SME Policies?

SME interventions can benefit from using impact evaluations in various ways. These evaluations can:

- Clarify the effect that interventions have on firms' outcomes and whether that impact achieved the expected objectives;
- Help to improve existing programs by comparing alternative design choices (for instance, comparing the performance of loan contracts with weekly versus monthly payments);
- Improve program targeting by identifying which firms benefit the most, or what barriers prevent others from gaining from interventions;
- Help prioritize resources by identifying the most cost-effective policies; and
- Make it possible to trace the different stages of an intervention so that evaluators are able to distinguish which key step in the program is not working as expected.

Unlike minimal-standard monitoring or simple before-and-after comparisons, impact evaluations isolate the effect of an intervention from all other factors that might alter the outcome of interest.

III. Menu of SME Finance Policies

While the impact evaluation methods presented in this Framework can serve to evaluate a broad set of SME and financial inclusion interventions, the Framework's main focus are SME Finance policies. More concretely, the SME finance reforms and interventions that the Framework covers are those examined in the GPFI SME Finance Policy Guide, which are classified in three groups: (1) regulatory and supervisory frameworks; (2) financial infrastructure; and (3) public interventions. Table 1 provides examples of these policies by type of intervention.

TABLE 1. EXAMPLES OF SME FINANCE POLICIES

INTERVENTION TYPE	EXAMPLES
1. Regulatory and supervisory frameworks	
Frameworks to promote competition	Regulations enabling entry of new banks Regulatory framework for licensing requirements
2. Financial infrastructure	
Insolvency regime	Bankruptcy reforms
Credit information systems	Introduction of credit bureaus, credit registries
Equity Investment	Reforms encouraging venture capital, angel funds
Accounting and auditing standards for SMEs	Reforms facilitating business registration procedures
3. Public interventions	
Public credit guarantee (PCG) schemes	Funds for guarantee to SMEs
Lending by state-owned financial institutions	Micro and SME finance programs
Apexes and other wholesale funding facilities	Direct lending in the form of grants
SME capacity, creditworthiness	Business/financial literacy training for entrepreneurs
Value-chain organization projects	Subsidies to promote technology transfer to SMEs

IV. Implementing an Impact Evaluation

The key challenge in evaluating the impact of any program is to ensure that observed outcomes are a direct result of the program itself and would not have occurred without the program. Without credibly addressing this, the impact evaluation may assign the outcome to the program, when in reality it could have occurred without it.

To clearly see the issue at stake, suppose a program affects some SMEs but not others. In essence, two questions must be addressed to resolve the issue: (1) How would SMEs who have participated in the program have done without the program? and (2) How would those who have not participated in the program have fared if they had participated? These questions are referred to as counterfactual because neither of these scenarios occurred in reality and thus are unobservable.

Observing the same SME over time will not, in most cases, give a reliable estimate of the impact the program had on it because many other things may have changed at the same time the program was introduced. The solution to this problem is to estimate the average impact of the program rather than the impact on each firm. One way to do that is to compare the average impact on the group that has participated in the program (also known as the "treatment" group) with an outcome for a similar group that has not (the "comparison" or "control" group).

The challenge is to ensure that this comparison group is identical to the treatment group in all ways, except for participating in the program. For example, to evaluate the impact of access to finance on SME productivity, it is not sufficient to compare those with a loan to those without one because SMEs that obtain a loan may be fundamentally different from those that do not. While controlling observables (such as size, age, and industry) may reduce these differences, some of the important differences are more difficult to observe—such as the entrepreneurial talent of the owners, their different risk preferences, or their social support network. Observed differences in performance between these two groups may be attributed to these latent (that is, unobservable) differences in characteristics rather than to access to finance.

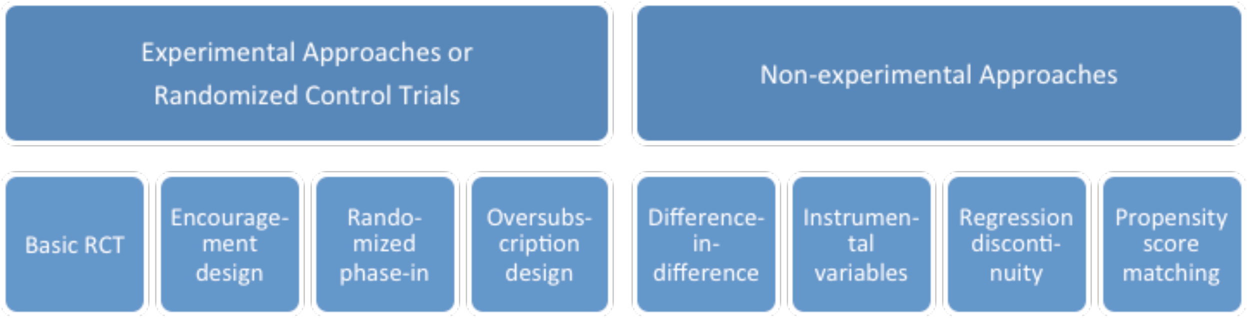
Impact evaluation techniques deal with these issues by identifying a proper counterfactual group to compare with the group of SMEs that were affected by the policy and, in this way, estimate as cleanly as possible the effect of the policy of interest. In general, impact evaluation approaches can be classified in two broad groups: experimental and non-experimental (see Figure 1).[2] Experimental methodologies randomly assign the intervention between the group that participates in the program or policy (the treatment group) and the group that does not (the control group) to ensure that any difference between these groups is attributed to the intervention. There are different ways to conduct randomized assignment. The most common randomized approaches that will be discussed in the Framework are described in Box 1. Non-experimental approaches—such as difference-in-difference, instrumental variables, regression discontinuity, and propensity score matching—identify a control group and then use statistical techniques to ensure that the impact estimate is properly measured. Sections V and VI of the Framework describe in detail each of these methods, their assumptions, and their advantages

[2] While minimal standard monitoring is not considered a rigorous impact evaluation method, it is a widely used approach to monitor the effect of policies on the targeted subjects. Section VII describes this method in more detail.

and disadvantages, providing examples of specific SME finance policies that were evaluated with them. Appendix 4 summarizes the main assumptions and characteristics of the evaluation methods discussed in this paper.

While discussing in detail qualitative assessments is beyond the focus of this Framework, it is worth mentioning that these types of analysis are an important complement to the findings reached through a rigorous impact evaluation. Qualitative assessments are commonly based on opinions of program participants and stakeholders about the policy, its success, and its limitations. Through surveys, interviews, focus groups, and/or case studies, qualitative evaluations collect additional information that sheds light on the satisfaction of participants, on the relevant mechanisms responsible for the impact of the intervention, and on general feedback to adjust and improve the operation of the policy or intervention. The OECD Framework for the evaluation of SME policies by Storey and Potter (2007) provides an in-depth review of these assessments.

FIGURE 1. EVALUATION APPROACHES


Experimental Approaches or Randomized Control Trials
Non-experimental Approaches
Basic RCT
Encourage-ment design
Rando-mized phase-in
Oversubs-cription design
Difference-in-difference
Instrumen-tal variables
Regression disconti-nuity
Propensity score matching

BOX 1. TYPES OF RANDOM ASSIGNMENT

Basic assignment. The classic model for random assignment is to take a baseline survey and randomly assign some participants to the project. This can be done on the level of individuals, firms, schools, or villages.

Oversubscription design. In this design, all eligible candidates are allowed to apply to the program, and a subset of all applicants is randomly assigned to receive the program (via a lottery system, for example). This design is useful when resources are limited and the demand for a program or service exceeds supply. This design can also be useful in randomizing among marginal loan applicants, as in Karlan and Zinman (2010).

Randomized phase-in. Because of the resource constraints, some units (individuals or geographic areas) subject to the program cannot receive the treatment at the same time. In such cases, randomizing who receives the program first is a fair way to allocate the resources and also allows for an impact evaluation of the program's effectiveness.

Encouragement design. In this design, some individuals or firms are randomly "encouraged" (via financial incentives or marketing materials) to participate in the program, even though the program is available to the rest of the population.

Operational Aspects of an Impact Evaluation

Budget Considerations

The overall cost of implementing an impact evaluation usually represents a small fraction of the total cost of the intervention. While the cost of an impact evaluation varies, it is possible to generate reasonable estimates up-front based on understanding the main cost drivers. These costs can be broadly categorized into "technical assistance" and "data collection," with data collection being the most important cost driver, generally constituting approximately 60 to 80 percent of the cost of an impact evaluation. For instance, while the average World Bank impact evaluation costs $500k to $900k (Gertler et al. 2011) when data collection is required, the cost declines to $50k to $200k when administrative data can be used.

Administrative data consists of information collected for some official purpose, such as reporting to government agencies or maintaining records of program participants. While this data is not designed to perform evaluations, if the available indicators fit with the objectives of the evaluation, administrative data is a valid option to consider.[3]

Data Collection Costs

It is difficult to determine the costs of data collection precisely since these will depend on different variables such as the sample size needed, the type of data to be collected (household, individual, administrative), the length of the surveys, the frequency in which the data will be gathered, and the labor costs of each country. Yet the Alliance for Financial Inclusion (AFI) provides estimated survey costs for different types of surveys with different sample sizes. According to AFI, a nationally representative cross-sectional survey could range from $100,000 to $700,000 depending on the sample size (from 1,000 to 7,000 observations) and the country where the survey was conducted. Information from the Living Standards Measurement Study of the World Bank indicates that survey costs range from $150 to $300 per household, with a usual sample size between 3,000 to 5,000 households.

The possibility of using administrative data can thus greatly reduce impact evaluation costs, but assessing the availability of *relevant* data is critical and needs to consider the following factors:

1. Impact evaluations require data before and after the intervention in both control and treatment groups. Administrative data would need to be available over these time periods for these population groups.

2. The more time points available, the more accurate the results: data available at regular intervals for the indicators of interest improve precision, preferably available before, during, and after the intervention.

3. Access and confidentiality can be challenging: While administrative data may exist, accessing these data may be difficult for security reasons. In addition, the time required to access the data in a workable format needs to be factored into the process.

[3] For example, the Italian tax authority conducts a "Sector Studies" survey to collect information on SMEs activities, economic outcomes, and other variables with the objective of computing how much SMEs pay in taxes. These administrative data have been used in different evaluation projects. In Chile, the Suppliers Development Program, which seeks to strengthen the commercial linkages between small- and medium-sized local suppliers and their large firm customers, keeps records of all participating firms. These records were used in an evaluation to understand the effect of the program on SME productivity (Arraiz, Henriquez, and Stucchi 2011).

4. Available indicators dictate the questions that can be asked: The types of outcomes that can be monitored are restricted to the types of indicators collected in the administrative data. Evaluators must make sure that the available indicators allow them to monitor the main outcomes of interest for the evaluation.

5. Data format and quality: Administrative data are usually collected for purposes other than statistical analysis. As such, data may not necessarily be in a format that can be directly analyzed, requiring effort to clean and reframe for analysis purposes. The quality of these data also needs to be scrutinized if the evaluation team has not been involved in its collection.

In many cases, administrative data are not available in exactly the right format needed for the impact evaluation for any of the reasons described above. However, it is often possible to work with the office responsible for collecting the administrative data to adapt the data collection activities (for example, by adding specific questions to the larger survey or by including control group data collection). Ex ante evaluations allow the administrative data to be adapted to suit the needs of the evaluation, which is not possible when relying on historical data for an ex post evaluation.

Impact evaluation methods will only affect costs in as much as they influence the data requirements. For instance, approaches such as propensity score matching or regression discontinuity require information on a large set of subjects. Non-experimental methods such as difference-in-difference also require baseline data to ensure that the control and treatment groups are comparable. Though more limited in its precision, an RCT is the only method that does not specifically require a baseline to be conducted since, by definition, the control and treatment groups will be comparable. However, it is generally good practice to collect baseline data for any evaluation method used.

Additional costs of impact evaluations not necessarily associated with minimal-standard monitoring include monitoring costs of the evaluation (if planned ahead) and researchers' time, but these are usually a small part of the overall budget. In addition, minimal-standard monitoring does not need to collect data on the control groups. It does, however, need data on the periods before the intervention started and after it was rolled out.

Technical Assistance Costs

In addition to data collection costs, impact evaluation work requires budgeting staff time, travel arrangements, and dissemination costs. Contributions from researchers are not needed throughout the whole project timeframe; they mainly contribute work for the impact evaluation design and sampling, as well as data cleaning and analysis. However, most impact evaluations that include data collection need a constant presence in the field, such as a field coordinator, to monitor data collection efforts. Still, these costs are usually a smaller part of the overall budget compared to data collection.

In summary, the main budget items of an impact evaluation are:

- *Data collection.* The team should identify all primary and secondary data collection requirements and provide a budget for completion (minimum baseline and follow-up data), including qualitative and/or cost-analysis data collection requirements where applicable.
- *Impact evaluation team.* The budget should include all staff and consultant time for managing the impact evaluation, including design, implementation, and analysis.
- *Travel.* All necessary travel costs for required project supervision must be factored in, including airfare, accommodations, and food.
- *Specialists.* The budget should include any additional consultant time and travel for technical assistance (such as survey instrument development, data quality control, and data entry program development).

- *Dissemination plan.* Any costs associated with travel or logistics must be taken into account for at least one field-based presentation at baseline and one at follow-up, as well as any costs associated with producing written materials.

- *Miscellaneous.* The budget should include any additional costs related to the impact evaluation, such as payments for institutional review of the research protocol.

Time Considerations

Ideally, impact evaluations should be planned prior to the rollout of the program. Doing so allows the team to collect meaningful pre-intervention baseline data and organize the project implementation for an eventual RCT (allocation of treatment and control group) and helps stakeholders to reach consensus on the program objectives.

To identify the impact of any intervention, evaluators then need to allow sufficient time for the impacts to manifest. Both short-term and long-term impacts can be considered, depending on the intervention, objectives, and the theory of change backing the project design. The following factors need to be weighted to determine when to collect follow-up data (Gertler et al. 2011):

- Program cycle (including program duration), time of implementation, and potential delays.

- Expected time needed for the program to affect outcomes, as well as the nature of outcomes of interest.

- Policymaking cycles.

Often, performing an evaluation too soon after the intervention may miss the important long-term consequences. Also, the evaluation timeline must adapt to the timeline of the project rather than to the evaluation driving the timeline of the project. Evaluators therefore need to be flexible regarding the timing. A strong monitoring system can help track the progress of the actual implementation.

When sufficient budget is available, it is advised to conduct multiple surveys (midline and endline), which allow the evaluators to draw short-term and long-term conclusions. In addition, tracking the progress of the intervention with a midline survey may help to realign the program to improve the overall project outcomes. Follow-up surveys that measure long-term impacts after the program implementation often produce the most convincing evidence regarding program effectiveness.

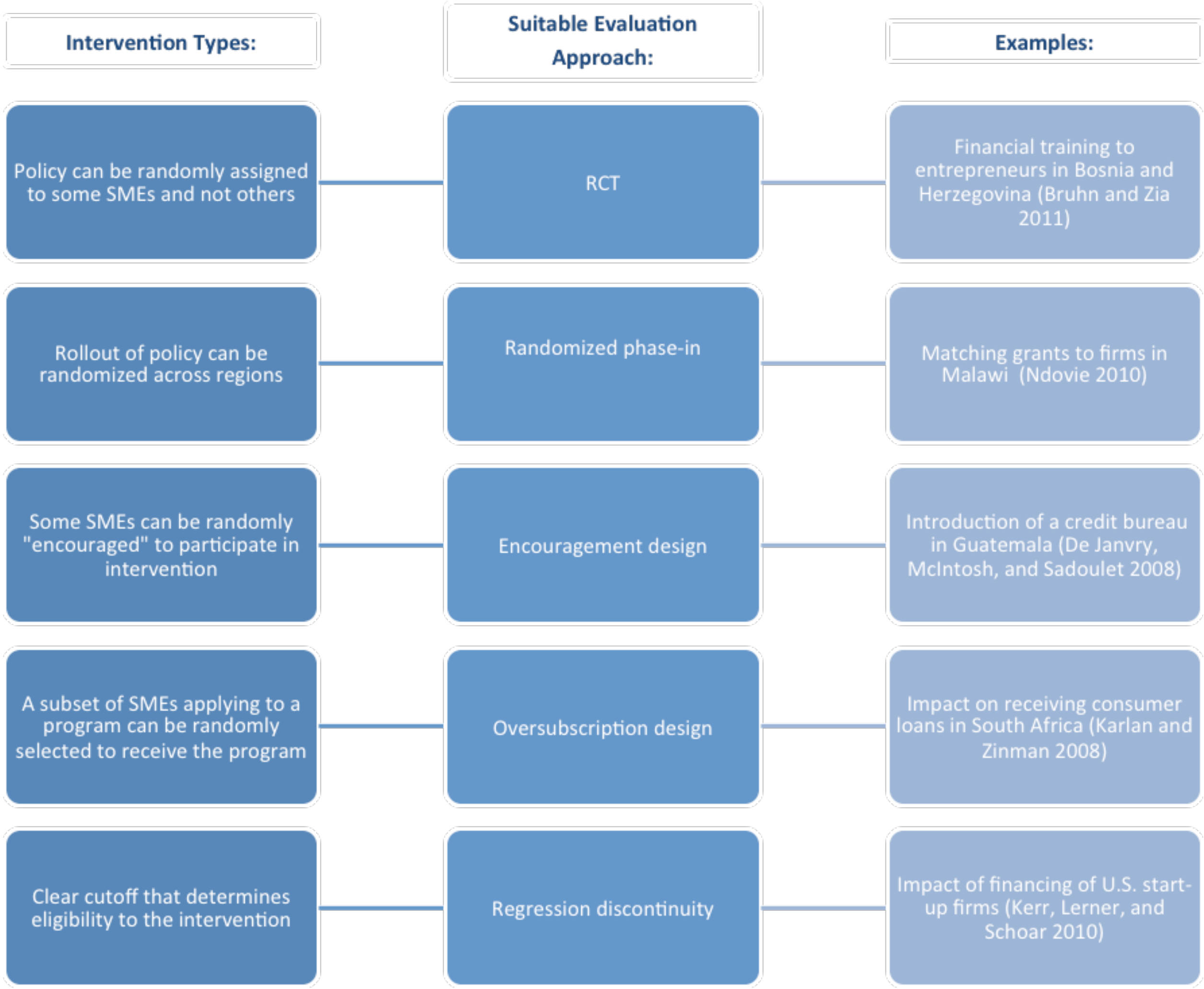
The timing of an evaluation must also account for when certain information is needed to inform decision making and must synchronize evaluation and data collection activities to key decision-making points. The production of results should be timed to inform budgets, program expansion, or other policy decisions.

Selecting an Impact Evaluation Method for an SME Finance Policy

The operational characteristics of the policy should guide the selection of the impact evaluation method. More concretely, there are two important components of the policy that matter when selecting an evaluation approach: i) who is eligible to the program and ii) how eligible subjects are selected to participate or receive the program. There is no “one size fits all” impact evaluation approach, and the best approach will differ with the situation and the policy’s characteristics. An additional factor to consider is whether the evaluation was planned ex ante (before the program has started) or is occurring ex post (during or after the program began).

SME finance impact evaluations that are planned in advance offer more options for evaluation methods than those conducted after the program or policy has been rolled out. Planning ahead has several advantages. For instance, the evaluator can carry out baseline analysis to establish appropriate comparison groups. Evaluators can also decide whether they need to collect specific data not covered in other sources. Under some circumstances, evaluators can introduce a randomization device to increase comparability of control and treatment groups and thus strengthen the evaluation results.

FIGURE 2. SUGGESTED DESIGNS FOR EVALUATIONS PLANNED AHEAD



Intervention Types:	Suitable Evaluation Approach:	Examples:
Policy can be randomly assigned to some SMEs and not others	RCT	Financial training to entrepreneurs in Bosnia and Herzegovina (Bruhn and Zia 2011)
Rollout of policy can be randomized across regions	Randomized phase-in	Matching grants to firms in Malawi (Ndovie 2010)
Some SMEs can be randomly "encouraged" to participate in intervention	Encouragement design	Introduction of a credit bureau in Guatemala (De Janvry, McIntosh, and Sadoulet 2008)
A subset of SMEs applying to a program can be randomly selected to receive the program	Oversubscription design	Impact on receiving consumer loans in South Africa (Karlan and Zinman 2008)
Clear cutoff that determines eligibility to the intervention	Regression discontinuity	Impact of financing of U.S. start-up firms (Kerr, Lerner, and Schoar 2010)

Planned evaluations can be used even in interventions in which no obvious control group was followed. For instance, national interventions that were implemented at the same time everywhere can still be evaluated with a rigorous method. Think of a nationwide intervention in which firms apply to participate in a program. Evaluators might plan ahead for an *encouragement device* to evaluate the intervention (such as reducing the cost of applying for randomly selected firms). Now think of this same intervention but with the additional constraint of limited fund availability, reducing the number of firms the program can accommodate. Evaluators can use an *oversubscription design* in which firms from the pool of applicants are randomly assigned to the program while the others are not.

Other very common interventions are those that take place simultaneously and at the national level. Evaluators can still find methods to evaluate the impact of these types of interventions. Think, for instance, of interventions trying to reduce the regulatory costs that SMEs face. We might expect these interventions to have a substantially higher effect on SMEs than on larger firms. If this is the case, then evaluators can plan ahead a *difference-in-difference* evaluation by comparing the performance of SMEs before and after the intervention with that of larger firms.

Finally, evaluators might be creative and utilize the lack of information among SMEs about new nationwide interventions. Let us suppose that a regulation to facilitate the requirements to open a business was implemented but not marketed to the public. Evaluators might then plan an encouragement design evaluation in which they randomly provide detailed information on the new regulation only to a subset of firms.

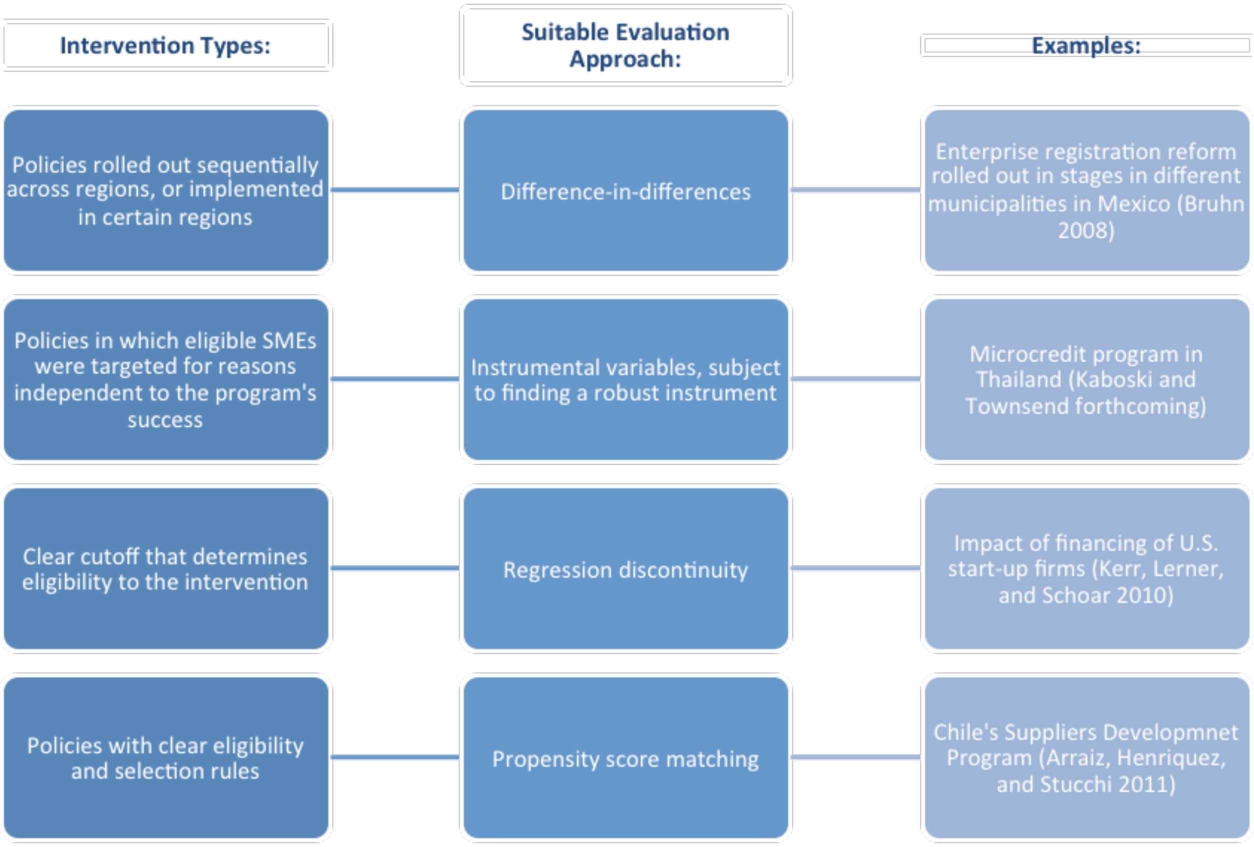
Figure 2 presents a method for selecting the most appropriate evaluation approach when the evaluation is planned ahead of the program implementation. While there is no unique mapping of evaluation approaches for interventions, in general, interventions that clearly distinguish participants from non-participants are good candidates for RCTs. Several public interventions might fall into this category, such as programs providing training or grants to SMEs. In other interventions, such as regulatory reforms, who receives the benefits and who does not might not be as clear. These types of interventions might be more suitable to evaluate approaches that randomize the rollout of the implementation sequentially throughout regions or that randomly provide an incentive to some groups to participate in the program.

Figure 3 presents a method to help evaluators select an approach for evaluations that were not planned before the intervention. If, for instance, a credit bureau was established in different regions over time, a difference-in-difference approach can evaluate its impact by comparing the outcomes over time on regions where the credit bureau started (the treatment group) with comparable regions where the credit bureau was not yet implemented (the control group).

Sections V and VI discuss in more detail the main features of each of the impact evaluation methods, providing examples of interventions evaluated using each approach and discussing their main assumptions, advantages, and disadvantages.

Appendix 3 discusses several examples of impact evaluations performed for various SME finance policies.

FIGURE 3. SUGGESTED DESIGNS FOR EVALUATIONS NOT PLANNED AHEAD



Intervention Types:	Suitable Evaluation Approach:	Examples:
Policies rolled out sequentially across regions, or implemented in certain regions	Difference-in-differences	Enterprise registration reform rolled out in stages in different municipalities in Mexico (Bruhn 2008)
Policies in which eligible SMEs were targeted for reasons independent to the program's success	Instrumental variables, subject to finding a robust instrument	Microcredit program in Thailand (Kaboski and Townsend forthcoming)
Clear cutoff that determines eligibility to the intervention	Regression discontinuity	Impact of financing of U.S. start-up firms (Kerr, Lerner, and Schoar 2010)
Policies with clear eligibility and selection rules	Propensity score matching	Chile's Suppliers Developmnet Program (Arraiz, Henriquez, and Stucchi 2011)

Steps in the Impact Evaluation Process

This section summarizes the recommended steps that an impact evaluation should follow.[4] We classified the main steps into four groups: pre-evaluation assessment, evaluation design, data collection, and analysis of results.

During the pre-evaluation assessment, the team must have a clear understanding of the intervention that will be evaluated. It is important to know its main operational characteristics, such as eligibility criteria for the program and how the eligible SMEs are selected for participation. *This information is crucial since these characteristics will be the main factors influencing the selection of the proper impact evaluation method.*

At this stage, the team should also identify the objectives for which the policy was designed. Was the policy intended to increase employment of SME workers? Was it planned to raise productivity of rural SMEs? Having clearly defined policy objectives will guide the evaluation team to decide which indicators to monitor throughout the evaluation. For instance, if the policy was intended to increase employment of SME workers, then a natural indicator to evaluate is the number of jobs. Evaluators should identify which indicators they plan to use, keeping in mind the data available to perform the evaluation.

During the evaluation design stage, the team must review if the indicators to monitor can be retrieved from data already available or if new data collection is needed. Since collecting data is the most expensive part of an impact evaluation, an effective way to maintain a tight budget is by using preexisting data whenever possible. Based on the intervention's characteristics and the type of data to be used, evaluators must decide on the most suitable impact evaluation approach (mainly, identify which subjects will constitute the treated and control groups). In the next section, we provide some guidelines on how to select the appropriate method.

Data collection is the third step, and it will apply in cases in which evaluators plan to collect new data. This includes the entire process from survey design, to piloting the questionnaires, conducting fieldwork, and validating the data.

TABLE 2. STEPS IN THE IMPACT EVALUATION PROCESS

I. Pre-evaluation assessment	■ Have a clear understanding of the characteristics of the intervention ■ Identify objectives of the intervention ■ Identify the outcomes/indicators to evaluate
II. Evaluation design	■ Review data available to perform evaluation and determine whether new data is needed ■ Select an impact evaluation method
III. Data collection (if needed)	■ Design survey ■ Pilot questionnaires ■ Conduct fieldwork ■ Process and validate data
IV. Analysis of results	■ Produce findings of the evaluation

[4] Gertler et al. (2011) provide an in-depth description of a roadmap for impact evaluations.

In the final stage, evaluators analyze the outcomes in the treatment and the control groups and produce the results. At this stage, the evaluators can determine the impacts of the intervention and present them to the appropriate policy makers.

Table 2 outlines the main activities to follow at each step of the impact evaluation process.

V. Impact Evaluation Methods—The Experimental Approach

In recent years, randomized experiments, also known as randomized control trials (RCTs), have increasingly become the preferred method of evaluation for many development economists (Duflo and Kremer 2006). The essence of the RCT design lies in randomly assigning some units (individuals or firms) to receive the "treatment" (that is, participation in the program) and others to serve as a control group. Such random assignment allows for a credible attribution of the outcomes observed to the program investigated.

The key reason the RCT methodology has gained so much popularity lies in its ability to address the *identification problem*—ensuring the outcome of the program or policy would not have occurred in such program or policy's absence.

Box 2 describes an RCT that evaluated a business training program targeted at entrepreneurs in Bosnia and Herzegovina. For a discussion of several prominent examples of RCT evaluations relevant for SME finance policies, see Appendix 3.

Key Assumptions

The key assumption of the RCT evaluation is the random assignment of subjects (such as SMEs) to

BOX 2. PUBLIC SECTOR INTERVENTION EVALUATION: BUSINESS TRAINING IN BOSNIA AND HERZEGOVINA

While access to finance has long been thought of as a constraint on SME growth, another set of constraints has recently emerged—business skills, or "managerial capital," which is thought to be lacking in many entrepreneurs. Thus, business training programs and managerial education have become an important focus for policy makers. Business training programs are a good example of interventions that can be evaluated using RCT because they can be randomly administered to a subset of the SMEs to create a clear control group.

A randomized evaluation of a comprehensive business and financial literacy training program for entrepreneurs ages 18 to 30 was conducted in Bosnia and Herzegovina (Bruhn and Zia 2011). The sample included small businesses with an average of two employees. The course covered basic business concepts and accounting skills, as well as investment and growth strategies, with a particular emphasis on the importance of up-front capital investment. The researchers randomly selected treatment and control groups, and performed baseline surveys in both groups. Similar to many other RCT studies, this study had a relatively low take-up rate: only 39 percent of those in the treatment group actually attended the business training course; others cited lack of time as the reason for nonattendance.

The authors found that the training program led to better business practices, such as separation of business and personal accounts and more favorable loan terms, greater investment, and some improvements in sales and profits (but only among a subsample of entrepreneurs with higher financial literacy). However, the program had no effect on firm survival or business start-up, or on loan default rates.

The type of information generated by such studies would enable policy makers to design effective financial literacy training programs and target the subsets of SMEs for which such training programs would be the most effective.

participate in the program. While such assignment is random by design, it must be assumed that SMEs cannot manipulate the program assignment (for example, by moving into or out of the affected areas). In addition, all those assigned to the control group must be credibly excluded from receiving any benefits from the intervention.

Strengths

Clear Comparison Group

The random assignment to participate in the program by design creates a valid comparison group since individuals or firms are randomly placed in the treatment group or the control group. Hence, placement does not depend on any preexisting characteristics that may influence the outcome of the program. In this case, one can be reasonably well assured that program participation is the only reason different average outcomes are observed in the two groups. In other words, when a randomized evaluation is correctly designed and implemented, it provides an unbiased estimate of the impact of the program in the study sample.

Baseline Data Not Necessary

RCT evaluations can be performed without detailed baseline data, which can save on the costs of data collection. Nevertheless, baseline data are often helpful to verify the assignment and also study how impacts differ for different subsamples, such as men and women.

Limitations

Not All Policies Are Suitable for RCT

For an RCT to work, there must be a clear distinction between the treatment and control groups. The best candidates for RCTs are programs that are targeted to individuals, firms, or local communities. For example, Ravallion (2009) argues that randomization is not suitable for a large subset of policies important for development economics because most often these policies apply to the whole country, the whole population, or all firms. Investigating such a policy using RCT is unlikely to be feasible because no group can be randomly selected not to receive the "treatment." Examples of such policies within the SME finance framework include most policies affecting legal, regulatory, and supervisory frameworks, as those policies most often are implemented on an economy-wide scale. However, such policies can often be evaluated using encouragement design, a type of RCT (see Box 3), or nonrandomized methods.

Sometimes policies that are intended to affect the whole economy may be designed to allow for randomization or for ex post program evaluation if the rollout happens in stages. For example, an

BOX 3. ENCOURAGEMENT DESIGN

Encouragement design is likely to be applicable for a wide variety of evaluations of SME-related policies. This method can be very useful for evaluating policies and interventions that are implemented at the country level, such as most changes in regulatory and supervisory frameworks. Such policies can be evaluated in a semi-randomized fashion. In this method, some units (such as firms or households) selected at random receive incentives to participate in a program that is available to all. Such encouragement can be in the form of information, marketing materials, or financial stimulus.

An example of an encouragement design mechanism can consist of reducing the cost of applications for a random subset of SMEs to a guarantee program. If firms receiving the encouragement are more likely to apply to the program, this mechanism will predict program participation. Moreover, as this program is assigned randomly, it will not be correlated with firms' access to credit, so the incentives can be used to evaluate the impact of the intervention.

enterprise registration reform was rolled out in stages in different municipalities in Mexico (Bruhn 2008). While the sequence of these events can be credibly seen as exogenous to the outcomes of interest, it was not done randomly. Nevertheless, Duflo and Kremer (2005) argue that randomly determining the order of phase-in may be a fair way to introduce a program and also will allow for RCT evaluation.[5]

An important limitation of RCT is that it cannot be used to randomly select the recipients of a loan, as financial institutions need to ensure that their recipients are creditworthy and that the loans will be repaid. Thus, the allocation of credit should not alter the risk-assessment process of the bank because it could undermine the viability of the SME finance program. An example of a design that takes this issue into account is Karlan and Zinman (2008). In their study, consumers first applied for loans, and then the pool of marginally rejected candidates was randomly assigned to receive a loan. Such studies may also help banks better refine their credit-scoring methodologies.

Another common issue with evaluating programs using randomized methods is that some individuals or firms must be restricted from access to the program. There may be political opposition to delaying program access to some people or firms, or there may be ethical considerations.

Finally, for an RCT evaluation to be feasible, evaluators need to obtain data on a sufficient number of treated versus untreated "units." If the units are individuals or firms, it is most likely that sufficient numbers can be found for a statistically valid comparison. But if, for example, the unit of analysis is financial institutions in a highly concentrated financial sector, then there might not be enough of them to compare one group to the other.

Power of the Design

One important issue with experimental design is the power to detect the program effect. The power of the design is the probability that a statistically significant result will be obtained. In other words, the power is the assurance that the result observed is unlikely due to pure chance. One way to address the issue of power is to ensure a sufficiently large sample size. Appendix 2 offers more details on the issues of power and take-up.

Take-up

Related to the problem of power is the take-up of the program, or the proportion of those affected by a policy or a program—whether individuals, households, or SMEs—that will actually use the program. Any program or intervention's impact will significantly depend on the take-up. For example, not all enterprises will chose to register formally or to obtain a loan even if they are assigned to the "treatment" group that offers a particular intervention. A program that increases the availability of finance may not have the desired impact if SMEs do not actually need more access (but perhaps suffer from high costs of access).

The first challenge with low take-up is that it increases the sample size needed to generate statistically significant differences. The second challenge is one of interpreting program impact (see Appendix 1 for a discussion of technical details), which means that program effects must be carefully interpreted to decide whether the parameter estimated is, in fact, one of policy interest.

[5] However, randomized phase-in may become problematic when the comparison group is affected by the expectation of future treatment. For example, in the case of a phased-in microcredit program, individuals in the comparison groups may delay investing in anticipation of cheaper credit once they have access to the program. In this case, the comparison group does not provide a valid counterfactual.

VI. Impact Evaluation Methods—Non-experimental Approaches

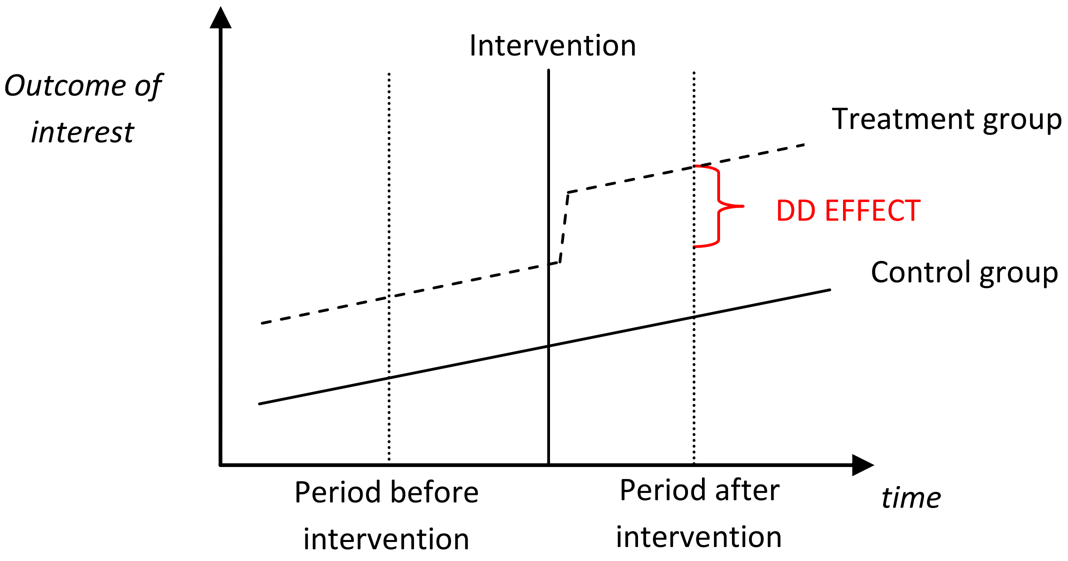
Difference-in-Difference

The difference-in-difference (DD) approach is one of the most popular methodologies used in impact evaluation, including assessments of SME finance policies. This methodology compares outcomes, before and after an intervention took place and between the group that received the intervention (treated group) and a control group. The function of the control group is to take into account changes over time that might also affect the treatment group's outcomes. Thus, by comparing the outcomes of the control group to the outcomes of the treated group, any factors affecting both groups in the same manner are canceled out. As with RCTs, the control group is used to infer what would have happened to the treated group if the intervention had not taken place.

To evaluate an intervention using DD, data on the outcomes of interest for the treatment and control groups are needed from periods before and after the intervention. Figure 4 illustrates the DD effect.[6]

The DD approach is well suited to evaluate SME interventions in which the implementation of the program took place at different stages (for example, a program that was rolled out across municipalities over time) or in which the implementation was targeted to some groups and not others (for example, a project targeting particular municipalities). The evaluator must understand the reasons for targeting specific groups, and whether the treatment group was selected to maximize the performance of the intervention, then DD estimates could produce biased results (see Box 4 for a DD impact evaluation example).

FIGURE 4. THE DD EFFECT



[6] The DD effect is computed through two subtractions. First, changes in the outcomes from periods before and after the policy was implemented are computed separately for both groups. Then, to net out any aggregate trend confounding the impact of the intervention, the gains of the treated group are subtracted from the outcomes' changes of the control group.

BOX 4. REGULATORY REFORM EVALUATION: BUSINESS REGISTRATION IN MEXICO

In 2002, the Mexican Federal Commission for Improving Regulation (COFEMER) implemented a new system that substantially reduced the number of procedures and days required to register a business. The objective of this system was to simplify business registration procedures in Mexico. Due to staff constraints, the system could not be implemented in all municipalities at the same time. While the system was launched in some municipalities in 2002, others were still in the process of setting it up in 2006. Interestingly, the timing of the implementation across municipalities had no particular pattern.

Bruhn (2011) used this exogenous variation on the timing of the implementation across municipalities to evaluate the impact of this new business registration reform on economic outputs. Using a difference-in-difference approach, she classified the municipalities that set up the system early as the treatment group. The control group consisted of municipalities with similar characteristics to those in the treatment group but where the system had not yet been implemented. As long as the changes in the economic outcomes over time would have been similar in the absence of the reform, this approach to examine the impact of this reform is valid.

To make sure this was the case, Bruhn examined whether the control municipalities could be used as a proper counterfactual by first establishing that these municipalities were comparable to the treated ones. Using data from periods before the reform, she showed that there were no statistically significant differences in the output data, which diminished concerns of selection bias issues between control and treatment municipalities. She also verified that both early and late adopters were geographically dispersed throughout Mexico, reducing the contagion issue by which firms from control municipalities could be benefiting from the reform.

Her findings suggest that the reform increased the number of registered businesses by 5 percent and employment in these industries by 2.8 percent. By increasing competition, the reform benefited consumers and hurt incumbent businesses: after the reform, the price level fell by 0.6 percent and the income of incumbent registered businesses declined by 3.2 percent.

Key Assumptions

The fundamental assumption of the DD estimator is that the control group trend is identical to the trend that the treated group would have had in the absence of treatment. While this assumption is not testable, its validity should always be carefully examined to ensure that the DD properly estimates the impact of the program. If data are available for several years preceding the treatment, then one straightforward way to assess the validity of this assumption is to analyze whether pretreatment trends were equal between groups. While this does not formally prove the identification assumption (which, as mentioned, is not testable), the equality of pretreatment trends suggests that the treated and control groups are, indeed, comparable and thus reinforces the credibility of the estimates.

However, this assumption might be violated when evaluating interventions in which firms self-select into the program. Take, for instance, an evaluation of a new state bank providing loans to SMEs in which firms have to apply for the loan. Using as a control group those firms that decided not to apply for the loan and as treatment those firms that did apply will very likely produce biased results. Firms that select into SME interventions do so because they expect some gains from their participation, while firms that decide not to participate are likely to expect no substantial gains from it. In this case, the control group is not a good representation of the treatment group. In contrast, if the state bank entered in some municipalities and not in others for logistical or political concerns, then a more robust comparison would be to use SMEs from municipalities without the state bank as the control group and SMEs from municipalities where the bank entered as the treatment group.

Strengths

DD Controls for Factors that Do Not Vary over Time

One benefit of this approach is that DD estimates control for all differences (observable and not) between control and treated groups that do not change over time, minimizing potential biases in impact estimates.

Limitations

The Key Assumption Is Not Testable

One of the main issues of this methodology is that its underlying assumption (of equal trends that the treatment and control groups would have had without the intervention) is not testable, and if it fails to hold, then the DD impact will be biased.

Targeted Interventions

The estimates could be biased if the intervention targeted groups that are expected to experience higher gains. For instance, if a microcredit intervention was implemented in villages with inherently high demand for credit, then the effect that the program would have in treated villages is potentially different from the effect that it would have in the control group, since the demand for loans in this group is lower. Therefore, it is important to understand the motives behind the intervention's implementation and the choice of the treated group.

Other Changes that Affect One Group and Not the Other

Another issue to consider is that this approach will fail to identify the impact of a policy if any change other than the intervention occurs over time affecting one group and not the other. When using DDs, one must be confident that such changes did not occur.

Instrumental Variables

The instrumental variables (IV) approach can be used to evaluate SME interventions in which firms, based on unobserved information, can select whether to participate in the program. Very often entrepreneurs self-select themselves to participate in SME finance projects. For example, an intervention providing public credit guarantees with the objective of increasing firms' access to credit may require that entrepreneurs apply for the guarantee. Firms that expect to benefit from having a public guarantee will apply, while firms that expect little or no benefit from the program will not.

To evaluate interventions of this type, an instrument or set of instruments is required. A valid instrument must be a strong predictor of participation in the intervention and must not be correlated with the outcome variable for reasons other than participation in the intervention (that is, it must be exogenous). In this example, an instrument must predict firms' choice to participate in the public guarantee program but must not influence firms' access to credit for reasons other than participation in the guarantee program.

Once an instrument is identified, the impact of an intervention is computed in two steps. In the first step, the instrument is used to predict program participation. In the second step, the predicted participation (which is independent of the outcome variable) is used to evaluate the intervention's impact.

Box 5 discusses an example of an IV impact evaluation that analyzed the effect of a microcredit program in Thailand.

Key Assumptions

The IV estimates are valid if the instrument:

- *Is a strong predictor of participation to the intervention.* In the microfinance evaluation example, the evaluators were interested in understanding the impact of credit on economic outcomes of Thai villages. Since the amount of credit injected in all villages through the program was the same, smaller villages ended up receiving a more intense credit injection than larger ones. The evaluators' instrument (interactions between the number of households in a village and the program years) is a good predictor of the intensity of credit received in each village because the number of households determined the intensity of the credit injection.

- *Is not correlated with the outcomes evaluated.* In the example above, the instrument used for the evaluation (number of households in each village during the program years) must not influence the consumption of Thai households, their investments, and overall asset and income growth except through the effect of the program.[7]

If these assumptions do not hold, the impact estimates will be biased.

Strengths

IV Controls for Unobserved Information

One benefit of the IV approach is that it controls for unobserved differences between participating and nonparticipating subjects. IV estimates isolate the effect of the intervention from unobserved information that influences self-selection into the program.

Baseline Data Are Not Needed

To estimate the IV impact, baseline data are not needed.

Limitations

Unplanned IV Evaluations Are Rare

Evaluations of an intervention in which an IV design was not planned ex ante are rare because finding a valid exogenous instrument that predicts participation is extremely challenging.

BOX 5. PUBLIC INTERVENTION EVALUATION: THAILAND MICROFINANCE FUND

During 2001 and 2002, a substantial microfinance initiative was implemented in Thailand: Thailand's Million Baht Village Fund Program. This public intervention consisted of injecting funds into all 77,000 Thai villages. The initial funds distributed were significant, corresponding to about 1.5 percent of Thai GDP in 2001. Each transfer was used to form an independent village bank for lending within the village. Importantly, every village, regardless of its characteristics, was eligible to receive the program. This program is among the largest government microfinance initiatives of its kind.

Kaboski and Townsend (forthcoming) evaluated the impact that Thailand's Million Baht Village Fund Program had on economic outputs of Thai villages using the IV approach. As each village received the same amount of money, regardless of the population of the village, smaller villages received a relatively more intense injection of credit. Due to the nature of the intervention, the expansion of credit in villages by the Thai Fund Program could be correlated with the number of households in a village during the program years. Using these interactions of number of households and the program years as instruments for the amount of credit received, the authors assessed the impact of this program. Their findings suggest that the Million Baht Village Fund injection of microcredit in villages did increase the overall credit in the economy. Households borrowed more, consumed more, and increased their earnings. A short-term effect of increasing future incomes and making business and market labor more important sources of income was also found. The increased borrowing and short-lived consumption response, despite no decline in interest rates, point to a relaxation of credit constraints. The increased labor income and especially wage rates indicate important spillover effects that may have also affected non-borrowers.

IV Estimates Only Local Effects

A second limitation of the IV approach is that it estimates only the local average treatment effects (LATE). This means that the IV estimates measure only the impact that the intervention had on those subjects that were affected by the instrument (Angrist and Kreuger 2001).[8] In many cases, these local firms are not necessarily the most important for national policy makers.

Regression Discontinuity

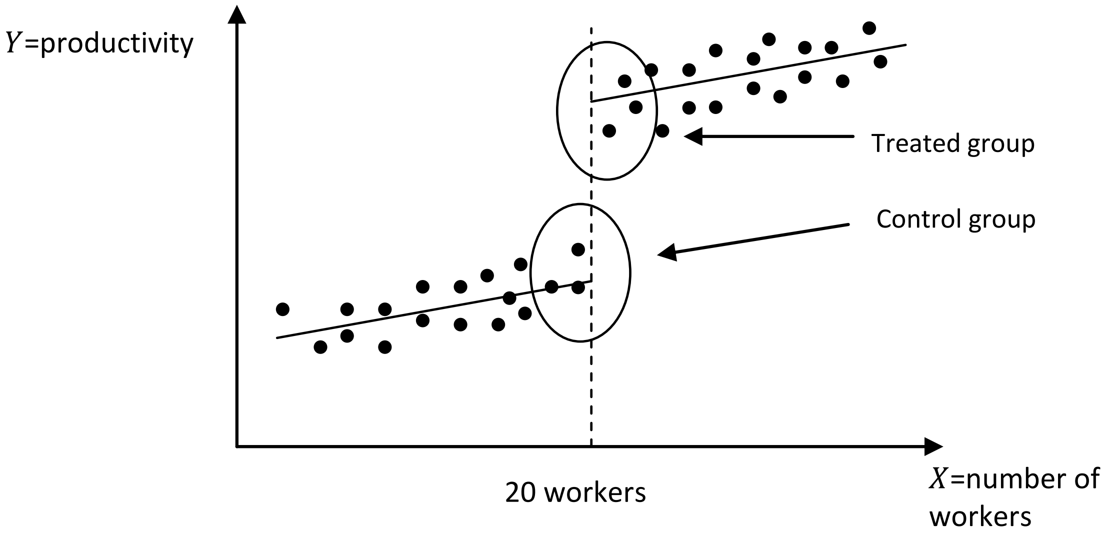
Regression discontinuity (RD) is a non-experimental approach used to evaluate interventions that have a defined cutoff for participation. For instance, a business training project aimed at increasing firms' productivity may be provided only to firms that employed more than 20 workers in the year before the intervention. This exogenous cutoff provides a design that allows the identification of the intervention's impact, since firms at the margin of the threshold would not differ substantially: there would be no reason to believe that a firm with 19 workers is different from a firm with 20 workers.

The assumption of this method is that at the margin of the cutoff, the assignment to the treatment and the control groups is close to random. By comparing the outcomes of treated firms (firms with 20 workers) with control firms (firms with 19 workers), evaluators can measure the intervention's effect (see Box 6 for an example).

Graphically, the outcome variable (that is, firms' productivity) should show a discontinuity at the cutoff value (that is, at 20 workers). Figure 5 illustrates this example.

One way to validate the RD estimates is to use pre-intervention data on the treatment and control groups and to analyze whether discontinuity exists between these two groups at the cutoff (Angrist and Pischke 2009). If no discontinuity is found for pre-intervention periods, then evidence supports that the discontinuity was generated by the intervention.

FIGURE 5. RANDOM DISCONTINUITY



[7] The instrument would violate this assumption if, even in the absence of credit, larger Thai villages might have experienced different trends in economic activity or business growth than smaller villages.

[8] See also Appendix 1 for a discussion of related issues that arise with RCT.

BOX 6. FINANCIAL INFRASTRUCTURE EVALUATION—ROLE OF ANGEL FUNDS IN U.S. START-UP FIRMS

Most equity funding of SMEs around the world comes from two sources: retained earnings and capital provided by personal savings, friends and family, and other “angel” investors.[1] Similar to venture capitalists, angel funds are investors for high-potential start-up investments, commonly structured as semiformal networks of high net worth individuals who decide to invest in projects of aspiring entrepreneurs based on their own assessments. To evaluate the impact of angel funds in U.S. start-up firms, Kerr, Lerner, and Schoar (2010) obtained information on prospective ventures from a large angel investment group. Using a regression discontinuity approach to evaluate the effect of angel funding on the performance of high-growth start-up firms, the authors compared firms that fall just above and just below the funding criteria of the angel group. The evaluation found a strong, positive effect of angel funding on the survival and growth of ventures.

[1]World Bank Enterprise Surveys: http://www.enterprisesurveys.org/.

Key Assumptions

The key assumption behind the RD approach is that the potential outcome (that is, firms’ productivity) may be associated with the cutoff variable (that is, number of workers), but in a smooth manner. In other words, in the absence of the intervention, this association should have been smooth at the cutoff. In this way, any discontinuity in the potential outcome at the cutoff is interpreted as a causal effect of the intervention. This is known as the continuity assumption (Van der Klaauw 2008).

Strengths

Baseline Data Are Not Needed

One benefit to using an RD design is that baseline data are not needed to estimate the impact. However, data from pre-intervention periods are strongly recommended to perform robustness checks on the validity of the discontinuity.

RD Estimates Are Comparable to Randomized Estimates

A second advantage of the RD approach is that from a methodological point of view, a solid RD design is comparable in internal validity to a randomized experiment.

Limitations

Independence of Threshold

The most important issue to consider when implementing an RD evaluation is the validity of the cutoff. If the cutoff was assigned with the objective of maximizing the intervention’s impact, then conclusions from the RD will be biased. The cutoff selected must be independent of the expected outcomes from the intervention. Suppose that in the example of the business training project, firms with at least 20 workers are concentrated in the most developed region of the country. Firms in this region are more likely to have access to finance. Thus, the effects of providing business training are likely to be higher if the cutoff is 20 workers, since these firms will also have access to better terms of credit, likely increasing their productivity, than if the cutoff were 15 or 10 workers.

Manipulation of the Assignment

Moreover, RD inferences will be invalid if firms are able to manipulate assignment into the program. For instance, if the cutoff is a specific number of employees, then firms can easily hire one more employee to participate in the intervention, prompting selection issues that contaminate the RD impact. As long as firms are unable to manipulate their eligibility into the program, the RD estimates are valid. Thus, RD design is more flexible than the IV approach since the IV methodology requires that the instrument is exogenous to the outcomes and that firms are not able to manipulate the assignment (Lee and Lemieux forthcoming).

Sufficient Observations Close to the Cutoff

A second issue of the RD approach is that in order to measure impact estimates, sufficient observations in close proximity to the cutoff must be available. In the business training example, sufficient firms with 18 to 22 workers (a number that is close to the cutoff of 20) would be needed to evaluate the RD effect.

Estimated Parameters Might Not Be the Most Important Ones

As in the case of the IV methodology, RD estimates can only estimate the average treatment effect of observations close to the cutoff (that is, the local treatment effect). This implies that it might be difficult to draw conclusions about the impact of the intervention for firms away from the cutoff of 20 workers.

Propensity Score Matching

Propensity score matching (PSM) is a non-experimental approach that can be used to analyze the impact of an SME intervention in which (1) the institutional arrangements that defined selection into the project are known by the evaluator and (2) a control group is not maintained. Under these circumstances, the PSM approach can identify a control group from the group of firms not participating in the program.

The intuition of this method is to find a control group whose observable characteristics are similar to the treated group but that did not participate in the intervention. The impact of the intervention will then be measured as the difference in outcomes between the treated group (that is, firms participating in the program) and control group (comparable firms not participating in the program). The approach matches treated firms to non-treated ones using propensity scores that summarize all observable information used to assign treatment (or eligibility to the program). Thus, PSM can be used to identify a control group that is statistically equivalent to the treatment group. As in all other approaches, the control group is used to infer what would have happened to intervention participants without it.

To compute the propensity score, one must estimate the conditional probability of participating in the intervention as a function of the observed characteristics.[9] These characteristics are then aggregated into the score. Once a control group is identified, the impact of an intervention is measured by the difference in outcomes between the treated and control groups (see Box 7 for an example).

Key Assumptions

The assumption underlying the PSM estimates is known as the conditional independence assumption. This assumption implies that after controlling for observable differences between the treated and control group, the outcome resulting in the absence of the intervention would be the same in both cases. Thus, conditional on the score, any differences between the treated and control group are attributed to the effect of the intervention.

In other words, this assumption implies that using observed information from SMEs is enough to identify a statistically equivalent control group. This assumption is unlikely to hold in SME interventions in which firms self-select to participate based on factors that are difficult to observe from the data, such as entrepreneurial attitudes, managers' skills, or risk aversion. If these unobserved factors are driving firms' participation in the program, then the PSM approach will fail to identify a proper control group.

[9] The conditional probability can be estimated through a probit or a logit model in which the dependent variable is an indicator variable equal to 1 if the subject participated in the intervention, and 0 otherwise. The independent variables are the observed characteristics that determined participation in the intervention.

Strengths

PSM Makes It Possible to Identify a Control Group When the Eligibility Criteria Are Known and Observed

The overall advantage of the PSM approach is that a control group can be identified when the selection process is known and observed.

The PSM approach is especially useful when several characteristics influence the eligibility for an intervention, since it provides a natural weighting scheme (the score) that yields unbiased estimates of the intervention effect (Dehejia and Wahba 2002).

Limitations

PSM Is Data Intensive

Data on sufficient firms and detailed information on their characteristics are needed to identify a control group that is statistically identical to the treated group.

PSM Does Not Control for Unobserved Self-selection

If unobservable characteristics also influence participation in the intervention and outcomes (self-selection issues such as the ones discussed in the example), then the PSM by itself is not an appropriate method. This could be the case when participating entrepreneurs or firms self-select in the intervention for reasons that also influence their performance. Evaluations using PSM in these situations tend to at least combine PSM with an alternative approach, such as DD, in order to remove the bias due to time-invariant unobservable characteristics (such as motivation, skills, or risk aversion).

Eligibility Criteria Must Not Be Associated with Participation in the Intervention

Another issue to take into consideration when using PSM is that information from the institutional arrangements of the intervention is needed to identify the participant selection characteristics (Caliendo and Kopening 2008). For valid PSM estimates, these variables must not be affected by participation in the intervention.

BOX 7. PUBLIC INTERVENTION EVALUATION: CHILE'S SUPPLIER DEVELOPMENT PROGRAM

In Chile, the Suppliers Development Program encouraged large firms to invest in the training of their SME suppliers, strengthening the linkage between large (potentially exporter) firms and SMEs. Large firms participating in the program were expected to provide professional advice, personnel training, technical assistance, or technology transfer to their SME partners. The program would then subsidize the cost of these activities. Each project participating in the program consisted of one large firm that sponsored the knowledge transfer and at least 20 SMEs in the agriculture and forestry sector, or at least 10 SMEs in other economic activity sectors.

An evaluation of the program was done by Arraiz, Henriquez, and Stucchi (2011). Administrative data allowed the evaluators to follow beneficiary and non-beneficiary firms for several years before and after the program was in place. To identify a control group, the evaluators estimated the propensity score using the probability of participating in the program with firms' information from 2002, the year before the beneficiaries started participating in the program. The score helped the evaluators determine a control group, which was composed of firms that did not take part in the program but that had similar probabilities of participating.

A concern of evaluators was that unobserved characteristics of firms (such as managers' skills or motivation) could have influenced their participation into the program and their success in it. In such cases, the PSM approach should be combined with other evaluation methods that control for unobserved information that might influence self-selection. The evaluators combined PSM with the DD approach, since DD estimates control for all unobserved information between the treated and control groups that do not change over time. After identifying their control group through PSM, the evaluators estimated the DD effect of the program. The evaluation found that both local SME suppliers and large firms benefited from participating in it. Local SMEs that participated in the program increased sales and employment. Large firms increased their sales and their likelihood of becoming exporters.

VII. Minimal Standard Monitoring

Minimal standard monitoring typically refers to before-and-after comparisons that monitor over time the performance of the subjects affected by an intervention. The main distinction between a minimal standard monitoring and an impact evaluation approach is that minimal standard monitoring does not follow a control group to learn what would have happened to the treatment group in the absence of the intervention.

Suppose, for instance, that evaluators are interested in analyzing the impact that a public credit program has on the profits of SMEs. To do a minimal standard monitoring, the only data needed would be information before and after the program on the profits of SMEs that participated in the program. The before-and-after effect is then measured by the difference in the average profits before and after the program.

An advantage of this approach is that evaluators only need to have information on the subjects of interest before and after the reform took place. Compared to rigorous impact evaluations, this approach demands the least amount of data.

A second advantage regards budget. While the difference in cost between rigorous impact assessments and before-and-after comparisons should not be substantially different if data collection is not needed, impact assessments still need to reserve budget for monitoring costs of the evaluation and researchers' time; whereas in minimal-standard monitoring, if these costs exist, then they should be lower.

The drawback of using before-and-after comparisons is that there is no control group that allows us to know what would have happened if firms had not received the intervention. With this method, the odds of falsely attributing an effect are large. This method can only identify how subjects change over time. Part of these changes might be attributed to the intervention, but any other factor changing over time parallel to the intervention (such as economic growth or changing macroeconomic conditions) will contaminate the evaluation. Therefore, we are not able to confidently measure and isolate the impact of the intervention.

VIII. Conclusions

As stated in the *SME Finance Policy Guide* (GPFI 2011), further work is needed on impact assessment techniques for SME finance policies and interventions. Only a handful of rigorous studies exist. More studies are needed on a wider range of policies in a number of different institutional settings to learn what works, where, and why. To identify good practice models, it is important to examine if the results of certain policies can be repeated in other environments.

This Framework is intended as a resource for policy makers and regulators to select adequate approaches to evaluate SME finance policies and interventions. While the focus of the Framework is on SME finance policies, the methods described can be applied to evaluate a broader set of SME interventions. The paper reviews a variety of impact evaluation methods—randomized experiments, difference-in-difference, propensity scoring, and regression discontinuity designs—and provides recommendations on how to map the various techniques to interventions spanning regulatory and supervisory frameworks, financial infrastructure programs, and public interventions.

It is important to understand and consider all possible evaluation options and not focus on any single approach, such as randomization. While randomization has many advantages, it is not necessarily the optimal choice in all situations, and it has its own limitations that need to be addressed in carefully planned and implemented studies. The impact evaluation studies should be driven by important policy questions rather than by methods of evaluation.

McKenzie (2010) argues that the SME sector is one area that is particularly full of unexploited possibilities for impact evaluations: “SME focused policies are typically carried out by governments and international financial institutions (IFIs) rather than NGOs, and are too expensive usually for researchers to fund the program on offer themselves. As a result, there is a real knowledge gap—and an opportunity to be grasped. If governments and operations staff at IFIs can work with researchers in evaluating the many projects being implemented, it should be possible to evaluate rigorously many of the policies being carried out for SMEs and to learn where modifications of existing strategies are needed.”

In summary, more work is needed to evaluate the wide variety of SME finance policies, and international organizations are well suited to fill in these knowledge gaps. As Duflo and Kremer (2005, p.342) state, “The benefits of knowing which programs work and which do not extend far beyond any program or agency, and credible impact evaluations are global public goods in the sense that they can offer reliable guidance to international organizations, governments, donors, and NGOs beyond national borders.”

References

Angrist, Joshua D., and Guido Imbens. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62 (2): 467–75.

Angrist, Joshua D., and Alan B. Kreuger. 2001. “Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments.” *Journal of Economic Perspectives* 15 (Fall): 69–85.

Angrist, J.D., and J.S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist’s Companion.* Princeton, NJ: Princeton University Press.

Arraiz, I., F. Henriquez, and R. Stucchi. 2011. “Impact of the Chilean Supplier Development Program on the Performance of SME and Their Large Firm Customers,” Working Paper, Inter-American Development Bank, Washington, DC.

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. “Household Decision Making and Savings Impacts: Further Evidence from a Commitment Savings Product in the Philippines,” Working Paper 939, Economic Growth Center, Yale University, New Haven.

Bauchet, Jonathan, C. Marshall, L. Starita, J. Thomas, and A. Yalouris. 2011. “Latest Findings from Randomized Evaluations of Microfinance,” Consultative Group to Assist the Poor Report No 2, Washington, DC, December. http://www.cgap.org/gm/document-1.9.55766/FORUM2.pdf.

-----------. 2008. “License to Sell: The Effect of Business Registration Reform on Entrepreneurial Activity in Mexico,” *Policy Research Working Paper* 4538, World Bank, Washington, DC.

Bruhn, Miriam. 2011. “License to Sell: The Effect of Business Registration Reform on Entrepreneurial Activity in Mexico.” *Review of Economics and Statistics* 93(1): 382–386.

Bruhn, Miriam, and Bilal Zia. 2011. “Stimulating Managerial Capital in Emerging Markets—The Impact of Business and Financial Literacy for Young Entrepreneurs,” Policy Research Working Paper 5642, World Bank, Washington, DC.

Bruhn, Miriam, and I. Love. 2009. “The Economic Impact of Banking the Unbanked: Evidence from Mexico,” Policy Research Working Paper 4981, World Bank, Washington, DC.

Burgess, R., and R. Pande. 2005.”Can Rural Banks Reduce Poverty? Evidence from the Indian Social Banking Experiment.” *American Economic Review* 95 (3): 780–95.

Caliendo, M., and S. Kopening. 2008. “Some Practical Guidance for the Implementation of Propensity Score Matching.” *Journal of Economic Surveys* 22: 31–72.

Cole, Shawn, T. Sampson, and B. Zia. 2011. “Prices or Knowledge? What Drives Demand for Financial Services in Emerging Markets?” *Journal of Finance* 66 (6): 1933–67.

———. 2009. “Financial Literacy, Financial Decisions, and the Demand for Financial Services: Evidence from India and Indonesia.” Working Paper 09-117, Harvard Business School, Cambridge, MA.

De Mel, Suresh, D. McKenzie, and C. Woodruff. 2008a. “Are Women More Credit Constrained? Experimental Evidence on Gender and Microenterprise Returns.” *American Economic Journal: Applied Economics* 1(3): 1-32.

———. 2008b. “Returns to Capital: Results from a Randomized Experiment.” *Quarterly Journal of Economics* 123 (4): 1329–72.

Dehejia, R., and S. Wahba. 2002. “Propensity Score Matching Methods for Non-Experimental

Causal Studies." *Review of Economics and Statistics* 84 (1): 151–61.

De Janvry, Alain, C. McIntosh, and E. Sadoulet (2008) "The Supply- and Demand-Side Impacts of Credit Market Information," San Diego, University of California–San Diego, unpublished.

Duflo, Esther, and Michael Kremer. 2005. "Use of Randomization in the Evaluation of Development Effectiveness." In *Evaluating Development Effectiveness*, ed. Osvaldo Feinstein, Gregory K. Ingram, and George K. Pitman, 205–32. New Brunswick, NJ: Transaction Publishers.

Duflo, Esther, and Emmanuel Saez. 2003. "The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment." *Quarterly Journal of Economics* 118(3): 815–842.

Gertler, Paul J., Sebastian Martinez ,Patrick Premand, Laura B. Rawlings, and Christel M. J. Vermeersch. 2011. "Impact Evaluation in Practice." World Bank, Washington, DC.

Global Partnership for Financial Inclusion (GPFI). 2011. "SME Finance Policy Guide." Paper on Behalf of the Global Partnership for Financial Inclusion. IFC, Washington, DC.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1): 5–86.

Kaboski, Joseph P., and Robert M. Townsend. Forthcoming. "A Structural Evaluation of a Large-Scale Experimental Microfinance Initiative." *Econometrica*.

Karlan, Dean, and Jonathan Zinman. 2010. "Expanding Credit Access: Using Randomized Supply Decisions to Estimate the Impacts." *Review of Financial Studies* 23 (1): 433–64.

———. 2009. "Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment." Econometrica 77(6): 1993–2008.

———. 2008. "Credit Elasticities in Less-Developed Economies: Implications for Microfinance," American Economic Review 98(3): 1040–68.

Kerr, W. R., J. Lerner, and A. Schoar. 2010. "The Consequences of Entrepreneurial Finance: A Regression Discontinuity Analysis." Working Paper 10-086, Harvard Business School, Cambridge, MA.

Lee, D. S., and T. Lemieux. Forthcoming. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*.

McKenzie, David, and Christopher Woodruff. 2008. "Experimental Evidence on Returns to Capital and Access to Finance in Mexico." *World Bank Economic Review* 22(3): 457–82.

McKenzie, David. 2010. "Impact Assessments in Finance and Private Sector Development: What Have We Learned and What Should We Learn?" *World Bank Research Observer* 25(2): 209-33.

Ndovie. 2010. "Malawi Business Environment Strengthening Technical Assistance Project (BESTAP) Impact Evaluation." Presentation, Dakar.

Ravallion, Martin. 2009. "Should the Randomistas Rule?" *The Economists' Voice* (February). www.bepress.com/ev.

Storey, D. J., and J. Potter. 2007. "OECD Framework for the Evaluation of SME and Entrepreneurship Policies and Programme." Organisation for Economic Co-operation and Development (OECD), Paris.

Todd, Petra E., and Kenneth I. Wolpin. Forthcoming. "Structural Estimation and Policy Evaluation in Developing Countries." *Annual Review of Economics*.

Van der Klaauw, W. 2008. "Regression Discontinuity Analysis: A Survey of Recent Developments in Economics." Labor 22 (2): 219–45.

Winters, P., L. Salazar, and A. Maffioli. 2010. "Designing Impact Evaluations for Agricultural Projects," Impact Evaluation Guidelines, Strategy Development Division, Technical Notes IDB-TN-198, Inter-American Development Bank, Washington, DC.

World Bank, 2012. "Impact Evaluation Toolkit." World Bank, Washington, DC.

Appendix 1. General Concerns

As discussed throughout the Framework, each method has its limitations; however, a number of concerns apply to all impact evaluation methods. In this section we review such general concerns.

Biases—Selection, Attrition, and Spillovers

All impact evaluations face *selection bias* and need to have a credible way of addressing it. RCTs are best for addressing selection bias because they randomly assign units to be treated. However, several other sources of bias may still crop up in an RCT and may also be an issue in other types of impact evaluation.

One common problem is that the mere fact of being assigned to participate in a program (whether or not such assignment is done randomly) may cause

BOX 8. CHANGES IN BEHAVIOR IN RESPONSE TO PROGRAM ASSIGNMENT EXPERIMENT

One common concern with impact evaluations is that they can change the behavior of treatment and control groups. For example, if the treatment group receives a loan or a training program while the control group does not, then the treatment group may see this as a positive boost to entrepreneurs' morale, which may have an effect on their effort. This would contaminate the pure impact of the loan because the impact may be due to a short-term boost in morale and increased effort, and not to the additional finance or training content.

On the other side, individuals or firms in the control group may change their behavior in response to not being assigned into the program. For example, if some areas are affected and others are not, then individuals may move across the border into (or out of) the affected areas. In a delayed phase-in situation, when one area receives an intervention while another expects to receive it in the future, the possibility that the intervention is coming in the future is likely to affect behavior in the control group. Another example would be a program that involves collecting accounting data on firms as part of the baseline analysis. Here, the firms that are not in the treatment group may still change their behavior because their accounting data are collected and observed by the evaluators.

Thus, even if randomized methods have been employed and the intended allocation of the program was random, then the differences in behavior may contaminate this random assignment and produce biased results. Other approaches may also be subject to such sources of bias.

One advantage of experiments is that they can explicitly address any possible changes in behavior. For example, in Ashraf, Karlan, and Yin's 2006 study of a commitment savings account, the change in behavior for those who received information about the new account could come simply because of the reminder about the importance of savings. To deal with this possibility, the researchers introduced another treatment group that received marketing on the existing savings product, which also served as a reminder about the usefulness of savings. Thus, the possibility that the outcome for the new type of account was simply due to the change in savings behavior could be eliminated by adding this third group.[10]

[10] However, adding another group affects the issues of power, discussed above. This may explain why Ashraf, Karlan, and Yin (2006) find insignificant estimates for the coefficients on the third group.

the treatment or comparison group to change its behavior, which may contaminate the results of the experiment (see Box 8).

In addition, there may be *spillover* effects from those participating in a program in comparison to those that do not. For example, a program designed to enhance financial literacy of entrepreneurs may have spillover effects on those not receiving the program so that their literacy increases as well. This can easily happen if both treated and non-treated entrepreneurs belong to the same business association or have other social connections. Spillovers may also come from redistribution of resources by the government. For example, if some villages are positively affected by the experiment but others are not, then the local government may find other ways to channel resources to unaffected villages (Ravallion 2009).

If the spillover effects on non-treated individuals are generally positive, then the impact estimates will be smaller than they would have been without spillovers. This problem affects both randomized and nonrandomized evaluations. In some cases the experiments can be designed to directly measure the spillovers. For example, in their study of information and 401(k) participation, Duflo and Saez (2003) randomized the offer of getting an incentive to attend an information session at two levels. First, a set of university departments were randomly chosen for treatment, and then a random set of individuals within treatment departments were offered the prize. This allowed the authors to explore both the direct effect on attendance and plan enrollment of being offered an incentive and the spillover effect of being in a department in which others had been offered incentives.

Finally, there could be differences in attrition rates (that is, dropout) between treatment and control groups, which may also affect the results.[11]

Scaling Up and Systematic Effects

Many program evaluations, especially RCTs, are often of small-scale interventions and might have a different impact if implemented on a large scale.[12] For example, capital grants or directed loan programs for SMEs offered by governmental financial institutions may crowd out private sector loans. In the long run, capital grants may skew incentives of microentrepreneurs who will be waiting for grants rather than efficiently running their businesses. Such effects may be particularly important for assessing the welfare implications of scaling up a program. Scaling up programs raises several other issues (see Box 9).

Another example would be a small-scale training program that improves participants' chances to obtain a job. However, scaling up such a program may not necessarily raise aggregate employment because in a world with a fixed number of jobs, a training program could only redistribute the jobs (see Imbens and Wooldridge 2009).

[11] Attrition refers to a situation in which individuals or firms leave the sample observed by researchers. This could be due to closures for firms or a move for individuals or firms, or simply refusing to participate in subsequent surveys. If there are systematic differences in the attrition rates in the two groups, then the results may be biased in either direction. For example, if improving access to finance allows the weakest firms to survive, then the differences in attrition will make the group with access look weaker because it has a higher proportion of the weakest firms.

[12] In technical terms, RCTs estimate what are known as partial equilibrium treatment effects, which may differ from general equilibrium treatment effects (Duflo and Kremer 2005).

BOX 9. SCALING UP SMALL INTERVENTIONS

Scaling up a small program raises several additional issues.

Incentives. Most of the RCTs have been implemented by nongovernmental organizations (NGOs) or researchers, who are highly motivated to achieve the best possible outcome of the experiment. In addition, researchers often select the best NGOs to work with and test some of the products highly relevant to NGOs' work and image. Thus, experiments are often done under a set of ideal conditions, which may not be possible to replicate or scale up. The outcomes might be significantly different when the same program is implemented by government officials with a very different set of incentives (Ravallion 2009).

Allocation of resources. It is plausible that significantly more resources are allocated to the program during an experimental phase than would have been under a more realistic situation or in a less favorable context. Alternatively, such bias could go another way if the first phase of an experiment does not produce significant results because of ineffective implementation. However, the knowledge generated from the first phase would make subsequent phases more effective. Thus, it is important to understand the institutional and implementation factors that may make the same program successful in one place but not another.

Different outcomes. In an experimental setting, some firms with potentially low impact are mixed in with firms with potentially high impact from the same program because of the random assignment. If the program is scaled up, then the most likely takers will be firms with potentially high impact. Thus, the outcomes of a national program can be fundamentally different from those of an experiment because of the different types of individuals or firms participating (Ravallion 2009).

External Validity

In impact evaluation discussions, it is common to see references to the internal and external validity of the evaluation. *Internal validity* refers to ensuring that the measured impact is indeed caused by the intervention being tested, while *external validity* refers to the confidence that the impact measured in a specific study would carry over to other samples or populations.

RCTs in general have a good track record for ensuring internal validity (aside from the issues discussed above, which often can be addressed). However, RCTs often are criticized on the basis of their external validity (that is, transferability of the results to other situations, such as different samples of firms, or variations in policies or countries). For example, a specific program that was found effective for one type of firm in one country may not be effective for different types of firms in the same country or for the same type of firm in other countries. Alternatively, a program that had some minor variation from the one being tested may or may not be effective in the exact same situation as the one tested. While issues of external validity arise with other evaluation techniques, they more often appear in the context of RCTs.

One way to address external validity concerns is to replicate the evaluations in various settings. It is important to test how robust different programs are in different settings to produce valuable implementation knowledge. However, extensive replication is expensive and time-consuming.[13]

Another way to alleviate external validity concerns is to couple experiments with the theory of why the program is expected to work (see Duflo and Kremer 2005).

[13] In addition, researchers are unlikely to be interested in running the same program in different settings because the lack of novelty will greatly reduce the chances of publication.

Appendix 2. Size and Power of RCT

Sample sizes, as well as other design choices, will affect the power of an experiment. For example, if there are too few units in treatment or control groups, then the comparison of averages may not produce statistically significant results simply due to small sample. This can lead to erroneous conclusions. For example, the program may be deemed to have a significant effect when it actually does not, or the program may be deemed ineffective when it actually is effective.

The issue of power in RCTs can be addressed by ensuring a sufficient number of observations in each group and optimally dividing the proportion of individuals in treated and control groups based on the relative costs of treatment versus data collection. The larger the expected difference between treatment and control groups (that is, the effect size), the smaller the sample size needed for equal power.

Larger sample sizes are needed when there are several treatment groups and the researcher is interested in detecting the differences between various treatments in addition to detecting differences between treatment and control groups. Moreover, if researchers are interested in the effect of the program on a subgroup—for example, impact on female entrepreneurs relative to males—then the experiment must have enough power for this subgroup. This is nontrivial, especially in samples where female entrepreneurship is significantly less likely, which is not uncommon. Stratification methods can be used to ensure sufficient number of female entrepreneurs in the sample.

In some situations, the evaluation design concerns individuals or firms within the groups (for example, by randomly selecting villages and treating all individuals in a village), as the errors are likely to be correlated within the group. The larger the groups that are randomized, the larger the total sample size needed to achieve a given power.

Low take-up exacerbates the issues of power because it reduces the number of units on which to base the statistical analysis. For example, consider a program such as a new loan product or a business training that aims to raise the profits by 25 percent of microenterprises undertaking the program. A randomized experiment that offered the program to half the firms and used a single follow-up survey to estimate its impact would require a sample size of 670 firms if take-up was 100 percent, but would need a sample size of 2,700 with 50 percent take-up and 67,000 with 10 percent take-up.[14]

Thus, one solution to the problem of low take-up is to employ a very large sample so that the resulting sample will still contain enough firms or households to enable the researchers to detect a program impact of a given size. An example of a randomized experiment with sample sizes of this magnitude is seen in Karlan and Zinman (2009), where 58,000 direct-mail offers were randomly sent by a South African lender, with 8.7 percent of those contacted applying for a loan. However, the downside is that this solution can be very expensive and therefore not feasible in many situations.

The second solution to the low-power problem is to restrict the study to a group of units for which take-up would be much higher. For example, a business training program could be advertised to all eligible firms, and then the number of slots available in the program could be randomly allocated among the

[14] In addition, researchers are unlikely to be interested in running the same program in different settings because the lack of novelty will greatly reduce the chances of publication.

group of interested firms. Presumably, the take-up would be higher if the firms have already expressed interest in the training. An example of such a design is seen in Karlan and Zinman (2008), in which consumers first apply for loans and then the pool of marginally rejected candidates (all of whom wanted a loan) is randomly assigned to receive a loan.

The advantage of the second approach is that it requires much smaller samples to detect a treatment impact. The downside is that the program impact estimated will apply only to the self-selected group of individuals or firms that expressed interest in the program, not to the general population. For example, policy makers might be interested in the effect of the loan program on all firms or on firms interested in taking up credit. But an evaluation such as Karlan and Zinman (2008), based on the marginal applicants, only informs researchers of the impact on those firms that fall within a narrow band in terms of their creditworthiness according to the specific credit-scoring model used by the bank. Such firms may be different in important ways from the general population of firms. Thus, this experiment cannot be used to evaluate the impact of credit on all firms that desire credit or on the poorest segments of the population.

Appendix 3. Examples of Impact Evaluations

A discussion of several evaluation approaches by type of intervention is provided below.

Regulatory and Supervisory Frameworks

Entry of a New Bank in Mexico (DD Evaluation)

Bruhn and Love (2009) evaluate the impact on economic activity of the opening of a major bank in Mexico. In 2002, Banco Azteca opened more than 800 branches across the country. Branches were opened on the same day inside all of the preexisting stores of its parent company, Grupo Elektra.

Since Azteca entered only in municipalities with a preexisting Elektra store, these municipalities were used as the treatment group, and municipalities with similar characteristics but no Elektra store were used as the control group. Employing a difference-in-difference approach, the authors analyze the effect that Azteca had by comparing outcomes before and after it opened in both treatment and control municipalities. The gains from the opening of Banco Azteca are then the difference between the changes over time in treated municipalities and control municipalities.

The authors find that this bank had a significant impact on the economic activity of individuals belonging to the informal sector. Its opening increased the proportion of informal business owners by 7.6 percent and led to a higher proportion of women working as wage earners. Additionally, Azteca's opening increased income by about 9 percent for women and by about 5 percent for men.

2. Bank Branching Regulation in India (IV Evaluation)

Between 1977 and 1990, the Reserve Bank of India mandated that in order to open a branch in a location that already had bank branches, Indian banks had to open four branches in locations without banks. This policy expanded the presence of banks in rural areas of Indian states.

Burgess and Pande (2005) used an instrumental variables approach to evaluate the impact of this policy on poverty outcomes. The instruments were the policy-induced trend reversals of a state's initial financial development in its rural branch growth. In other words, less financially developed states in 1961 were less likely to receive bank branches in the periods outside the reform and substantially more likely to receive them during the years of the reform. As these trend reversals were significant in the years of the reform and had no direct impact on poverty outcomes, these instruments proved to be valid.

The evaluation concluded that rural branch expansion in India significantly reduced rural poverty. The reductions in rural poverty were linked to increased savings and credit provision in rural areas. By promoting the expansion of financial services into rural areas, this intervention allowed rural households to rely on more efficient mechanisms to accumulate capital and to obtain loans for longer-term productive investments.

Financial Infrastructure

Credit Information in Guatemala (RCT Evaluation)

Availability of information to evaluate SME creditworthiness is among the key institutional constraints limiting expansion of SME finance. Credit registries and bureaus could be an effective way to generate such information, as they contain historical information on repayment rates and current information on obligations. Establishment of credit bureaus is one of the policies that is likely to have economy-wide impact and thus is difficult to evaluate using an RCT.

De Janvry et al. (2008) used encouragement design to examine the impact of the introduction of a credit bureau in Guatemala (see also Boxes 3 and 5). They found that the awareness of the existence of a credit bureau was very low in surveys conducted soon after its implementation. They therefore randomly informed a subset of 5,000 microfinance borrowers about the existence of the bureau and how it worked. They found that awareness of the bureau led to a modest and temporary increase in repayment rates and to microfinance groups ejecting their worst-performing members.

Public Sector Interventions

Financial Support to Microenterprises in Sri Lanka and Mexico (RCT Evaluations)

Financing support for SMEs—whether through lines of credit, directed credit, cofinancing, equity financing, or other forms of direct financial assistance—is a popular form of intervention. Such interventions are based on the premise that a lack of finance hampers entrepreneurs, market failures prevent them from obtaining necessary capital, and therefore an injection of finance can put them on a path of increasing returns. However, credibly evaluating such programs requires distinguishing those that received the financial injection from those that did not, which is difficult because of self-selection issues (that is, enterprises that end up receiving a loan or a grant are different on many parameters, often unobservable, from those that do not receive such assistance).

Two recent studies use RCT to evaluate the effectiveness of grants to enterprises. De Mel, McKenzie, and Woodruff (2008b) study microenterprises in Sri Lanka, and McKenzie and Woodruff (2008) replicate the same experiment in Mexico. Grants between US$100 and US$200 were given to a randomly selected subset of microenterprises in each country. The authors find that the grants substantially raise incomes for the average firm receiving a grant and estimate real returns to capital of 5.7 percent per month in Sri Lanka and 20 percent per month in Mexico, much higher than market interest rates in both countries. In addition, the returns are highest for high-ability, credit-constrained firm owners, which is consistent with the view that credit market failures prevent talented owners from getting their firms to an optimal size. Interestingly, these studies find that the impact was similar whether the grants were given in cash or in the form of equipment or raw materials. On the flip side, the studies found that while one-time grants succeed in raising the incomes of poor business owners, they do not lead to significant job creation. Another surprising result of these studies is that grants did not raise the incomes of self-employed women; subsequent research has attempted to understand the reason for this result (De Mel, McKenzie, and Woodruff 2008a).

Studies like these can help policy makers design more effective interventions; however, more evidence may be needed before recommending that policy makers implement grant programs on a wide scale. Specifically, replicating similar experiments in other countries and with a variety of populations would show whether such policies would prove beneficial in other environments. In addition, while a small-scale intervention may be very helpful to those receiving the grants, the general equilibrium effects of implementing such policies on a wider scale need to be properly understood and investigated.

Financial Literacy Programs in Indonesia and the Dominican Republic (RCT Evaluations)

Financial literacy has come to play an increasingly prominent role in financial reform in both developed and developing countries, and is portrayed in global policy circles as a solution for many recent crisis-related financial problems. Many countries have set up financial literacy panels that are charged with developing financial literacy programs.

A recent study in Indonesia was designed to evaluate the causal relationship between financial literacy and demand for financial services (Cole, Sampson, and Zia 2011). The authors offered seminars to randomly selected groups and educated participants on the benefits and the procedure for opening savings accounts. The authors found an average negligible effect of such programs on the opening of new accounts; however, among the uneducated and financially illiterate households, there was a significant increase in opening new accounts. Moreover, they found small incentive payments to have a much larger effect on getting individuals to open bank accounts and to be three times as cost-effective as financial education. This study suggests a need for more research on the most effective ways to encourage households and microenterprises to save.

Drexler, Fischer, and Schoar (2011) report on two randomized trials to test the impact of financial training on firm-level and individual outcomes for microentrepreneurs in the Dominican Republic. They found no significant effect from a standard, fundamentals-based accounting training; however, a simplified, rule-of-thumb training produced significant and economically meaningful improvements in business practices and outcomes.

Partial Credit Guarantees in Italy (DD Approach)

In 1996, to promote lending to small firms, the Italian government established the Fund for Guarantee to SME, or SGS, with the generic mandate of providing direct guarantees to lending banks, co-guarantees together with other guarantor institutions, and guarantees of last resort to mutual guarantee institutions. To apply for a guarantee, an SME did not need to assess its degree of financial need. Instead, the SME needed to comply with a number of eligibility criteria, such as belonging to a specific sector and having sound economic and financial conditions. These criteria were then summarized in a scoring system that the SGS used to order applications according to their guarantee merit. Importantly, the eligibility criteria limited the percentage of applications that were rejected on merit grounds.

This paper used a difference-in-difference approach to test the fund's role in widening credit access for SMEs and lessening their borrowing costs. Using data from the fund's books, the authors compared outcomes of guaranteed SMEs with nonguaranteed SMEs before and after the SGS was launched. Specifically, the authors examined whether borrowing costs and access to credit, measured as the value of bank debt, were substantially different for SMEs that participated in the program than those that did not. The difference-in-difference effect can be interpreted as a causal impact as long as the average outcomes for the participating SMEs and the other firms would have followed parallel paths over time in the absence of the program. While this assumption is impossible to test, an exercise was performed to compare how different these two groups were before the program. The results from this exercise found no significant differences between the control and the treatment groups, validating the control group as a proper counterfactual. The difference-in-difference results from the paper suggest that Italy's scheme reduced participating SMEs' borrowing costs by 16 to 20 percent. Moreover, SMEs' bank debt increased by 12.41 percent once the scheme was available.

APPENDIX 4. ASSUMPTIONS, STRENGTHS, AND LIMITATIONS OF DIFFERENT APPROACHES

Comparison of Impact Evaluation Approaches			
Approach	Key assumptions	Strengths	Limitations
Randomized control trials (RCTs)	■ Subjects cannot manipulate assignment into the program ■ Subjects in the control group must be credibly excluded from receiving benefits from the intervention or program	■ Clear comparison group, which allows for credible identification of the impact	■ Not all policies are suitable for RCT ■ Local effects measured by RCTs might be different from systematic effects when a program is scaled up ■ External validity
Difference-in-difference (DD)	■ Trend of the treated group must be identical to trend of control group in the absence of the intervention	■ DDs control for factors (observed and unobserved) that do not vary over time ■ Cost-effective impact evaluation method	■ DD estimates are invalid if changes over time occurred to one group but not the other, or if the two groups had different trends before the intervention
Instrumental variables (IV)	■ Instrument must be strongly associated with participation of the policy and must not be associated with the outcomes evaluated	■ IV estimates control for unobserved information that may influence self-selection into the program	■ If not planned ahead, IV evaluations are difficult to do ■ IV results estimate only local effects
Regression discontinuity (RD)	■ In the absence of the intervention, the cutoff variable should be associated with the outcome variable in a continuous manner	■ Baseline data not needed ■ Solid RD estimates are comparable to RCT estimates	■ RD effects will be biased if the cutoff was assigned to maximize the impact of the intervention, or if firms are able to manipulate their eligibility into the program ■ Sufficient observations surrounding the cutoff are needed ■ RD results estimate only local effects
Propensity score matching (PSM)	■ After controlling for observed differences, outcomes of treated group identical are to outcomes of control group in the absence of the intervention	■ PSM allows identification of a control group when the eligibility criteria depend on multiple variables	■ PSM is data intensive ■ PSM estimator not robust against bias caused by unobserved information associated with participation in the program (in this case, PSM can be combined with other approaches) ■ PSM should be used only in cases where the evaluator has a clear and detailed understanding of the eligibility criteria of an intervention
Minimal standard monitoring	■ Outcomes of treated subjects are not being affected by any other factor except by the policy of interest	■ Relatively easy method to implement—does not require significant technical capacity from the evaluating team or the principal investigator	■ Results should be treated with caution since other factors besides the policy may be contaminating the results

THE WORLD BANK