USAID
FROM THE AMERICAN PEOPLE

METHODOLOGICAL ISSUES IN CONDUCTING IMPACT ASSESSMENTS OF PRIVATE SECTOR DEVELOPMENT PROGRAMS

IMPACT ASSESSMENT PRIMER SERIES
PUBLICATION #2

PRIVATE SECTOR DEVELOPMENT IMPACT ASSESSMENT INITIATIVE

DECEMBER 2006

This publication was produced for review by the United States Agency for International Development. It was prepared by Lucy Creevey, an independent consultant, and Gary Woller, of Woller and Associates.

METHODOLOGICAL ISSUES IN CONDUCTING IMPACT ASSESSMENTS OF PRIVATE SECTOR DEVELOPMENT PROGRAMS

IMPACT ASSESSMENT PRIMER SERIES
PUBLICATION #2

PRIVATE SECTOR DEVELOPMENT IMPACT ASSESSMENT INITIATIVE

CONTENTS

INTRODUCTION

The objectives of assessing the impact of private sector development (PSD) programs impact assessment are, at least, three-fold:

1. to identify program outcomes, whether positive or negative,
2. to determine whether outcomes can be attributed to (or was caused by) the program, and
3. to provide an in-depth understanding of the various causal relationships and the mechanisms through which they operate.

The principle of attribution is what distinguishes impact assessment from program monitoring or program evaluation. Attribution requires comparing observed outcomes to a counterfactual, which is an estimate of what would have happened if the program had not been undertaken. In order to make the comparison of observed outcomes to counterfactuals, researchers may choose from a variety of research methodologies. The methodology chosen will determine to a large degree the credibility of assessment findings.

Each methodology has its advantages and disadvantages. Ideally impact assessments will use rigorous experimental or quasi-experimental quantitative methodologies complemented by appropriate qualitative methodologies using trained and experienced field researchers. Notwithstanding, resources for conducting rigorous *mixed-methodology* assessments may not exist, in which case evaluators must often adopt less rigorous approaches or use less qualified researchers. Even in these cases, however, impact assessments must satisfy certain *minimally acceptable* methodologies for the results to be credible. By minimally acceptable is meant the bare minimum of methodologies, which may be less costly, but which remain within the bounds of credible research.

Deciding which methodologies to adopt requires in turn knowledge about methodological options available. It also requires knowledge about a range of methodological issues that inevitably arise in designing impact assessments and how to deal with them. In both cases, it further requires knowledge about how choices made affect the credibility of the impact assessment findings.

In this light, this contribution to the Impact Assessment Primer Series reviews the methodological options available to impact assessment researchers—including both quantitative and qualitative methods, their relative strengths and weaknesses, and their related implications—and it reviews important methodological issues that arise in designing impact assessments, discusses their implications for assessment findings, and offers suggestions for dealing with them.[1] The article concludes with a description of minimally acceptable methodological standards for conducting impact assessments of PSD programs.

QUANTITATIVE METHODOLOGIES

The most commonly used quantitative methodologies are longitudinal and cross sectional surveys. Longitudinal surveys compare changes that have occurred among a group (or panel) of respondents over time. Longitudinal surveys are usually costly, as they require large sample sizes to ensure representativeness and entail at least two separate survey rounds (baseline and follow-up). They also involve logistical difficulties related to locating and interviewing persons from the baseline survey often two or more years after the initial baseline.

---

[1] For a more in-depth discussion of qualitative and quantitative methodologies, their roles, and relative strengths and weaknesses, see Lucy Creevey, (2006), "Collecting and Using Data for Impact Assessment," Impact Assessment Primer Series paper #3.

Cross-sectional surveys examine impact through a one-time survey. They tend to be more time efficient and less costly than longitudinal surveys, but they too have their limitations. In particular, measuring impact at one point in time limits the ability to measure change in a reliable way and establish plausible evidence of impact. Despite their cost and difficulties, longitudinal surveys are generally preferred to cross-sectional surveys in terms of generating credible and useful information about program impact.

Quantitative survey methodologies can be further grouped into three categories: experimental, quasi-experimental, and non-experimental methodologies.

EXPERIMENTAL METHODOLOGIES

An experiment is defined as a deliberate attempt to administer a treatment in order to observe the effects of that treatment. The experimental method employs random assignment of persons as clients, and thus into the treatment group, or non-clients, and thus into the control group. It may also involve randomized selection of areas of program operation or expansion.

The advantage of randomization is that the whole population of potential program recipients, as defined by the relevant program selection criteria, is included in the sampling frame, and a purely random selection of initial program recipients are compared to a purely random set of non-clients. Since they are drawn from the same population and satisfy the same program selection criteria, treatment and control group members selected via randomization techniques can be assumed to be similar on key observable and non-observable characteristics. Randomized sample selection is thus more scientific in nature, and it minimizes potential biases.

Experimental methods may also be cheaper and easier to do because it can be (and usually is) done with the collaboration of the program administrators who use the assessment results for their own program purposes. In addition, experimental methods do not require a baseline survey, because the evaluation rests on outcome differences between the two groups at a point of time that allows for the program to achieve significant impact. Randomization makes sense when program expansion is constrained and program administrators are indifferent between offering the program to one cohort of recipients or another and/or in one location or another.

Despite the methodological and cost advantages of experimental methods, randomized evaluations often cannot be fully achieved. To begin with, the program must have sufficient discretion over client and site selection to permit random assignment into the treatment and control groups. For practical and legal reasons, this is not always possible. An example would be if the program worked through program partners as the primary implementers but had insufficient influence over the partners to convince them to adopt randomized client selection. Program managers or partners, moreover, frequently have explicit reasons for selecting one group of persons over another (e.g., they belong to a specific target group) or for selecting one area over another (e.g., an area is more accessible). Guaranteeing that selection into treatment and control groups is random and not tied to population, geographic, or characteristics can be difficult.

Second, there is a question of morality or undue influence. Randomization systematically prevents certain people (because of having been chosen randomly) from receiving assistance, and many program managers, program partners, or other stakeholders will therefore not permit it.

Third, randomization can at times be more unwieldy than traditional quasi-experimental methods, and it would be very difficult to extend the impact assessment beyond a target group at one level of the value chain.

Fourth, it is not always possible to start the impact assessment concurrently with the program itself. Impact assessments frequently commence only after the program has started, ruling out randomized client selection, at least initially.[2] Randomization may still be employed in this case in conjunction with program expansion.

QUASI-EXPERIMENTAL METHODOLOGIES

Quasi-experimental methodologies were developed to deal with the messy world of field research, where it is not always practical, ethical, or even possible to randomly assign firms or persons to client and non-client groups. In contrast to experimental methods, quasi-experimental methods do not randomly assign units to treatment or control groups but compare groups that already exist. Treatment group members are selected via random sampling of known program clients, while control group members are selected via random sampling of known non-clients. Even with a valid control group, however, quasi-experimental methodologies cannot be regarded as definitive *proof* of impact, since the control group is only similar in certain key respects to the treatment groups, not identical.

Cross-sectional surveys are a form of quasi-experimental methodology that offer a lower-cost alternative to longitudinal surveys, while avoiding the problem of panel attrition (see below). The limitations of this approach have been outlined above, and it is definitely not the preferred research strategy. However there are circumstances when it may be considered. For example, a program may be well underway or almost completed (thus making a baseline not possible) when program stakeholders decide that it would be important to assess the impacts of the program's work.

NON-EXPERIMENTAL METHODOLOGIES

Non-experimental methodologies examine only program clients without comparison to a control group of non-clients. A common example is the "pre-post study," which examines client outcomes both before and after program benefits are received using retrospective perceptions (the client is asked to compare her current state to her state at an earlier point in time). Another common non-experimental methodology is "longitudinal studies," which examine changes among program clients over time and relate the changes back to a baseline taken at program entry or at some earlier point of program participation. Non-experimental methodologies lack rigor and have a high risk of leading to false conclusions. They should, therefore, be used sparingly and only where other feasible options do not exist.

QUALITATIVE METHODOLOGIES

If done appropriately, quantitative methodologies can achieve high coverage and representativeness, while yielding accurate data and generalizeable results. However, they have a limited ability to grasp qualitative information, complex relationships, reverse causality, and potential negative impacts. Creating understanding of these issues is the role of qualitative methodologies.

Qualitative methods include such techniques as focus group discussion, key informant interviews, case studies, rapid appraisals, and participant observation. These methods allow for in-depth analysis in order to capture complex causal relationships and contextual information against which results can be properly understood. They can be fairly flexible in their design, allowing the researcher to probe and add more questions as the research proceeds, in order to gain greater insight into a program's context and into the attitudes and reactions of people.

Qualitative methodologies, however, tend to be limited in coverage and are by nature non-representative. Accordingly, they produce less reliable results in terms of standardization of data, representativeness, and ability to generalize and

[2] Commencing the impact assessment after program launch is appropriate if the program is early into its implementation cycle, participants have not had much time working with the program, and no impacts are expected to have occurred.

quantify impact. Qualitative methodologies alone generally are not sufficient to establish attribution. For this reason, they have been widely used to improve program impact rather than prove it.

MIXED METHODOLOGIES

In order to achieve a balance between credibility and cost, many researchers have resorted to a mix of quantitative and qualitative methodologies. They seek to estimate the amount, pattern or direction of change that can be plausibly associated with a PSD program and attribute it to program intervention with a high degree of confidence. Combining surveys with one or more qualitative methodologies can yield credible results and provide richer insights, while optimizing on time and cost. Many impact assessments are relying increasingly on triangulation, through the integration of findings yielded by three different methods of data collection (such as a combination of surveys, focus group discussions, and in-depth interviews). Triangulation resolves issues related to sample size and allows for cross-checking of findings to validate data and ensure the reliability of responses.

All else equal, the use of mixed methodologies is the preferred approach to impact assessment, as it is the approach most likely to yield credible in-depth findings.

IMPORTANT ISSUES IN IMPACT ASSESSMENT METHODOLOGY

There are a number of important methodological issues to consider in designing an impact assessment. These include, among others, the following:

- Research team selection
- Treatment and control group selection
- Sample size and panel attrition
- Spillover impacts
- Program expansion

These issues are discussed in greater depth in this section.

RESEARCH TEAM SELECTION

The choice of field researchers to conduct the impact assessment will also affect the reliability of the findings. It is strongly recommended that trained and experienced field researchers who are not involved in the program do the work. They should also have sufficient background on the country and on the program so as to understand the program context.

A trained external research team—which may include one or more researchers from outside the country, an in-country team, or a combination—is more likely to be objective about the program outcomes (having nothing to gain or lose from the results). It is also more likely to have the specific research skills needed than program personnel who are unlikely to have the required training in survey design, sampling, data processing or data analysis or to be skilled in conducting qualitative research.

The obvious disadvantages to hiring external researchers, especially those from the US or the EU, are cost, availability and timing. Foreign researchers may make the cost of the assessment significantly more than an in-country team, if, in fact, the appropriate in-country researchers are available at the time the research is scheduled to be done, and if they will commit to being available for all the different phases of the research. If there are highly trained researchers in the country available for this work, the costs may be greatly reduced, although not necessarily. If only a few local

researchers (or research firms) are available and are already working for other NGOs and donor agencies, they may charge high fees as well.[3]

Alternatively, the cost can be reduced by using program staff for some of the work. This approach might work if staff have a strong interest in learning whether program interventions have had the desired impacts and, particularly, which strategies appeared the most successful among which groups and why. The feasibility of this approach will also be greater to the extent the assessment work can be integrated into the relevant program monitoring activities.

An external researcher could design or help design the research plan and work with staff members who may administer the survey, conduct interviews, or run the focus group discussions. In general, however, program staff will not have time to do research that is not directly part of their program work, while the potential for biasing the process and findings using program staff can never be fully discounted. Regardless, if program impact is to be demonstrated, then the research must be as professionally and scientifically done as possible and with the least obvious bias. This typically rules out an extensive role for program staff in conducting the impact assessment. A general recommendation is that the impact assessment be carried out by a qualified and independent local research partner under the supervision of an assessment expert or experts (local or international) hired to manage the impact assessment process.

TREATMENT AND CONTROL GROUP SELECTION

Impact assessments that address program clients alone run the risk of false attribution. Attributing observed changes to program participation requires examination of a control group of non-clients in addition to a treatment group of program clients. Selecting and interviewing a valid control group can be difficult and costly, depending on the nature of the program. Despite its cost and difficulty, the importance of selecting a valid control group cannot be overstated. Absent a valid control group, observed outcomes cannot be credibly attributed to the program.

To create a valid control group, it is vitally important to match control group members to treatment group members on key observable (e.g., age, gender, education) and unobservable (e.g., entrepreneurship, risk taking, attitudes). It is also important to match the groups on geographic characteristics (e.g., climate and soil conditions for farmers or city size and characteristics for urban service providers). Under certain conditions, well-matched treatment and control groups can achieve levels of rigor comparable to random assignment, but, unlike random assignment, doing quasi-experimental assessments increases the risk of misleading results due to so-called *selection bias* in creating the treatment and control groups.

Failure to match treatment and control groups appropriately creates *selection bias*. In principle, selection bias can either inflate or deflate measured impact, but the more common case in quasi-experimental studies is that impact is overstated because program participants have advantages (related to location or personal characteristics) that would have led to higher performance on impact variables, even if they had not participated in the program. Some degree of selection bias is likely to be present in any quasi-experimental study. The aim is to keep it small enough that it does not invalidate the assessment findings.

SAMPLE SIZE AND PANEL ATTRITION

The purpose of the baseline and follow up surveys in a longitudinal study is to systematically identify important impacts on program clients by comparing observed outcomes among clients to those of a control group of non-clients. Both baseline and follow-up surveys must contain a representative sample of clients and non-clients, including

[3]A case in point is an ongoing impact assessment conducted by the PSD Impact Assessment Initiative in Zambia where local researcher firms charge as high as $400-$500 per day for their top-level researchers.

a statistically significant number of respondents within key subgroups that the researchers wish to compare. These subgroups may include different production areas, women versus men, those from remote rural areas versus those from peri urban or urban areas, those with larger versus medium and small businesses, and so forth.

This is not easy in practice. In the interim between the baseline and follow-up survey, people drop out of the survey, for a variety of reasons. People move, die, may not otherwise be found, or lose interest in participating; a phenomenon known as *panel attrition*. The cost and difficulty of tracking down respondents for the follow-up survey varies widely.

In light of inevitable panel attrition, researchers must decide how large a sample is required to produce credible findings. A rough estimate is that the baseline survey must be as much as 50% to 75% larger than the minimum statistically valid number needed in the follow-up survey so as to have enough respondents in the desired subgroups.

Panel attrition presents a clear, if not easily resolved, tradeoff. Larger samples cost more, but smaller samples run the risk of not having a statistically significant set of results reflecting the full comparisons needed to assess impact. If there are not the funds to create a sufficiently large sample, then is the survey is probably not worth doing. However, the research team can decide to restrict the goals of its research so that the research itself—in particular the longitudinal study—can be reasonably affordable. Obviously this does not mean omitting investigation of the impacts of central or key program activities, but it may mean restricting how far up and down the value chain the investigation is carried or limiting how fine the distinctions are among subgroups. Researchers should consult a statistician with experience drawing up samples in the relevant country for advice on sampling and sample size. Based on this advice, the program can decide whether sufficient resources exist to carry out the survey.

SPILLOVER IMPACTS

Program activities at the firm and household levels often involve the dissemination of information through training, advice, and other forms of learning. In such cases, there is likely to be spillover impacts as program participants pass useful information and practices on to their friends, relatives, and neighbors. Spillover impacts can also be negative. An example would be if business formation and growth by program clients siphoned sales away from competitors' businesses.

Spillover impacts make impact assessment more difficult, because they blur the distinction between program participants and non-participants. The result is either systematic underestimation or overestimation of true program impacts, depending on whether the spillover impacts are positive or negative. One approach to limit spillover impacts is to locate control sites physically distant from treatment sites, but doing so risks introducing other differences (e.g., climate, soil conditions, access to markets, and infrastructure development) that create selection bias. Another approach to account for spillover impacts is to conduct interviews, focus group discussions, or other qualitative methods with key informants who presumably possess knowledge about potential program spillovers, both positive and negative.

PROGRAM EXPANSION

It is also important to ensure that the program does not expand to control group sites prior to the conclusion of the impact assessment. Expanding the program to control group sites during the assessment period, in effect, converts the control group into the treatment group thereby contaminating the assessment and invalidating the results.

Avoiding this outcome requires that researchers collaborate with program administrators regarding expansion plans and the selection of control group sites.

MINIMALLY ACCEPTABLE METHODOLOGICAL STANDARDS

The more rigorous the research methodologies and the more qualified the research team, the more reliable and credible are the impact assessment findings. This advantage derives from choosing highly trained (often more expensive) experts to design and carry out the impact assessment, setting up a longitudinal survey, choosing a representative sample with appropriate treatment and control groups large enough to account for expected panel attrition and to allow analysis of relevant subgroups, and supporting the quantitative work with qualitative research. The costs of this set of choices are more than justified if the audience includes those who must decide whether the type of program under investigation is worth pursuing on a broader scale or in other locations. In this case, it is crucial to know as much as possible and with as much certainty as possible about program impacts.

Every deviation from these rigorous methodological choices introduces the risk of reducing credibility. Many program evaluations have been done that are not credible impact assessments, although labeled as if they were.[4] While these evaluations may be useful in a variety of ways (e.g., evaluating the effectives of program administration), they cannot demonstrate with any reliability that the program actually caused the observed changes, either with regards to program clients or other actors up and down the value chain.

To ensure credibility of impact assessment findings, it is therefore necessary to establish a set of minimally acceptable methodological standards. To be credible, an impact assessment will satisfy the following minimally acceptable methodological standards:

1. It will include observations on a group of participants (treatment group) and a matched group of non-participants (control group).
2. It will assess the status of both treatment and control group members at a time after impacts can reasonably be expected to have occurred (follow-up).
3. It will be based on a causal (logical) model in which clearly stated hypotheses link program activities to expected impacts.
4. It will be rigorous, in that all methodologies used are well documented and their weaknesses identified.
5. It will use data collection methods that follow accepted good practice.
6. It will use analytical methods that are appropriate, in that they match the type of data collected.
7. If a quasi-experimental methodology is used, it will include data on both treatment and control group members before impact could have occurred (baseline).

Generally, well-done experimental and quasi-experimental impact assessments satisfy all of the minimally acceptable methodological standards. Well-done cross-sectional impact assessments generally satisfy criteria 1- 6, but do less well on criterion 7. Satisfaction of criterion 7 is theoretically possible using retrospective perceptions, but these would need to be carefully constructed and cross-checked using qualitative methodologies. Generally, cross-sectional assessments are recommended only when longitudinal assessments are not possible, and researchers need to be both cognizant and up-front about their limitations.

---

[4] A review of over 70 PSD program evaluations by the PSD Impact Assessment Initiative found that no more than four or five of the evaluations (many self-styled as "impact assessments") satisfied minimally acceptable methodological standards. See Lily Zandniapur, Jennefer Sebstad, and Donald Snodgrass, (2004), Review of Evaluations of Selected Enterprises Development Projects," microREPORT #3, Washington, DC: USAID.

Although not required to satisfy the above criteria, it is further recommended that impact assessments use a mixed methodology approach. Relative to solely quantitative or solely qualitative methods, mixed methods impact assessments provide a broader and more in-depth understanding of program impact and its underlying causal relationships, in addition to a richer understanding of the mechanisms through which these causal relationships operate.

U.S. Agency for International Development
1300 Pennsylvania Avenue, NW
Washington, DC 20523
Tel: (202) 712-0000
Fax: (202) 216-3524
www.usaid.gov