Evaluating the Poverty Impact of Projects: A Handbook for Practitioners

Judy L. Baker
(jbaker2@worldbank.org)

December 15, 1999

LCSPR/PRMPO
The World Bank

Acknowledgements

This handbook was prepared by Judy L. Baker with input from background papers written by the following team members: Gillette Hall (case studies, lessons learned), Julia Lane (case studies, lessons learned), Martin Ravallion (analyst's introduction), Laura Rawlings (implementation issues); and the World Bank PREM Poverty Group work on impact evaluation carried out by Kene Ezemenari, Gloria Rubio, Anders Rudqvist and K. Subbarao. Background research was provided by Matthew Fleming and Samir Stewart. The handbook was jointly supported through LCSPR and PRMPO under the leadership of Norman Hicks, Guillermo Perry, and Michael Walton.

The work has benefited greatly from the comments received by the team members mentioned above, as well as Sabina Alkire, Michael Bamberger, Soniya Carvalho, Wendy Cunningham, Shahidur Khandker, Norbert Schady, and Quentin Wodon.

Table of Contents

List of Boxes

List of Tables

List of Annexes

Foreword

Despite the billions of dollars spent on development assistance each year, there is still very little known about the actual impact of projects on the poor. There is broad evidence on the benefits of economic growth, investments in human capital, and the provision of safety nets for the poor. But for a specific program or project in a given country, is the intervention producing the intended benefits and what was the overall impact on the population? Could the program or project be better designed to achieve the intended outcomes? Are resources being spent efficiently? These are the types of questions that can only be answered through an impact evaluation, an approach which measures the outcomes of a program intervention in isolation of other possible factors.

Many Governments, institutions, and project managers are reluctant to carry out impact evaluations because they are deemed to be expensive, time consuming, technically complex, and because the findings can be politically sensitive, particularly if they are negative. Many evaluations have also been criticized because the results come too late, do not answer the right questions, or were not carried out with sufficient analytical rigor. A further constraint is often the limited availability and quality of data.

Yet with proper and early planning, the support of policy makers, and a relatively small investment compared to overall project cost, a rigorous evaluation can be very powerful in assessing the appropriateness and effectiveness of programs. Evaluating impact is particularly critical in developing countries where resources are scarce and every dollar spent should aim to maximize its' impact on poverty reduction. If programs are poorly designed, do not reach their intended beneficiaries, or are wasteful, with the right information they can be redesigned, improved, or eliminated if deemed necessary. The knowledge gained from impact evaluation studies will also provide critical input to the appropriate design of future programs and projects.

This handbook seeks to provide project managers and policy analysts with the tools needed for evaluating project impact. It is aimed at readers with a general knowledge of statistics. For some of the more in-depth statistical methods discussed, the reader is referred to the technical literature on the topic. Chapter 1 presents an overview of concepts and methods, Chapter 2 discusses key steps and related issues to consider in implementation, Chapter 3 illustrates various analytical techniques through a case study, and Chapter 4 includes a discussion of lessons learned from a rich set of 'good practice' evaluations of poverty projects which have been reviewed for this handbook. The case studies, included in Annex I, were selected from a range of evaluations carried out by the World Bank, other donor agencies, research institutions, and private consulting firms. They were chosen for their methodological rigor, attempting to cover a broad mix of country settings, types of projects, and evaluation methodologies. Also included in the Annexes are samples of the main components that would be necessary in planning any impact evaluation – sample terms of reference, a budget, impact indicators, a log frame, and a matrix of analysis.

While the techniques used in impact evaluation are similar across sectors and population subgroups, the illustrations of methodologies and case examples in the handbook focus on assessing the impact of projects targeted to the poor. Poverty impact can include a wide range of projects and evaluation questions such as measuring the impact of micro-finance programs on household income, the impact of a training program on employment, the impact of a school feeding program on student attendance, or the impact of the construction of rural roads on household welfare.

Regardless of the project type or questions being addressed, the design of each impact evaluation will be unique depending on factors such as the type of data available, local capacity, and timing and budget concerns. Finally, evaluations which will yield high quality, credible and generalizable results for policy makers will require strong financial and political support, early and careful planning, participation of stakeholders in the design of the objectives and approach of the study, adequate data, a suitable mix of methodologies including both quantitative and qualitative techniques, the rigorous application of these techniques, and communication between team members throughout the process.

Chapter 1: Defining concepts and techniques for impact evaluation

A comprehensive evaluation is defined in the literature as an evaluation that includes monitoring, process evaluation, cost-benefit evaluation and impact evaluation. Yet each of these components is distinctly different. *Monitoring* will help to assess whether a program is being implemented as was planned. A program monitoring system enables continuous feedback on the status of program implementation, identifying specific problems as they arise. *Process evaluation* is concerned with how the program operates, and focusing on problems in service delivery. *Cost-benefit* or *cost effectiveness* evaluations assess program costs (monetary or non-monetary), in particular their relation to alternative uses of the same resources and to the benefits being produced by the program. And finally, *impact evaluation* is intended to determine more broadly if the program had the desired effects on individuals, households and institutions and if those effects are attributable to the program intervention. Impact evaluations can also explore unintended consequences, whether positive or negative, on beneficiaries. Of particular interest for this handbook is the extent to which project benefits reach the poor, and the impact that these benefits have on their welfare. Some of the questions addressed in impact evaluation include: How did the project affect the beneficiaries? Were any improvements a direct result of the project, or would they have improved anyway? Could program design be modified to improve impact? Were the costs justified?

These questions cannot, however, be simply measured by the outcome of a project. There may be other factors or events that are correlated with the outcomes that are not caused by the project. To ensure methodological rigor, an impact evaluation must estimate the *counterfactual*, that is, what would have happened had the project never taken place or what otherwise would have been true. For example, if a recent graduate of a labor training program becomes employed, is it a direct result of the program or would that individual have found work anyway? To determine the counterfactual, it is necessary to net out the effect of the interventions from other factors – a somewhat complex task. This is accomplished through the use of *comparison* or *control* groups, those who do not participate in a program or receive benefits, which are subsequently compared to the *treatment* group, individuals who do receive the intervention. *Control* groups are selected randomly from the same population as the program participants, whereas the *comparison* group is more simply the group that does not receive the program under investigation. Both the comparison and control group should resemble the treatment group in every way, with the only difference between groups being program participation.

Determining the counterfactual is at the core of evaluation design. This can be accomplished using several methodologies which fall into two broad categories, *experimental* designs (randomized), and *quasi experimental* designs (non randomized). It is, however, quite tricky to net out the program impact from the counterfactual conditions which can be affected by history, selection bias, and contamination. Qualitative and participatory methods can also be used to assess impact, with these techniques often providing critical insights into beneficiaries perspectives, the value of programs to beneficiaries, the processes which may have affected outcomes, and a deeper interpretation of results observed in quantitative analysis. The strengths and weaknesses of each of these methods are discussed in more detail below. As the reader will find, no technique is perfect and thus the evaluator must make decisions about the tradeoffs for each method chosen. Early and careful planning will, however, provide many more methodological options in designing the evaluation.

Experimental designs

Experimental designs, also known as randomization, are generally considered the most robust of the evaluation methodologies. By randomly allocating the intervention among eligible beneficiaries, the assignment process itself creates comparable treatment and control groups that are statistically equivalent to one another, given appropriate sample sizes. This is a very powerful outcome because, in theory, the control groups generated through random assignment serve as a perfect counterfactual, free from the troublesome selection bias issues that exist in all evaluations. The main benefit of this technique is the simplicity in interpreting results – the program impact can be measured by the difference between the means of the samples of the treatment group and the control group on the outcome being evaluated. One example is the Kenya textbooks evaluation where evaluators selected a random allocation of program sites, administered a baseline survey, created control groups, and then administered the treatment, which in this case was the delivery of textbooks. Having control and treatment groups then allowed the evaluators to clearly determine the impact of textbooks on student learning.

While experimental designs are considered the optimum approach to estimating project impact, in practice there are several problems. First, randomization may be unethical due to the denial of benefits or services to otherwise eligible members of the population for the purposes of the study. An extreme example would be the denial of medical treatment which can turn out to be life-saving to some members of population. Second, it can be politically difficult to provide an intervention to one group and not another. Third, the scope of the program may mean that there are no non-treatment groups such as with a project or policy change that is broad in scope – examples include an adjustment loan, or programs administered at a national level. Fourth, individuals in control groups may change certain identifying characteristics during the experiment which could invalidate or contaminate the results. If, for example, people move in and out of a project area, they may move in and out of the treatment or control group. Alternatively people who were denied a program benefit may seek it through alternative sources, or those being offered a program may not take up the intervention. Fifth, it may be difficult to assure that assignment is truly random. An example of this might be administrators who exclude high risk applicants to achieve better results. And finally, experimental designs can be expensive and time consuming in certain situations, particularly in the collection of new data.

With careful planning, some of these problems can be addressed in the implementation of experimental designs. One way is with the random selection of beneficiaries. This can be used to provide both a politically transparent allocation mechanism and the basis of a sound evaluation design, as budget or information constraints often make it impossible to accurately identify and reach the most eligible beneficiaries. A second is bringing control groups into the program at a later stage once the evaluation has been designed and initiated. Using this technique, the random selection determines *when* the eligible beneficiary receives the program, not *if* they receive it. This was done in the evaluation of a nutrition program in Colombia which provided the additional benefit of addressing questions regarding the necessary time involved for the program to become effective in reducing malnutrition (McKay, 1978). Finally, randomization can be applied within a sub-set of equally-eligible beneficiaries, while reaching all of the most eligible and denying benefits to the least eligible, as was done with education projects in the El Chaco region for the Bolivia social fund evaluation (Pradhan, Rawlings and Ridder, 1998). However, if this latter suggestion is implemented, one must keep in mind that the results produced from the evaluation will only be applicable to the group from which the randomly-generated sample was selected.

Quasi-experimental designs

Quasi-experimental (non-random) methods can be used to carry out an evaluation when it is not possible to construct treatment and comparison groups through experimental design. These techniques generate comparison groups which resemble the treatment group, at least in observed characteristics, through econometric methodologies which include: matching methods, double difference methods, instrumental variables methods, and reflexive comparisons (see Box 1.2). Using these techniques, the treatment and comparison groups are usually selected *after* the intervention using non-random methods. Therefore, statistical controls must be applied to address differences between the treatment and comparison groups and/or sophisticated matching techniques must be used to construct a comparison group that is as similar as possible to the treatment group. In some cases, a comparison group is also chosen before the treatment though the selection is not randomized.

The main benefit of quasi-experimental designs is that they can draw on existing data sources and are thus often quicker and cheaper to implement, and can be performed after a program has been implemented, given sufficient existing data. The principle disadvantages of quasi-experimental techniques are that (i) the reliability of the results is often reduced as the methodology is less robust statistically; (ii) the methods can be statistically complex; and (iii) there is a problem of selection bias. When generating a comparison group rather than randomly assigning one, there are many factors which can affect the reliability of results. Statistical complexity requires considerable expertise in the design of the evaluation, and analysis and interpretation of the results. This may not always be possible, particularly in some developing country circumstances.

The third problem of *bias* relates to the extent to which a program is participated in differentially by subgroups of a target population, thus affecting the sample, and ultimately the results. There are two types of bias, those due to differences in *observables* or something in the data, and those due to differences in *unobservables*, (not in the data) often called *selection bias* (Box 1.1). An observable bias could include the selection criteria through which an individual is targeted such as geographic location, school attendance, or participation in the labor market. Unobservables which may bias program outcomes could include individual ability, willingness to work, or family connections, and a subjective (often politically driven) selection process of individuals into a program. Both types of biases can yield inaccurate results, including under and over-estimates of actual program impacts, negative impacts when actual program impacts are positive (and vice-versa), and statistically insignificant impacts when actual program impacts are significant and vice-versa.[1] It is possible to control for bias through statistical techniques such as matching and instrumental variables, but it is very difficult to fully remove them and remains a major challenge for researchers in the field of impact analysis.

Box 1.1: The Problem of Selection Bias

Selection bias relates to unobservables which may bias outcomes (e.g. individual ability, pre-existing conditions, etc.) Randomized experiments solve the problem of selection bias by generating an experimental control group of people that would have participated in a program but who were randomly denied access to the program or treatment. The random assignment does not remove selection bias, but instead balances the bias between the participant and non-participant samples. In quasi-experimental designs, statistical models (e.g. matching, double differences, instrumental variables) approaches this by modeling the selection processes to arrive at an unbiased estimate using nonexperimental data. The general idea is to compare program participants and nonparticipants holding selection processes constant. The validity of this model depends on how well the model is specified.

A good example is the wages of women. The data represents women who choose to work. If this decision were made, we could ignore the fact that not all wages are observed and use ordinary regression to estimate a wage model. Yet the decision by women to work is not made randomly – women who would have low wages may be unlikely to chose to work because there personal reservation wage is greater than the wage offered by employers. Thus the sample of observed wages for women would be biased upwards.

This can be corrected for if there are some variables that strongly affect the chances for observation (the reservation wage), but not the outcome under study (the offer wage). Such a variable might be the number of children at home.

Source: Greene, 1997.

Among quasi-experimental design techniques, matched comparison techniques are generally considered a second-best alternative to experimental design. The majority of the literature on evaluation methodology is centered around the use of this type of evaluation reflecting both the frequency of use of matched comparisons and the many challenges posed by having less-than-ideal comparison groups. In recent years there

[1] See, for example, LaLonde, 1986; Fraker and Maynard, 1987; LaLonde and Maynard, 1987; Friedlander and Robins, 1995).

have been substantial advances in *propensity score matching* techniques (Rosenbaum and Rubin, 1985; Jalan and Ravallion, 1998). This method is very appealing to evaluators with time constraints and working without the benefit of baseline data given that it can be used with a single cross-section of data. This technique, is, however, dependent on having the right data as it relies on oversampling program beneficiaries during the fielding of a larger survey and then 'matching' them to a comparison group selected from the larger core sample of the overall effort, often a national household survey. Given the growth in the applications of large surveys in developing countries, such as the multipurpose Living Standards Measurement Studies, this evaluation method seems particularly promising. A good example is the evaluation of a public works program, Trabajar, in Argentina (Jalan and Ravallion, 1998, Annex 1.1 and Chapter 4).

Box 1.2: Summary of quantitative methods for evaluating program impact

The main methods for impact evaluation are discussed below. As no method is perfect, it is always desirable to triangulate.

Experimental or Randomized Control Designs

- *Randomization*, in which the selection into the treatment and control groups is random within some well-defined set of people. In this case there should be no difference (in expectation) between the two groups besides the fact that the treatment group had access to the program. (There can still be differences due to sampling error; the larger the size of the treatment and control samples the less the error.)

Non-Experimental or Quasi-Experimental Designs

- *Matching methods or constructed controls*, in which one tries to pick an ideal comparison that matches the treatment group from a larger survey. The most widely used type of matching is *Propensity Score Matching*, where the comparison group is matched to the treatment group on the basis of a set of observed characteristics, or using the "propensity score" (predicted probability of participation given observed characteristics); the closer the propensity score, the better the match. A good comparison group comes from the same economic environment and was administered the same questionnaire by similarly trained interviewers as the treatment group.
- *Double difference* or *difference-in-differences* methods, in which one compares a treatment and comparison group (first difference), before and after a program (second difference). Comparators should be dropped in cases where propensity scores are used and if they have scores outside the range observed for the treatment group.
- *Instrumental variables or statistical control* methods, in which one uses one or more variables which matter to participation, but not to outcomes given participation. This identifies the exogenous variation in outcomes attributable to the program – recognizing that its placement is not random but purposive. The "instrumental variables" are first used to predict program participation, then one sees how the outcome indicator varies with the predicted values.
- *Reflexive comparisons*, in which a "baseline" survey of participants is done before the intervention, and a follow-up survey done after. The baseline provides the comparison group, and impact is measured by the change in outcome indicators before and after the intervention.

Qualitative Methods

Qualitative techniques are also used for carrying out impact evaluation with the intent to determine impact by the reliance on something other than the counterfactual to make a causal inference (Mohr, 1995). The focus instead is on understanding processes, behaviors, and conditions as they are perceived by the individuals or groups being studied (Valadez and Bamberger, 1994). For example, qualitative methods and particularly participant observation can provide insight into the ways in which households and local communities perceive a project and how they are affected by it. Because measuring the counterfactual is at the core of impact analysis techniques, qualitative designs have generally been used in conjunction with other evaluation techniques. The qualitative approach uses relatively open-ended methods during design, collection of data, and analysis.[2] Qualitative data can also be quantified. Among the methodologies used in qualitative impact assessments are the techniques developed for rapid rural assessment which rely on participants knowledge of the conditions surrounding the project or program being evaluated, or participatory evaluations where stakeholders are involved in all stages of the evaluation, determining the objectives of the study, identifying and selecting indicators to be used, and participating in data collection and analysis.[3]

The benefits of qualitative assessments are that they are flexible, can be specifically tailored to the needs of the evaluation using open-ended approaches, they can be carried out quickly using rapid techniques, and can greatly enhance the findings of an impact evaluation through providing a better understanding of stakeholders’ perceptions and priorities, and the conditions and processes which may have affected program impact.

Among the main drawbacks are the subjectivity involved in data collection, the lack of a comparison group, and the lack of statistical robustness given mainly small sample sizes making it difficult to generalize to a larger, representative population. The validity and reliability of qualitative data are very dependent on the methodological skill, sensitivity, and training of the evaluator. If field staff are not sensitive to specific social and cultural norms and practices, and non-verbal messages, the data collected may be misinterpreted. And finally, without a comparison group, it is impossible to determine the counterfactual and thus causality of project impact.

Integrating Quantitative and Qualitative Methods

While there is an extensive literature on quantitative versus qualitative methods in impact evaluation, there is also a growing acceptance on the need for integrating the two approaches. Impact evaluations using quantitative data from statistically representative samples are better suited to assessing causality using econometric methods or reaching

[3] For a detailed discussion on participatory methods see World Bank, 1996, The World Bank Participation Sourcebook.

generalizable conclusions. However, qualitative methods allow the in-depth study of selected issues, cases or events and can provide critical insights into beneficiaries' perspectives, the dynamics of a particular reform or the reasons behind certain results observed in a quantitative analysis. There are significant tradeoffs in selecting one technique over another.

Integrating quantitative and qualitative evaluations can often be the best vehicle for meeting the project's information needs. In combining the two approaches, qualitative methods can be used to inform the key impact evaluation questions, survey questionnaire or the stratification of the quantitative sample, and analysis of the social, economic and political context within which a project takes place, while quantitative methods can be used to inform qualitative data collection strategies, to design the sample to inform the extent to which the results observed in the qualitative work can be generalized to a larger population using a statistically representative sample, and , statistical analysis can be used to control for household characteristics and the socio-economic conditions of different study areas, thereby eliminating alternative explanations of the observed outcomes.

There are several benefits of using integrated approaches in research discussed in Bamberger (1999), which also apply to impact evaluations. Among them:

- *Consistency checks can be built in through the use of triangulation procedures that permit two or more independent estimates to be made for key variables* (such as income, opinions about projects, reasons for using or not using public services, specific impact of a project, etc.).
- *Different perspectives can be obtained.* For example, while researchers may consider income or consumption to be the key indicators of household welfare, case studies may reveal that women are more concerned about vulnerability (defined as the lack of access to social support systems in times of crises), powerlessness, or exposure to violence.
- *Analysis can be conducted on different levels.* Survey methods can provide good estimates of individual, household, and community-level welfare, but they are much less effective for analyzing social processes (social conflict, reasons for using or not using services, and so on) or for institutional analysis (how effectively health, education, credit, and other services operate, and how they are perceived by the community). There are many qualitative methods designed to analyze issues such as social process, institutional behavior, social structure, and conflict.
- *Opportunities can be provided for feedback to help interpret findings.* Survey reports frequently include references to apparent inconsistencies in findings, or to interesting differences between communities or groups which cannot be explained by the data. In most quantitative research, once the data collection phase is completed it is not possible to return to the field to check on such questions. The greater flexibility of qualitative research means that it is often possible to return to the field to gather additional data. Survey researchers also use qualitative methods to check on

outliers—responses that diverge from the general patterns. In many cases the data analyst has to make an arbitrary decision as to whether a household or community that reports conditions that are significantly above or below the norm should be excluded (on the assumption that it reflects a reporting error) or the figures adjusted. Qualitative methods permit a rapid follow-up in the field to check on these cases.

In practice, the integration of quantitative and qualitative methods should be carried out during each step of the impact evaluation. Chapter 2 mentions many opportunities for doing this. For illustration, the Nicaragua School Autonomy Reform Case provides a good example of integrated methods. Quantitative methods following a quasi-experimental design were used to determine the relationship between decentralized management and learning, and to generalize results for different types of schools. In addition, qualitative techniques including a series of key informant interviews and focus group discussions with different school-based staff and parents, were utilized to analyze the context in which the reform was introduced, examine the decision-making dynamics in each school, and assess the perspectives of different school community actors on the autonomy process (see Annex 1.11)

Other Approaches to Impact Evaluation

Two other topics are particularly relevant to the discussion of evaluating the poverty impact of projects: approaches to measuring the impact of structural adjustment programs, and theory based evaluations. Both incorporate many of the methodologies discussed above, though use a different approach.

Evaluating Structural Adjustment Programs. There has been substantial debate on the impact of structural adjustment programs on the poor. Much of the evidence used to support this debate is, however, based on deficient assumptions and methods. As with other projects, the policy changes under structural adjustment projects must be i) compared with relevant counterfactuals that would respond to the same macroeconomic constraints, and ii) analyzed in the context of the local economic structure and based on empirical information from household surveys. This, however, is very difficult for three reasons. First, policy changes may have economy-wide impact making it impossible to find comparison groups which are unaffected. Second, because of exogenous factors, lags, feedbacks, and substitutions, any changes in the well-being of the poor must be interpreted with extreme caution. And third, it is difficult to predict what would have happened if adjustment had not taken place – what alternative policies a government might have pursued, and what the resulting impact would have been on the poor.

In the literature, several approaches have been used, each with their own shortcomings. The techniques are in many cases similar to those described in Box 1.2, though as seen below, estimating the counterfactual requires vast assumptions which may substantially affect the validity of the results. This is most viably handled by isolating specific policy changes which would affect the population, such as exchange rate policies, trade policies, reductions in public expenditures, and reductions in public sector employment. Yet even with this approach, it can be difficult to isolate the impact of specific policies. For examples, see Killick (1995), Poppele, et. al (1999), Bourguignon, et. al (1991), and Sahn, et. al, (1996).

Box 1.3: Summary of methods used to evaluate adjustment policies

Approaches with no counterfactual

- *Qualitative* studies which assess conditions of the population (often identifying vulnerable subgroups), before, during and after adjustment policies are implemented through focus groups, interviews, and other qualitative techniques.

- *Before and After*, which compares the performance of key variables during and after a program with those prior to the program. The approach uses statistical methods to evaluate whether there is a significant change in some essential variables over time. This approach often gives biased results because it assumes that had it not been for the program, the performance indicators would have taken their pre-crisis period values.

Approaches which generate a counterfactual using multiple assumptions

- *Computable General Equilibrium Models* (CGE) which attempt to contrast outcomes in treatment and comparison groups through simulations. These models seek to trace the operation of the real economy, and are generally based on detailed social accounting matrices (SAMs) collected from data on national accounts, household expenditure surveys, and other survey data. CGE models do produce outcomes for the counterfactual, though the strength of the model is entirely dependent on the validity of the assumptions. This can be problematic as data bases are often incomplete and many of the parameters have not been estimated by formal econometric methods. CGE models are also very time consuming, cumbersome, and expensive to generate.

- *With and without comparisons,* which compare the behavior in key variables in a sample of program countries to their behavior in non-program countries (a comparison group). This is an approach to the counterfactual question, using the experiences of the comparison group as a proxy for what would otherwise have happened in the program countries. It is, however, quite difficult to achieve a true comparison group. The method assumes that only the adoption of an adjustment program distinguishes a program country from the comparison group and that the external environment affects both groups the same.

- *Statistical Controls* consist of regressions that control for the differences in initial conditions and policies undertaken in program and non-program countries. The approach identifies the differences between program and non-program countries in the pre-program period, and then controls these differences statistically to identify the isolated impacts of the programs in the post-reform performance.

Theory Based Evaluation. The premise of theory based evaluations is that programs and projects are based on explicit or implicit theory about how and why a program will work. The evaluation would then be based on assessing each theory and assumptions about a program during implementation rather than at a midpoint or after the project has been completed. In designing the evaluation, the underlying theory is presented as many microsteps, with the methods then constructed for data collection and analysis to track the unfolding of assumptions. If events do not work out as expected, the evaluation can say with a certain confidence where, why, and how the breakdown occurred.

The approach puts emphasis on the responses of people to program activities. Theories direct the evaluator's attention to likely types of near-term and longer-term effects. Among the advantages are first, that the evaluation provides early indications of program effectiveness during project implementation. If there are break downs during implementation, it is possible to fix it along the way. Second, the approach helps to explain how and why effects occurred. If events work out as expected, the evaluation can say with a certain confidence how the effects were generated. By following the sequence of stages, it is possible to track the microsteps that led from program inputs through to outcomes.

The shortcomings of the approach are similar to many of the other methodologies. In particular, i) identifying assumptions and theories can be inherently complex; ii) evaluators may have problems in measuring each step unless the right instruments and data are available, iii) problems in testing the effort because theory statements may be too general and loosely constructed to allow for clear-cut testing; and iv) there may be problems of interpretation making it difficult to generalize from results (see Weiss).

An example of theory-based technique is being piloted by the Operations and Evaluation Department of the World Bank to evaluate the impact of social investment funds on community-level decision making processes, traditional power structures and relationships, and community capacity, trust, and well-being. This will be based on the theory that priority groups can effectively implement a project and operate and maintain the investment created by the project. A set of main assumptions, and sub assumptions has been set out and will then tested using existing household survey data, as well as a specially designed survey instrument for a smaller sample, and focus groups and other PRA techniques. The information from each of these data sources will be triangulated in the analysis.

Cost Benefit or Cost Effectiveness Analysis

While this type of analysis is not strictly concerned with measuring impact, it enables policymakers to measure program efficiency by comparing alternative interventions on the basis of the cost of producing a given output. It can greatly enhance the policy implications of the impact evaluation and therefore should be also included in the design of any impact evaluation.[4]

Cost Benefit analysis attempts to measure the economic efficiency of program costs versus program benefits, in monetary terms. For many projects, especially in the social sectors, it is not possible to measure all the benefits in monetary terms. For example, the benefits of a program to provide school inputs (textbooks, classroom furniture, preschool programs) would be increased learning. Instead of measuring monetary outcomes, learning achievement scores could be used to quantify the benefits. This would require *cost-effectiveness* analysis. The concepts for both types of analysis are the same.

The main steps of cost benefit and cost effectiveness analysis are to identify all project costs, benefits, and then compute a cost/effectiveness ratio. In calculating costs, the value of the intervention itself should be included, as well as all other costs such as administration, delivery, investment costs (discounted to the net present value), the monetary value of freely provided goods or services, the social costs such as environmental deterioration, health hazards, etc. Benefits can be monetary such as gain in income, or the number of units delivered, test scores, or health improvements. When benefits cannot be quantified, it is possible to use subjective indicators such as ranking or weighting system. This approach, however, can be tricky in interpreting subjective scores.

Once the costs and benefits have been determined, the cost-effectiveness ratio (R) is then : R=Cost/Unit (or benefit). This ratio can then be compared across interventions to measure efficiency. In theory, this technique is quite straightforward. In practice, however, there are many caveats involved in identifying and quantifying the costs and benefits. It is important to ensure that appropriate indicators are selected, that the methodologies and economic assumptions used are consistent across ratios, and that the ratios are indeed comparable. And as with other techniques used in impact analysis, measuring cost effectiveness can be best carried out when included in the evaluation design from the earliest stages. This allows for the collection of the necessary cost and benefit information and ensuring consistency.

Choosing a Methodology

Given the variation in project types, evaluation questions, data availability, cost, time constraints, and country circumstances, each impact evaluation study will be different and will require some combination of appropriate methodologies, both quantitative and qualitative. The evaluator must carefully explore the methodological options in designing the study, with an aim to produce the most robust results possible. Among quantitative methods, experimental designs are considered the optimal approach with matched comparisons a second-best alternative. Other techniques, however, can

[4] For a more complete discussion of cost benefit and cost effectiveness analysis see “Handbook on Economic Analysis of Investment Operations, World Bank, 1996.

also produce reliable results, particularly with a good evaluation design and high quality data.

The evidence from the 'best practice' evaluations reviewed for this handbook highlights that the choice of impact evaluation methodologies is not mutually exclusive. Indeed, stronger evaluations often combine methods to ensure robustness and to provide for contingencies in implementation. Joining a 'with and without' approach with a 'before and after' approach that uses baseline and follow-up data is one combination strongly recommended from a methodological perspective (Subbarao et al 1999). Having baseline data available will allow evaluators to verify the integrity of treatment and comparison groups, assess targeting and prepare for a robust impact evaluation. This is true even for randomized control designs. Although randomization ensures equivalent treatment and comparison groups at the time of randomization, this feature should not influence evaluators into thinking that they do not need baseline data. Indeed, baseline data may be crucial to reconstructing why certain events took place and controlling for these events in the impact assessment.

Incorporating cost benefit or cost-effectiveness analysis is also strongly recommended. This methodology can enable policymakers to compare alternative interventions on the basis of the cost of producing a given output. This is particularly important in the developing country context where resources are extremely limited.

Finally, combining quantitative and qualitative methods is the ideal, as it will provide the quantifiable impact of a project, as well as an explanation of the processes and interventions that yielded these outcomes. While each impact evaluation will have unique characteristics requiring different methodological approaches, a few general qualities of a best practice impact evaluation include :

- An estimate of the counterfactual has been made by:
  - ➢ using random-assignment to create a control group (experimental design);
  - ➢ appropriately and carefully using other methods such as matching to create a comparison group (quasi-experimental design).

- To control for pre and post-program differences in participants, and to establish program impacts, there are relevant data collected at:
  - ➢ baseline; and
  - ➢ follow-up (including sufficient timeframe to allow for program impacts).

- The treatment and comparison groups are of sufficient sizes to establish statistical inferences with minimal attrition.

- Cost benefit or cost-effectiveness analysis is included to measure project efficiency.

- Qualitative techniques are incorporated to allow for the triangulation of findings.

Chapter 2: Key steps in designing and implementing impact evaluations[5]

Undertaking an impact evaluation study can be quite challenging and costly with implementation issues arising at every step of the way. These challenges highlight the importance of a well-designed study, a committed and highly qualified team, and good communication between the evaluation team members. By incorporating the evaluation early into the design of a project, it will be possible to obtain results in a timely way so that the findings can be used for mid-project adjustments of specific components.

Regardless of the size, program type, or methodology used for the evaluation, there are several key steps to be carried out as outlined below (Box 2.1). This chapter will provide a discussion of these steps, as well as a discussion of the many issues that may arise in implementation. The sequencing of these steps is critical, particularly in ensuring the collection of necessary data before the project begins implementation. Early planning provides the opportunity to randomize, to construct ex-ante matched comparisons, to collect baseline data, and to identify upcoming surveys that could be used in a propensity score matching approach.

All of the design work and initial data collection should be done during project identification and preparation. Ideally, some results will be available during the course of project implementation so they can feed into improving the project design if necessary. A good example of how a project incorporated evaluation plans from the earliest stages is illustrated in the Uganda Nutrition and Early Childhood project (see Chapter 4).

Determining whether or not to carry out an evaluation

A first determination is whether or not an impact evaluation is required. As discussed above, impact evaluations differ from other evaluations in that they are focused on assessing causality. Given the complexity and cost in carrying out impact evaluation, the costs and benefits should be assessed, as well as consideration if another approach would be more appropriate such as monitoring of key performance indicators or a process evaluation.[6] And perhaps the most important inputs to the decision of whether or not to carry out an evaluation are having strong political and financial support.

The additional effort and resources required for conducting impact evaluations are best mobilized when the project is innovative, replicable, involves substantial resource allocations, and has well-defined interventions. For example, the impact evaluation of the Bolivian Social Investment Fund met each of these criteria. First, the new social fund model introduced in Bolivia was considered innovative and replicable; second, the social

[5] This chapter draws heavily on a paper prepared by Laura Rawlings, *Implementation Issues in Impact Evaluation*, Draft, 1999.

[6] These approaches should not be seen as substitutes for impact evaluations, indeed they often form critical complements to impact evaluations.

fund has been responsible for roughly 25 percent of all public investments in Bolivia since the beginning of the evaluation; and third because the interventions were well-defined by the social fund menu of sub-projects.

Box 2.1 Main steps in designing and implementing impact evaluations

During Project identification/preparation

1. Determining whether or not to carry out an evaluation
2. Clarifying objectives of the evaluation
3. Exploring data availability
4. Designing the evaluation
5. Forming the evaluation team

6. If data will be collected:
   - i. Sample Design and Selection
   - ii. Data Collection Instrument Development
   - iii. Staffing and Training Fieldwork Personnel
   - iv. Pilot Testing
   - v. Data Collection
   - vi. Data Management and Access

During Project Implementation

7. Ongoing data collection
8. Analyzing the data
9. Writing up the findings and discussing them with policymakers and other stakeholders
10. Incorporating the findings in project design

Impact evaluations should also be prioritized if the project in question is launching a new approach such as a pilot program which will later be under consideration for expansion based on the results of the evaluation, or the new World Bank Learning and Innovation Loans. This rationale made the Nicaraguan school autonomy reform a good candidate for an impact evaluation. The evaluation study accompanied the government’s testing of a new decentralized school management model from its pilot stage in the mid-1990s through its expansion to almost all secondary schools and about half of all primary schools today. The evaluation was managed by a closely coordinated international team including local staff from the Ministry of Education’s research and evaluation unit and the World Bank’s Primary Education Project coordination office in Managua. Their involvement assured that the evaluation informed key policy decisions regarding the modification and expansion of the pilot.

Another important consideration is to ensure that the program that is to be evaluated is sufficiently developed to be subject to an impact evaluation. Pilot projects and nascent reforms are often prone to revisions regarding their content, as well as how, when and by whom they will be implemented. These changes can undermine the coherence of the evaluation effort, particularly experimental designs and other types of prospective evaluations that rely on baseline and follow-up data of clearly established treatment and control groups. Under circumstances where the policies to be evaluated are still being defined, it may be advisable to avoid using an impact evaluation in order to allow for flexibility in the project.

Gaining support from policy makers and financiers for an impact evaluation can be challenging, but is a prerequisite for proceeding. They must be convinced that the evaluation is a useful exercise addressing questions that will be relevant to decisions concerning the evaluated program's refinement, expansion or curtailment. They must also be convinced of the legitimacy of the evaluation design and therefore the results, particularly when the results are not as positive as anticipated.

Financing for an impact evaluation remains a difficult issue for program managers and client counterparts alike. The financing issue is compounded by the fact that data on evaluation costs are usually difficult to obtain. And perhaps the stickiest issue arises from the public good value of the evaluation – if the results of the evaluation are going to be used to inform policies applied outside of the national boundaries within which the evaluation is conducted, as is often the case, why should an individual country bear the cost of the evaluation? Among the case studies which had information on sources of funding, the information show that countries often assume the majority, but not the entirety, of the evaluation costs. As is discussed more fully in Chapter 4, many of the cases reviewed suggest that successfully implementing an impact evaluation requires not only a substantial resource commitment from the client countries, but also the involvement of World Bank staff, or external researchers and consultants, necessitating resources beyond those provided by the country.

Clarifying evaluation objectives

Once it has been determined that an impact evaluation is appropriate and justified, establishing clear objectives and agreement on the core issues that will be the focus of the evaluation up front will contribute greatly to its success. Clear objectives are essential to identifying information needs, setting output and impact indicators, and constructing a solid evaluation strategy to provide answers to the questions posed. The use of a logical (log) framework approach provides a good and commonly-used tool for identifying the goals of the project and the information needs around which the evaluation can be constructed.

The log frame, increasingly used at the World Bank, is based on a simple four by four matrix which matches information on project objectives, with how performance will be tracked using milestones and work schedules, what impact project outputs will have on a beneficiary institution or system and how that will be measured, and how inputs are used to deliver outputs (see Annex 5 for examples). In other words, it is assumed that the project's intended impact is a function of the project's outputs, as well as a series of other factors. The outputs, in turn, are a function f the project's inputs and factors outside the project. Quantifiable measures should then be identified for each link in the project cycle. This approach does not preclude the evaluator from also looking at the unintended impacts of a project, but serves to keep the objectives of the evaluation clear and focused. Qualitative techniques are also useful in eliciting participation in the clarification of objectives of the evaluation and resulting impact indicators.

Although a statement of the objective would seem on the face of it, to be one of the easiest parts of the evaluation process, it can be extremely difficult. For example, statements that are too broad do not lend themselves to evaluation. The objective statement in the Mexico PROBECAT evaluation (Annex 1.9) that the evaluation is about "the effect of the PROBECAT training program on labor market outcomes." would be more precise if it were narrowed down to the effect of PROBECAT on hours worked, hourly earnings, monthly salary and time to first job placement for different types of workers. The Mexico PROGRESA evaluation provides a good example of how creating a clear outline and delineating multiple objectives by creating from the start a clear outline and a separate discussion of each component – with objectives detailed in sub categories (Annex 1.10). This was particularly important because the intervention was quite complex, with the evaluation having to address not only the program impact, but also aspects of program operations - targeting and timing.

Reviewing other evaluation components such as cost-effectiveness or process evaluations may also be important objectives of a study and can complement the impact evaluation. Cost-effectiveness may be of particular concern for policymakers whose decision it will be to curtail, expand or reform the intervention being evaluated. On issues related to service delivery, a process evaluation may be relevant to assess the procedures, dynamics, norms and constraints under which a particular program is carried out.

Exploring data availability

Many types of data can be used to carry out impact evaluation studies. These can include a range from cross sectional or panel surveys to qualitative open-ended interviews. Ideally this information is available at the individual level to ensure that true impact can be assessed. Household level information can conceal intrahousehold resource allocation, which affects women and children, because they often have more limited access to household productive resources. In many cases, the impact evaluation will take advantage of some kind of existing data or piggy-back on an ongoing survey which can save considerably on costs. With this approach, however, problems may arise in the timing of the data collection effort and with the flexibility of the questionnaire design. Box 2.4 highlights some key points to remember in exploring the use of existing data resources for the impact evaluation.

With some creativity, it may be possible to maximize existing information resources. A good example is with the evaluation of the Honduran Social Investment Fund (see Chapter 4). This study used a module from the national income and expenditure survey in the social fund questionnaire thereby allowing social fund beneficiaries' income to be compared to national measures to assess poverty targeting (Walker et al 1999).

At the most basic level, data on the universe of the population of interest will be required as a basis from which to determine sample sizes, construct the sampling frame and select the sample. Other types of data which may be available in a given country and can be used for different impact evaluations include[7]:

- Household income and expenditure surveys
- Living Standards Measurement Studies (LSMS)
- Labor market surveys
- Records of cooperatives, credit unions and other financial institutions
- School records on attendance, repetition, examination performance
- Public health records on infant mortality, incidence of different infectious diseases, number of women seeking advice on contraception, or condom consumption
- Specialized surveys conducted by universities, NGOs, consulting groups
- Monitoring data from program administrators
- Project case studies

Using Existing Survey Data. Many surveys may also be in the planning stages or are ongoing. If a survey measuring the required indicators is planned, the evaluation may be able to oversample the population of interest during the course of the general survey (for example to use for the propensity score matching approach) as was done for the Nicaraguan Social Investment Fund evaluation and the Argentine Trabajar workfare program evaluation (Jalan and Ravallion 1998). Conversely, if a survey is planned that will cover the population of interest, the evaluation may be able to introduce a question or series of questions as part of the survey, or add a qualitative survey to supplement the quantitative information. For example, the Credit with Education program in Ghana, included a set of qualitative interviews with key stakeholders as well as with non-participant and participant focus groups which provided qualitative confirmation of the quantitative results (Annex 1.6). The evaluation assessed the impact of the program on the nutritional status and food security of poor households. Quantitative data included specific questions on household income and expenditure, and skills level, while qualitative data focused on women's empowerment – status and decision making in the household, social networks, self-confidence, etc.

[7] See Valadez and Bamberger

Box 2.2: Key points for identifying data resources for impact evaluation

- Know the program well. It is risky to embark on an evaluation without knowing a lot about the administrative/institutional details of the program; that information typically comes from the program administration.
- Collect information on the relevant "stylized facts" about the setting. The relevant facts might include the poverty map, the way the labor market works, the major ethnic divisions, other relevant public programs, etc.
- Be eclectic about data. Sources can embrace both informal, unstructured, interviews with participants in the program as well as quantitative data from representative samples. However, it is extremely difficult to ask counter-factual questions in interviews or focus groups; try asking someone who is currently participating in a public program: "what would you be doing now if this program did not exist?" Talking to program participants can be valuable, but it is unlikely to provide a credible evaluation on its own.
- Ensure that there is data on the outcome indicators and relevant explanatory variables. The latter needs to deal with heterogeneity in outcomes conditional on program participation. Outcomes can differ depending on whether one is educated, say. It may not be possible to see the impact of the program unless one controls for that heterogeneity.
- Depending on the methods used, data might also be needed on variables that influence participation but do not influence outcomes given participation. These instrumental variables can be valuable in sorting out the likely causal effects of non-random programs (Box 1.2).
- The data on outcomes and other relevant explanatory variables can be either quantitative or qualitative. But it has to be possible to organize the information in some sort of systematic data structure. A simple and common example is that one has values of various variables including one or more outcome indicators for various observation units (individuals, households, firms, communities).
- The variables one has data on and the observation units one uses are often chosen as part of the evaluation method. These choices should be anchored to the prior knowledge about the program (its objectives of course, but also how it is run) and the setting in which it is introduced.
- The specific source of the data on outcomes and their determinants, including program participation, typically comes from survey data of some sort. The observation unit could be the household, firm, geographic area, depending on the type of program one is studying.
- Survey data can often be supplemented with useful other data on the program (such as from the project monitoring data base) or setting (such as from geographic data bases).

Designing the Evaluation

Once the objectives and data resources are clear, it is possible to begin the design phase of the impact evaluation study. The choice of methodologies will depend on the evaluation question, timing, budget constraints, and implementation capacity. The pros and cons of the different design types discussed in Chapter 1 should be balanced to determine which methodologies are most appropriate and how quantitative and qualitative techniques can be integrated to complement each other.

Even after the evaluation design has been determined and built into the project, evaluators should be prepared to be flexible and make modifications to the design as the project is implemented. In addition, provisions should be made for tracking the project interventions if the evaluation includes baseline and follow-up data so that the evaluation effort is parallel with the actual pace of the project.

In defining the design, it is also important to determine how the impact evaluation will fit into the broader monitoring and evaluation strategy applied to a project. All projects must be monitored so that administrators, lenders and policymakers can keep track of the project as it unfolds. The evaluation effort, as argued above, must be tailored to the information requirements of the project.

Evaluation question. The evaluation questions being asked are very much linked to the design of the evaluation in terms of the type of data collected, unit of analysis, methodologies used, and timing of the various stages. For example, in assessing the impact of textbooks on learning outcomes, it would be necessary to tailor the evaluation to measuring impact on students, classrooms and teachers during a given school year. This would be very different than measuring the impact of services provided through social fund investments which would require data on community facilities and households. The case studies in Annex I provide the other examples of how the evaluation question can affect the evaluation design.

In clarifying the evaluation questions, it is also important to consider the gender implications of project impact. At the outset this may not always be obvious, however, in project implementation there may be secondary effects on the household, which would not necessarily be captured without specific data collection and analysis efforts.

Timing and budget concerns. The most critical timing issue is whether it is possible to begin the evaluation design before the project is implemented and when the results will be needed. It is also useful to identify up front, at which points during the project cycle information from the evaluation effort will be needed so that data collection and analysis activities can be linked. Having results in a timely manner can be crucial to policy decisions such as during a project review, around an election period, or when decisions regarding project continuation are being made.

Some methods require more time to implement than others. Random assignment and before and after methods (e.g. reflexive comparisons) methods take longer to implement than ex-post matched comparison approaches. When using before and after approaches that utilize baseline and follow-up assessments, time must allowed for the last member of the treatment group to receive the intervention, and then usually, more time is allowed for post-program effects to materialize and be observed. Grossman (1994) suggests that twelve to eighteen months after sample enrollment into the intervention is a typical period to allow before examining impacts. In World Bank projects with baselines, waiting for both the intervention to take place and the outcomes to materialize can take years. For example, in the evaluation of the Bolivian Social Investment Fund which relied on baseline data collected in 1993, follow-up data was not collected until 1998 because of the time needed for the interventions (water and sanitation projects, health clinics, and schools) to be carried out and for effects on the beneficiary population’s health and education outcomes to take place. A similar period of time has been required for the evaluation of a primary education project in Pakistan which used an experimental design with baseline and follow-up surveys to assess the impact of community schools on student outcomes, including academic achievement.

The timing requirements of the evaluation cannot drive the project being evaluated. By their very nature, evaluations are subject to the time frame established by the rest of the project. Evaluations must wait on projects that are slow to disburse and generate interventions. And even if projects move forward at the established pace, some interventions take longer to carry out, such as infrastructure projects. The time frame for the evaluation is also sensitive to the indicators selected since many take longer to manifest themselves in the beneficiary population, such as changes in fertility rates or educational achievement.

Implementation Capacity. A final consideration in the scale and complexity of the evaluation design is the implementation capacity of the evaluation team. Implementation issues can be very challenging, particularly in developing countries where there is little experience with applied research and program evaluations. The composition of the evaluation team is very important, as well as their experience with different types of methodologies, and their capacity vis-à-vis other activities being carried out by the evaluation unit. This is particularly relevant when working with public sector agencies with multiple responsibilities and limited staff. Awareness of the unit’s workload is important in order to assess not only how it will affect the quality of evaluation being conducted, but also in terms of the opportunity cost of the evaluation with respect to other efforts for which the unit is responsible. There are several examples of evaluation efforts which were derailed when key staff were called onto other projects and thus were not able to implement the collection of data on schedule at the critical point in time (such as a point during the school year, agricultural season, etc.). Such situations can be avoided through coordination with managers in the unit responsible for the evaluation to ensure that a balance is achieved with respect to the timing of various activities, as well as the distribution of staff and resources across these activities. Alternatively, it can be preferable to contract a private firm to carry out the evaluation (discussed below).

Formation of the evaluation team

There is a range of skills needed in evaluation work. The quality and eventual utility of the impact evaluation can be greatly enhanced with coordination between team members and policy makers from the outset. It is therefore important to identify team members as early as possible, agree upon roles and responsibilities, and establish mechanisms for communication during key points of the evaluation.

Among the core team is the evaluation manager, analysts (both economist and other social scientist), and for evaluation designs involving new data collection, a sampling expert, survey designer, fieldwork manager and fieldwork team, and data managers and processors.[8] Depending on the size, scope and design of the study, some of these responsibilities will be shared or other staffing needs may be added to this core team. In some cases where policy analysts may not have had experience integrating quantitative and qualitative approaches, it may be necessary to spend additional time at the initial team building stage to sensitize team members and ensure full collaboration. The broad responsibilities of team members include:

- *Evaluation manager* - responsible for establishing the information needs and indicators for the evaluation (that are often established with the client using a logical framework approach), drafting terms of reference for the evaluation, selecting the evaluation methodology, and identifying the evaluation team. In many cases, the evaluation manager will also carry out policy analysis.
- *Policy analysts* – an economist is needed for the quantitative analysis, as well as a sociologist or anthropologist for ensuring participatory input and qualitative analysis at different stages of the impact evaluation. Both should be involved in writing the evaluation report.
- *Sampling expert* - can guide the sample selection process. For quantitative data, the sampling expert should be able to carry out power calculations to determine the appropriate sample sizes for the indicators established, select the sample, review the results of the actual sample versus the designed sample, and incorporate the sampling weights for the analysis. For qualitative data, the sampling expert should guide the sample selection process in coordination with the analysts, ensuring that the procedures established guarantee that the correct informants are selected. The sampling expert should also be tasked with selecting sites and groups for the pilot test and will often need to be paired with a local information coordinator responsible for collecting data for the sampling expert from which the sample will be drawn.
- *Survey designer* – this could be a person or team, whose responsibility is designing the data collection instruments, accompanying manuals and codebooks, and coordinating with the evaluation manager(s) to ensure that the data collection instruments will indeed produce the data required for the analysis. This person/team should also be involved in pilot testing and refining the questionnaires.
- *Fieldwork manager/staff* – the manager should be responsible for supervising the entire data collection effort, from planning the routes for the data collection to forming and scheduling the fieldwork teams generally composed of supervisors and interviewers. Supervisors generally manage the fieldwork staff (usually interviewers, data entry operators and drivers) and are responsible for the quality of data collected in the field. Interviewers administer the questionnaires. In some cultures, it is necessary to ensure that male and female interviewers carry out the surveys and that they are administered separately for men and women.
- *Data managers and processors* - design of the data entry programs, enter the data, check its validity, provide the needed data documentation and produce basic results that can be verified by the data analysts.

[8] For a comprehensive guide to designing and implementing surveys, please see Grosh and Muñoz, 1996.

In building up the evaluation team, there are also some important decisions that the evaluation manager must make about local capacity, and the appropriate institutional arrangements to ensure impartiality and quality in the evaluation results. First, is whether there is local capacity to implement the evaluation, or parts of it, and what kind of supervision and outside assistance will be needed. Evaluation capacity varies greatly from country to country and although international contracts that allow for firms in one country to carry out evaluations in another country are becoming more common[9], the general practice for World Bank-supported projects seems to be to implement the evaluation using local staff while providing a great deal of international supervision. Therefore, it is necessary to critically assess local capacity and determine who will be responsible for what aspects of the evaluation effort. Regardless of the final composition of the team, it is important to designate an evaluation manager who will be able to work effectively with the data producers as well as the analysts and policymakers using the data and the results of the evaluation. If this person is not based locally, it is recommended that a local manager be designated to coordinate the evaluation effort in conjunction with the international manager.

Second is whether to work with a private firm or public agency. Private firms can be more dependable with respect to providing results on a timely basis, but capacity building in the public sector is lost and often private firms are understandably less amenable to incorporating elements into the evaluation that will make the effort costlier. Whichever counterpart or combination of counterparts is finally crafted, a sound review of potential collaborators' past evaluation activities is essential to making an informed choice.

And third, is what degree of institutional separation to put in place between the evaluation providers and the evaluation users. There is much to be gained from the objectivity provided by having the evaluation carried out independently of the institution responsible for the project being evaluated. However, evaluations can often have multiple goals, including building evaluation capacity within government agencies and sensitizing program operators to the realities of their projects once these are carried out in the field. At a minimum, the evaluation users who can range from policymakers in government agencies in client countries to non-governmental organizations, bilateral donors and international development institutions, must remain involved enough in the evaluation to ensure that the evaluation process is recognized as being legitimate and that the results produced are relevant to their information needs. Otherwise, the evaluation results are less likely to be used to inform policy. In the final analysis, the evaluation manager and his or her clients must achieve the right balance between involving the users of evaluation and maintaining objectivity and legitimacy of the results.

Data Development

Having adequate and reliable data is a necessary input to evaluating project impact. High quality data is essential to the validity of the evaluation results. As

[9] One example is the Progresa evaluation being carried out by the International Food and Policy Research Institute's (IFPRI).

discussed above, assessing what data exist is a first important step before launching any new data collection efforts. Table 2.1 links the basic evaluation methodologies with data requirements. Most of these methodologies can incorporate qualitative and participatory techniques in the design of the survey instrument, in the identification of indicators, and in input to the identification of controls, variables used for matching or in instrumental variables.

Table 2.1: Evaluation methods and corresponding data requirements

Method	*Data Requirement*		*Use of Qualitative Approach*
	Minimal	*Ideal*	
Experimental or randomized controls	Single project cross-section with and without beneficiaries	Baseline and follow-up surveys on both beneficiaries and non-beneficiaries. Allows for control of contemporaneous events, in addition to providing control for measuring impact. (This allows for a difference in difference estimation)	• Inform design of survey instrument, sampling • Identify indicators • Data collection and recording using: ➢ Textual data ➢ Informal or semi-structured interviews ➢ Focus groups or community meetings ➢ Direct observation ➢ Participatory methods ➢ Photographs • Triangulation • Data Analysis
Non-Experimental Designs			
a) Constructed controls or matching	Large survey, census, national budget or LSMS type survey—that over samples beneficiaries	Large survey, and smaller project-based household survey, both with two points in time to control for contemporaneous events.	
b) Reflexive comparisons and double difference	Baseline and follow-up on beneficiaries	Time series or panel on beneficiaries and comparable non-beneficiaries	
c) Statistical control or instrumental variable	Cross-section data representative of beneficiary population with corresponding instrumental variables	Cross-section and time series representative of both the beneficiary and non-beneficiary population with corresponding instrumental variables.	

Source: Adapted from Ezemenari, Rudqvist, and Subbarao, 1999 and Bamberger, 1999.

For evaluations which will generate their own data, there are the critical steps of designing the data collection instruments, sampling, fieldwork, data management and data access. This section does not outline the step by step process of how to undertake a survey, but rather provides a brief discussion of these steps. Some of the discussion in this section, notably regarding sampling and data management, is more relevant to evaluations based on the collection and analysis of larger-scale sample surveys using quantitative data than for evaluations using qualitative data and small sample sizes.

Deciding What to Measure. The main output and impact indicators should be established when planning the evaluation, possibly as part of a logical framework approach. To ensure that the evaluation is able to assess outcomes during a period of time relevant to decision-makers' needs, a hierarchy of indicators might be established, ranging from short-term impact indicators such as school attendance to longer-term indicators such as student achievement. This ensures that even if final impacts are not picked up initially, program outputs can be assessed. In addition, the evaluator should plan on measuring the delivery of intervention as well as taking account of exogenous factors that may have an effect on the outcome of interest.

Evaluation managers can also plan to conduct the evaluation across several time periods, allowing for more immediate impacts to be picked up earlier, while still tracking final outcome measures. This was done in the Nicaragua School Reform evaluation where the shorter-term impact of the reform on parental participation and student and teacher attendance were established, and the longer-term impacts on student achievement are still being assessed.

Information on the characteristics of the beneficiary population not strictly related to the impact evaluation but of interest in the analysis might also be considered, such as their level of poverty or their opinion of the program. In addition, the evaluator may also want to include cost measures in order to do some cost-effectiveness analysis or other complementary assessments not strictly related to the impact evaluation.

The type of evaluation design selected for the impact evaluation will also carry data requirements. These will be specific to the methodology, population of interest, impact measures and other elements of the evaluation. For example if an instrumental variable approach (one of the types of matched comparison strategies) is to be used, the variable(s) that will serve as the instrument to separate program participation from the outcome measures must be identified and included in the data collection. This was done for the Bolivian Social Investment Fund impact evaluation where knowledge of the social fund and the presence of NGO's (non-governmental organizations) were used as instrumental variables in assessing the impact of social fund interventions.

It can be useful to develop a matrix for the evaluation, listing the question of interest, the outcome indicators that will be used to assess the results, the variable, and the source of data for the variable. This matrix can then be used to review questionnaires and plan the analytical work as was done in the evaluation of the Nicaragua Social Investment Fund (see Annex 6).

Developing Data Collection Instruments and Approaches. Developing appropriate data collection instruments that will generate the required data to answer the evaluation questions can be tricky. This will require having the analysts involved in the development of the questions, in the pilot test and in the review of the data from the pilot test. Involving both the field manager and the data manager during the development of the instruments, as well as local staff, preferably analysts who can provide knowledge of the country and the program can be critical to the quality of information collected (Grosh and Muñoz, 1996). It is also important to ensure that the data collected can be disaggregated by gender to explore the differential impact of specific programs and policies.

Quantitative evaluations usually collect and record information either in a numeric form, or as pre-coded categories. With qualitative evaluations, information is generally presented as descriptive text with little or no categorization. The information may include an individuals' responses to open-ended interview questions, notes taken during focus groups, or the evaluator's observations of events. Some qualitative studies use the pre-coded classification of data as well (Bamberger, forthcoming). The range of data collection instruments and their strengths and weaknesses are summarized in Table 2.2, with the most commonly used technique being questionnaires.

The responses to survey questionnaires can be very sensitive to design, thus it is important to ensure that the structure and format are appropriate, preferably undertaken by experienced staff. For example, the utility of quantitative data has often been severely handicapped for simple mechanical reasons such as the inability to link data from one source to another, as was the case in a national education assessment in one country where student background data could not be linked to test score results making it impossible to assess the influence of student characteristics on performance or to classify the tests scores by students' age, gender, socio-economic status or educational history.

For both qualitative and quantitative data collection, even experienced staff must be trained to collect the data specific to the evaluation and all data collection should be guided by a set of manuals that can be used as orientation during training and a reference during the fieldwork. Depending on the complexity of the data collection task, the case examples show that training can range from three days to several weeks.

Pilot testing is an essential step, as it will reveal whether the instrument can reliably produce the required data and how the data collection procedures can be put into operation. The pilot test should mimic the actual fieldwork as closely as possible. For this reason, it is useful to have data entry programs ready at the time of the pilot to test their functionality as well as to pilot test across the different populations and geographical areas to be included in the actual fieldwork.

Table 2.2 Main Data Collection Instruments for Impact Evaluation

Technique	*Definition and Use*	*Strengths*	*Weaknesses*
Case studies	Collecting information that results in a story which can be descriptive or explanatory and serve to answer the questions how and why.	-Can deal with a full variety of evidence from documents, interviews, observation -Can add explanatory power when focusing on institutions, processes, programs, decisions and events	-Good case studies are difficult to do -Require specialized research and writing skills to be rigorous -Findings not generalizable to population -Time-consuming -Difficult to replicate

Focus Groups	Holding focused discussions with members of target population who are familiar with pertinent issues before writing a set of structured questions. The purpose is to compare the beneficiaries perspectives with abstract concepts in the evaluation's objectives.	-Similar advantages to interviews (below) -Particularly useful where participant interaction is desired -A useful way of identifying hierarchical influences	-Can be expensive and time consuming -Must be sensitive to mixing of hierarchical levels -Not generalizable
Interviews	The interviewer asks questions to one or more persons, and records the respondents' answers. Interviews may be formal or informal, face to face or by telephone, or closed- or open- ended.	-People and institutions can explain their experiences in their own words and setting -Flexible to allow the interviewer to pursue unanticipated lines of inquiry and to probe into issues in-depth -Particularly useful where language difficulties are anticipated -Greater likelihood of getting input from senior officials	-Time consuming -Can be expensive -If not done properly, the interviewer can influence interviewees' response
Observation	Observing and recording situation in a log or diary. This includes who is involved, what happens, when, where, and how events occur. Observation can be direct (observer watches and records), or participatory (the observer becomes part of the setting for a period of time).	-Provides descriptive information on context and observed changes	-Quality and usefulness of data highly depend on the observer's observational and writing skills -Findings can be open to interpretation -Does not easily apply within a short time frame to process change
Questionnaires	Developing a set of survey questions whose answers can be coded consistently.	-Can reach a wide sample, simultaneously -Allows respondents time to think before they answer -Can be answered anonymously -Impose uniformity by asking all respondents the same things -Make data compilation and comparison easier	-The quality of responses is highly dependent on the clarity of questions -If sent, sometimes difficult to persuade people to complete and return questionnaire -Can involve forcing institutional activities and people's experiences into predetermined categories.
Written document analysis	Reviewing documents such as records, administrative data bases, training materials, and correspondence.	-Can identify issues to investigate further and provide evidence of action, change, and impact to support respondents' perceptions -Can be inexpensive	-Can be time consuming

Source: Adapted from Taschereau, 1998.

Sampling. Sampling is an art best practiced by an experienced sampling specialist. The design need not be complicated, but it should be informed by the sampling specialist's expertise in the determination of an appropriate sampling frames, sizes and selection strategies.[10] The sampling specialist should be incorporated in the evaluation process from the earliest stages to review the available information needed to select the sample and determine whether any enumeration work will be needed which can be time consuming.

As with other parts of the evaluation work, coordination between the sampling specialist and the evaluation team are important. This becomes particularly critical when conducting matched comparisons because the sampling design becomes the basis for the "match" that is at the core of the evaluation design and construction of the counterfactual. In these cases, the sampling specialist must work closely with the evaluation team to develop the criteria that will be applied to match the treatment and comparison groups. For example in the evaluation of the Nicaragua school autonomy reform project, autonomous schools were stratified by type of school, enrollment, length of time in the reform and location and matched to a sample of non-autonomous schools using the same stratifications except length of time in the reform. This can be facilitated by having a team member responsible for the data collection work assist the sampling specialist in obtaining the required information including data on the selected outcome indicators for the power calculations (an estimate of the sample size required to test for statistical significance between two groups), a list of the population of interest for the sample selection, and details on the characteristics of the potential treatment and comparison groups important to the sample selection process.

There are many trade-offs between costs and accuracy in sampling which should be made clear as the sampling framework is being developed. For example, conducting a sample in two or three stages will reduce the costs of both the sampling and the fieldwork, but the sampling errors and therefore the precision of the estimates will be increased.

Once the outcome variables and population(s) of interest have been determined by the evaluation team a first step for the sampling specialist would be to determine the power calculations.[11] Since the power calculation can be performed using only one outcome measure, and evaluations often consider several, some strategic decisions will need to be made regarding which outcome indicator to use when designing the sample.

After developing the sampling strategy and framework, the sampling specialist should also be involved in selecting the sample for the fieldwork and the pilot test to ensure that the pilot is not conducted in an area that will be included in the sample for the fieldwork. Often initial fieldwork will be required as part of the sample selection

[10] The discussion on sampling included here refers primarily to issues related to evaluations that collect quantitative data from larger, statistically representative samples.

[11] See Valdez and Bamberger for a discussion of the power calculation process, p. 382-384.

procedure. For example, an enumeration process will be required if there are no up-to-date maps of units required for the sample (households, schools, etc.) of if a certain population of interest needs to be pre-identified so that it can be selected for the purpose of the evaluation, such as malnourished children.

Once the fieldwork is concluded, the sampling specialist should provide assistance on determining sampling weights to compute the expansion factors and correct for sampling errors and non-response.[12] And finally, the sampling specialist should produce a sampling document detailing the sampling strategy, including: (i) from the sampling design stage: the power calculations using the impact variables; the determination of sampling errors and sizes; the use of stratification to analyze populations of interest; (ii) from the sample selection stage: an outline of the sampling stages and selection procedures; (iii) from the fieldwork stage to prepare for analysis: the relationship between the size of the sample and the population from which it was selected, non-response rates and other information used to inform sampling weights; and any additional information that the analyst would need to inform the use of the evaluation data. This document can be used to maintain the evaluation project records and should be included with the data whenever it is distributed to help guide the analysts in using the evaluation data.

Questionnaires. The design of the questionnaire is important to the validity of the information collected. There are four general types of information required for an impact evaluation (Valadez and Bamberger). These include:

- Classification of nominal data with respondents classified according to whether they are project participants or belonging to the comparison group.
- Exposure to treatment variables recording not only the services and benefits received, but also the frequency, amount, and quality. Assessing quality can be quite difficult.
- Outcome variables to measure the effects of a project. These include immediate products, sustained outputs or the continued delivery of services over a long period, and project impacts such as improved income, employment, etc.
- Intervening variables which are factors that affect participation in a project or the type of impact produced such as individual, household or community characteristics. These variables can be important for exploring biases.

The way in which the question is asked, as well as the ordering of the questions, are also quite important in generating reliable information. A relevant example is the measurement of welfare which would be required for measuring the direct impact of a project on poverty reduction. Asking an individual about their income level would not necessarily yield accurate results on their level of economic well being. As discussed in the literature on welfare measurement, questions on expenditures, household

[12] Grosh and Muñoz (1996) provide a detailed discussion of sampling procedures as part of household survey work. Kish (1965) is considered one of the standard textbooks in the sampling field.

composition, assets, gifts and remittances, and the imputed value of home grown food and owner-occupied housing are generally used to capture the true value of household and individual welfare. The time recall used for expenditure items, or the order in which these questions are asked, can significantly affect the validity of the information collected.

Among the elements noted for a good questionnaire are keeping it short and focused on important questions, ensuring that the instructions and questions are clear, limiting the questions to those needed for the evaluation, including a ‘no opinion’ option for closed questions to ensure reliable data, and using sound procedures to administer the questionnaire which may indeed be different for quantitative and qualitative surveys.

Fieldwork Issues. Working with local staff who have extensive experience in collecting data similar to that needed for the evaluation can greatly facilitate fieldwork operations. These staff can provide not only the required knowledge of the geographical territory to be covered, but their knowledge can also be critical to developing the norms used in locating and approaching informants. Field staff whose expertise is in an area other than the one required for the evaluation effort can present problems, as was the case in an education evaluation in Nicaragua that used a firm specialized in public opinion polling to conduct a school and household survey. The expertise that had allowed this firm to gain an excellent reputation based its accurate prediction of improbable election results was not useful for knowing how to approach school children or merge quantitative data sets. This lack of expertise created substantial survey implementation problems that required weeks of corrective action by a joint team from the Ministry of Education and the World Bank.

The type of staff needed to collect data in the field will vary according to the objectives and focus of the evaluation. For example a quantitative impact evaluation of a nutrition program might require the inclusion of an anthropometrist to collect height for weight measures as part of a survey team whereas the impact evaluation of an educational reform would most likely include staff specialized in the application of achievement tests to measure the impact of the reform on academic achievement. Most quantitative surveys will require at least a survey manager, data manager, field manager, field supervisors, interviewers, data entry operators and drivers. Depending on the qualitative approach used, field staff may be similar with the exception of data entry operators. The skills of the interviewers, however, would be quite different with qualitative interviewers requiring specialized training particularly for focus groups, direct observation, etc.

Three other concerns are useful to remember when planning survey operations. First, it is important to take into consideration temporal events that can affect the operational success of the fieldwork and/or the external validity of the data collected, such as the school year calendar, holidays, rainy seasons, harvest times or migration patterns. Second, it is crucial to pilot test data collection instruments, even if they are adaptations of instruments that have been used previously, both to test the quality of the instrument with respect to producing the required data and to familiarize fieldwork staff with the dynamics of the data collection process. Pilot tests can also serve as a proving ground for the selection of a core team of field staff for carrying out the actual survey. Many experienced data collectors will begin with 10-20 percent more staff in the pilot test than will be used in the actual fieldwork and then select the best performers from the pilot to form the actual data collection teams. Finally, communications are essential to field operations. For example, if local conditions permit their use, field work can be enhanced by providing supervisors with cellular phones so that they can be in touch with the survey manager, field manager and other staff to answer questions and keep them informed of progress.

Data Management and Access. The objectives of a good data management system should be to ensure the timeliness and quality of the evaluation data. Timeliness will depend on having as much integration as possible between data collection and processing so that errors can be verified and corrected prior to the conclusion of fieldwork. The quality of the data can be ensured by applying consistency checks to test the internal validity of the data collected both during and after the data are entered and by making sure that proper documentation is available to the analysts who will be using the data. Documentation should consist of two types of information: (i) information needed to interpret the data including codebooks, data dictionaries, guides to constructed variables, and any needed translations; and (ii) information needed to conduct the analysis which is often included in a basic information document that contains a description of the focus and objective of the evaluation, details on the evaluation methodology, summaries or copies of the data collection instruments, information on the sample, a discussion of the fieldwork, and guidelines for using the data.

It is recommended that the data produced by evaluations be made openly available given the public good value of evaluations and the possible need to do additional follow-up work to assess long-term impacts by a team other than the one that carried out the original evaluation work. To facilitate the data sharing process, at the outset of the evaluation an open data access policy should be agreed upon and signed establishing norms and responsibilities for data distribution. An open data access policy puts an added burden on good data documentation and protecting the confidentiality of the informants.[13]

Analysis, Reporting and Dissemination

As with other stages of the evaluation process, the analysis of the evaluation data, whether quantitative or qualitative, requires collaboration between the analysts, data producers, and policy makers to clarify questions and ensure timely, quality results.

[13] If panel data are collected from the same informants over time by different agencies, the informants will have to be identified to conduct the follow-up work. This requirement should be balanced against the confidentiality norms that generally accompany any social sector research. One possible solution is to make the anonymous unit record data available to all interested analysts, but ask researchers interested in conducting follow-up work to contact the agency in charge of the data in order to obtain the listing of the units in the sample, thereby giving the agency an opportunity to ensure quality control in future work through contact with the researchers seeking to carry it out.

Problems with the cleaning and interpretation of data will almost surely arise during analysis and require input from various team members.

Some of the techniques and challenges of carrying out quantitative analysis based on statistical methods are included in Chapter 3. There are also many techniques for analyzing qualitative data (see Miles and Huberman). While a detailed discussion of these methods is beyond the scope of this handbook, two commonly used methods for impact evaluation are mentioned - *content analysis* and *case analysis* (Taschereau). *Content analysis* is used to analyze data drawn from interviews, observations, and documents. In reviewing the data, the evaluator develops a classification system for the data organizing information based on: i) the evaluation questions for which the information was collected; ii) how the material will be used; and iii) the need for cross-referencing the information. The coding of data can be quite complex and may require many assumptions. Once a classification system has been set up, the analysis phase begins, also a difficult process. This involves looking for patterns in the data, and moving beyond description, toward developing an understanding of program processes, outcomes and impacts. This is best carried out with the involvement of team members. New ethnographic and linguistic computer programs are also now available designed to support the analysis of qualitative data.

Case analysis is based on case studies designed for in-depth study of a particular group or individual. The high level of detail can provide rich information for evaluating project impact. The process of collecting and analyzing the data are carried out simultaneously, as evaluators make observations as they are collecting information. They can then develop and test explanations, and link critical pieces of information.

Whether analyzing the quantitative or qualitative information, a few other general lessons related to the analysis, reporting and dissemination can also be drawn from the case examples in Annex 1.

First, analysis commonly takes longer than anticipated, particularly if the data are not as clean or accessible at the beginning of the analysis, if the analysts are not experienced with the type of evaluation work, or if there is an emphasis on capacity building through collaborative work. In the review of the case studies considered for this article, the most rapid analysis took approximately one year after producing the data and the longer analysis close to two years. The case in Chapter 3 illustrates some of the many steps involved in analysis and why it can take longer than anticipated.

Second, the evaluation manager should plan to produce several products as outputs from the analytical work, keeping in mind two elements. The first is to ensure that the timing of outputs around key events when decisions regarding the future of the project will be made, such as mid-term reviews, elections or closings of a pilot phase. The second is the audience for the results. Products should be differentiated according to the audience for which they are crafted, including government policymakers, program managers, donors, the general public, journalists, and academics.

Third, the products will have the most policy relevance if they include clear and practical recommendations stemming from the impact analysis. These can be broken into short and long-term priorities, and when possible, should include budgetary implications. Decision makers will be prone to look for the ‘bottom line’.

Finally, the reports should be planned as part of a broader dissemination strategy, which can include presentations for various audiences, press releases, feedback to informants, and making information available on the web. Such a dissemination strategy should be included in the initial stages of the planning process to ensure that it is included in the budget and that the results reach the intended audience.

Chapter 3: Applying analytical methods for impact evaluation: A case study[14]

This case study is based on a hypothetical anti-poverty program, PROSCOL, which provides cash transfers targeted to poor families with school-age children in one region of a given developing country. The case is intended to illustrate the analytical steps involved in carrying out an impact evaluation and the options an analyst may face, with the process applicable to any type of anti-poverty program. In exploring how to go about evaluating the impact of the program, the policy analyst makes several common errors along the way, seeking input on specific topics from the specialized skills of colleagues - a statistician, economist, econometrics professor, and sociologist. Among the analytical steps that the analyst goes through in the case are:

Identifying the questions to be addressed in the impact evaluation
Assessing data resources
Taking a first look at the data
Understanding biases
Learning about forgone income
Adding control variables
Understanding the importance of exogeneity
Exploring better ways to form a comparison group – propensity score matching
Learning about biases due to unobservables
Reviewing what could have been done with a baseline survey: double differences
Using instrumental variables
Testing the various methodologies
Incorporating input from the field
Planning for future work

Description of the hypothetical program, PROSCOL

The PROSCOL program identifies families eligible for participation using various poverty proxies which include the number of people in the household, the education of the head, and various attributes of the dwelling. PROSCOL pays a fixed amount per school-age child to all selected households on the condition that the children attend 85 percent of their school classes, which has to be verified by a note from the school. Households must keep their children in school until 18 years of age.

This program was introduced 12 months ago, is financed by the World Bank, and operates out of the Ministry of Social Development. In an effort to assess PROSCOL's impact on poverty in order to help determine whether the program should be expanded to include the rest of the country, or be dropped, the World Bank has requested an impact evaluation by the Ministry of Finance. The request was to the Ministry of Finance so as to

[14] This chapter draws heavily on a background paper by Martin Ravallion, *The Mystery of the Vanishing Benefits: Ms. Speedy Analyst's Introduction to Evaluation*, Policy Research Working Paper No. 2153, 1999.

help assure an independent evaluation, and to help develop capacity for this type of evaluation in a central unit of the government ? close to where the budgetary allocations are being made.

Identifying the questions to be addressed in the impact evaluation

The first step for the analyst in the Ministry of Finance assigned to the task of carrying out the PROSCOL evaluation is to clarify which project objectives will be looked at in evaluating impact. The project has two policy goals: the cash transfers aim to reduce current poverty, and by insisting that transfer recipients keep their kids in school the program aims to reduce future poverty by raising education levels among the current population of poor children. Two pieces of information would therefore be needed about the program to assess impact. First, are the cash transfers mainly going to low-income families? And second, how much is the program increasing school enrollment rates?

Assessing data resources

To carry out the evaluation the analyst has two main resources. The first is a report based on qualitative interviews with program administrators and focus groups of participants. It is not clear, however, whether those interviewed were representative of PROSCOL participants, or how poor they are relative to those who were not picked for the program and were not interviewed. The report says that the children went to school, but it is possible that they may have also gone to school if the program had not existed. While this report is an important start, it does not tell the analyst how poor PROSCOL participants are and what impact the program has on schooling. The second resource is a recent independent national household survey carried out by the country's Bureau of Statistics, called the Living Standards Survey (LSS). The LSS included a random sample of 10,000 households, and asked about household incomes by source, employment, expenditures, health status, education attainments, and demographic and other attributes of the family. The survey had incorporated a question on whether or not the sampled household had participated in PROSCOL, and a line item for money received from PROSCOL in the listing of income sources.

Taking a first look at the data

The analyst then proceeds with obtaining the raw LSS data set to focus on assessing who is benefiting from the program. She uses a statistical software package such as SPSS or SAS to generate a cross-tab of the average amount received from PROSCOL by household deciles, where the deciles are formed by ranking all households in the sample according to their income per person. In calculating the latter, the analyst decides to subtract any monies received from PROSCOL as a good measure of income in the absence of the program with the intent of identifying who gained according to their pre-intervention income.

The cross-tab suggests that the cash transfers under the program are quite well targeted to the poor. By the country's official poverty line, about 30 percent of the population in the Northwest is poor. From the table, calculations show that the poorest 30 percent of the survey sample receive 70 percent of the PROSCOL transfers. At first glance, this appears to be a positive result.

The next question is about the impact on schooling. This is looked at through a cross tabulation with average school enrollment rates of various age groups for PROSCOL families versus non-PROSCOL families. This suggests almost no difference between the two; the average enrollment rate for kids aged 6-18 is about 80 percent in both cases. The analyst then calculates average years of schooling at each age, with the results plotted separately for PROSCOL families and non-PROSCOL families. This shows that the two figures are not identical, but they are very close. At this stage, the analyst wonders if there was really no impact on schooling, or if the approach is wrong.

Understanding biases

With this uncertainty the analyst next seeks input from a senior statistician to explore why the results suggest that PROSCOL children are no more likely to be in school than non-Proscol children. The statistician hypothesizes that the results may have a serious bias. In order to assess program impact, we need to know what would have happened without the program. Yet the analyst has not accounted for this – instead the non-PROSCOL families are used as the comparison group for inferring what the schooling would be of the PROSCOL participants if the program had not existed.

In other words, P_i denotes PROSCOL participation of the i'th child. This can take two possible values, namely P_i =1 if the child participates in PROSCOL and P_i =0 if he/she does not. If the i'th child does not participate, then its level of schooling is S_{0i} which stands for child i's schooling S when P=0. If the child does participate then its schooling is S_{1i}. Its gain in schooling due to PROSCOL is S_{1i}-S_{0i}. The gain for ith child who participates (P=1) is then:

$$G_i = S_{1i} - S_{0i} \mid P_i = 1$$

The '$\mid$' stands for 'given that' or 'conditional on' and is needed to make it clear that the calculation is the gain for a child who actually participated. If one wants to know the average gain, this is simply the mean of all the G's which gives the sample mean gain in schooling amongst all those who participated in PROSCOL. As long as this mean is calculated correctly (using the appropriate sample weights from the survey) it will provide an unbiased estimate of the true mean gain. The latter is the 'expected value' of G, and it can be written as:

$$G = \mathrm{E}(S_{1i} - S_{0i} \mid P_i = 1)$$

This is another way of saying 'mean'. However, it need not be exactly equal to the mean calculated from the sample data, given that there will be some sampling error. In the evaluation literature, $\mathrm{E}(S_{1i}-S_{0i} \mid P_i=1)$ is sometimes called the '*treatment effect' or 'the average treatment effect on the treated*'. In this case, PROSCOL is considered the treatment.

The statistician points out to the analyst that she has not calculated G, but rather the difference in mean schooling between children in PROSCOL families and those in non-PROSCOL families. This is the sample estimate of:

$$D = \mathrm{E}(S_{1i} \mid P_i=1) - \mathrm{E}(S_{0i} \mid P_i=0)$$

There is a simple identity linking the D and G, namely:

$$D = G + B$$

This term ‘B’ is the bias in the estimate, and it is given by:

$$B = \mathrm{E}(S_{0i} \mid P_i=1) - \mathrm{E}(S_{0i} \mid P_i=0)$$

In other words, the bias is the expected difference in schooling without PROSCOL between children who did in fact participate in the program, and those who did not. This bias could be corrected if $\mathrm{E}(S_{0i} \mid P_i=1)$ were known, but it isn’t possible to even get a sample estimate of that. One can’t observe what the schooling would have been of children who actually participated in PROSCOL had they not participated; that is missing data – also called a ‘*counterfactual*’ mean.

This bias presents a major concern. In the absence of the program, PROSCOL parents may well send their children to school less than do other parents. If so, then there will be a bias in the calculation. Going back to the original evaluation questions, we are interested in the extra schooling due to PROSCOL. Presumably this only affects those families who actually participate. In other words, we need to know how much less schooling could be expected without the program. If there is no bias, then the extra schooling under the program is the difference in mean schooling between those who participated and those who did not. Thus the bias arises if there is a difference in mean schooling between PROSCOL parents and non-PROSCOL in the absence of the program.

To eliminate this bias, the best approach would be to assign the program randomly. Then participants and non-participants will have the same expected schooling in the absence of the program, i.e., $\mathrm{E}(S_{0i} \mid P_i=1) = \mathrm{E}(S_{0i} \mid P_i=0)$. The schooling of non-participating families will then correctly reveal the counterfactual, i.e., the schooling that we would have observed for participants had they not had access to the program. Indeed, random assignment will equate the whole distribution, not just the means. There will still be a bias due to sampling error, but for large enough samples one can safely assume that any statistically significant difference in the distribution of schooling between participants and non-participants is due to the program.

Within the existing design of the program, it is clear that participation is not random. Indeed, it would be a serious criticism of PROSCOL to find that it was. The very fact of its purposive targeting to poor families, who are presumably less likely to send their kids to school, would create bias.

This raises the question, if PROSCOL is working well then we should expect participants to have worse schooling in the absence of the program. Then $E(S_{0i} | P_i = 1) < E(S_{0i} | P_i = 0)$ and the analysts' original calculation will underestimate the gain from the program. We may find little or no benefit even though the program is actually working well.

The analyst now realizes that the magnitude of this bias could be huge. Suppose that poor families send their kids to work rather than school; because they are poor and cannot borrow easily, they need the extra cash now. Non-poor families send their kids to school. The program selects poor families, who then send their kids to school. One observes negligible difference in mean schooling between PROSCOL families and non-PROSCOL families; indeed, $E(S_{1i} | P_i = 1) = E(S_{0\,i} | P_i = 0)$ in expectation. But the impact of the program is positive, and is given by $E(S_{0i} | P_i = 0) - E(S_{0i} | P_i = 1)$. The failure to take account of the program's purposive, pro-poor, targeting could well have led to a very substantial under-estimation of PROSCOL's benefits from the analyst's comparison of mean schooling between PROSCOL families and non-PROSCOL families.

Learning about forgone income

The analyst next shows the results of her cross-tab of amounts received from PROSCOL against income to another colleague, an economist in the Ministry of Finance. The economist raises a main concern – that the gains to the poor from PROSCOL have been clearly overestimated because foregone income has been ignored. Children have to go to school if the family is to get the PROSCOL transfer, thus they will not be able to work, either on the family business or in the labor market. For example, children aged 15-18 can earn two-thirds or more of the adult wage in agriculture and construction. PROSCOL families will lose this income from their childrens' work. This foregone income should be taken into account when the net income gains from the program are calculated. And this net income gain should be subtracted, not the gross transfer, to work out pre-intervention income. This will also matter to determining how poor the family would have been is in the absence of the PROSCOL transfer. The current table, therefore, might greatly overstate the program's gains to the poor.

The analyst wonders why she should factor out the foregone income from child labor, assuming that less child labor is a good thing. The economist highlights that she should look at the gains from reducing child labor, of which the main gain is the extra schooling, and hence higher future incomes of currently poor families. The analyst has produced tables which reflect the two main ways PROSCOL reduces poverty: by increasing the current incomes of the poor, and by increasing their future incomes. The impact on child labor matters to both, but in opposite directions, thus PROSCOL faces a trade off.

This highlights why it is important to get a good estimate of the impact on schooling; only then will it be possible to determine the foregone income. It is for example, possible that the extra time at school comes out of non-work time.

With regard to the second cross-tab, the main concern raised by the economist is that there is no allowance for all the other determinants of schooling, besides participation in PROSCOL. The economist suggests running a regression of years of schooling on a set of control variables as well as whether or not the child's family was covered by PROSCOL. For the i'th child in the sample let:

$$S_i = a + bP_i + cX_i + e_i$$

Here a, b and c are parameters, X stands for the control variables, such as age of the child, mother's and father's education, the size and demographic composition of the household and school characteristics, while e is a residual that includes other determinants of schooling, and measurement errors. The estimated value of b gives you the impact of PROSCOL on schooling.

Note that if the family of the i'th child participates in PROSCOL then P=1 and so its schooling will be $a + b + cX_i + e_i$. If it does not participate, then P=0 and so its schooling will be $a + cX_i + e_i$. The difference between the two is the gain in schooling due to the program, which is just b.

Adding control variables

As suggested, the analyst next runs a regression with and without the control variables. When it is run without them, the results show that the estimated value of b is not significantly different from zero (using the standard t-test given by the statistical package). These results look very similar to the first results, taking the difference in means between participants and nonparticipants — suggesting that PROSCOL is not having any impact on schooling. However, when several control variables are included in the regression, there is a positive and significant coefficient on PROSCOL participation. The calculation shows that by 18 years of age, the program has added two years to schooling.

The analyst wonders why these control variables make such a difference? And are the right controls being used? She next visits her former Econometrics Professor and shows him her regressions. His first concern related to the regression of schooling on P and X is that it does not allow the impact of the program to vary with X; the impact is the same for everyone, which does not seem very likely. Parents with more schooling would be more likely to send their children to school, so the gains to them from PROSCOL will be lower. To allow the gains to vary with X, let mean schooling of non-participants be $a_0 + c_0X_i$ while that of participants is $a_1 + c_1X_i$, so the observed level of schooling is:

$$S_i = (a_1 + c_1X_i + e_{1i})P_i + (a_0 + c_0X_i + e_{0i})(1 - P_i)$$

where e_0 and e_1 are random errors, each with means of zero and uncorrelated with X. To estimate this model, it is necessary to add an extra term for the interaction effects between program participation and observed characteristics to the regression already run. Thus the augmented regression is:

$$S_i = a_0 + (a_1 - a_0)P_i + c_0X_i + (c_1 - c_0)P_iX_i + e_i$$

where $e_i = e_{1i}P_i + e_{0i}(1 - P_i)$. Then $(a_1 - a_0) + (c_1 - c_0)X$ is the mean program impact at any given value of X. If the mean X in the sample of participants is used, then it will give the mean gain from the program.

Understanding the importance of exogeneity

A second concern raised by the Econometrics Professor is in how the regression has been estimated. In using the regress command in the statistical package, Ordinary Least Squares (OLS), there is concern because the OLS estimates of the parameters will be biased even in large samples unless the right-hand side variables are exogenous. *Exogeneity* means that the right-hand-side variables are determined independently of schooling choices and so they are uncorrelated with the error term in the schooling regression. Because participation in the program was purposively targeted PROSCOL's participation is not exogenous. This can affect the calculation of the program's impact as follows:

The equation for years of schooling is:

$$S_i = a + bP_i + cX_i + e_i$$

The value of $a + b + cX_i + e_i$ was used as the estimate of the i'th household's schooling when it participates in PROSCOL, while $a + cX_i + e_i$ was used to estimate schooling if it does not participate. Thus the difference, b, is the gain from the program. However, in making this calculation the implicit assumption is that e_i was the same either way. In other words, the assumption was that e was independent of P .which would affect the calculation of the program's impact.

This highlights the bias due to non-random program placement which may also be affecting the estimate based on the regression model suggested earlier by the Economist ($S_i = a + bP_i + cX_i + e_i$). This may not, however, mean that the results will necessarily be completely wrong.

The Econometrics Professor clarifies this with an explicit equation for P, namely:

$$P_i = d + eZ_i + ?_i$$

where Z is several variables that include all the observed *'poverty proxies'* used for PROSCOL targeting. There will also be some purely random error term that influences participation; these are poverty proxies that are not in the data, and there will also have been mistakes in selecting participants that also end up in this $?$ term. This equation is linear, yet P can only take two possible values, 0 and 1. Predicted values between zero and one are acceptable, but a linear model cannot rule out the possibility of negative predicted values, or values over one. There are nonlinear models that can deal with this problem, but to simplify the discussion it will be easiest to confine attention to linear models.

There is a special case in which the above OLS regression of S on P and X will give an unbiased estimate of b. That is when X includes all the variables in Z that also influence schooling, and the error term ? is uncorrelated with the error term e in the regression for schooling. This is sometimes called '*selection on observables*' in the evaluation literature.

Suppose that the control variables X in the earlier regression for schooling include all the observed variables Z that influence participation P and ? is uncorrelated with e (so that the unobserved variables affecting program placement do not influence schooling conditional on X). This has then eliminated any possibility of P being correlated with e. It will now be exogenous in the regression for schooling. In other words, the key idea of selection on observables is that there is some observable X such that the bias vanishes conditional on X.

Adding the control variables to the regression of schooling on PROSCOL participation made a big difference because the X must include variables that were amongst the poverty proxies used for targeting, or were correlated with them, and they are variables that also influenced schooling. This, however, only works if the assumptions are valid. There are two problems to be aware of. Firstly, the above method breaks down if there are no unobserved determinants of participation; in other words if the error term ? has zero variance, and all of the determinants of participation also affect schooling. Then there is no independent variation in program participation to allow one to identify its impact on schooling; it is possible to predict P perfectly from X, and so the regression will not estimate. This problem is unlikely to arise often, given that there are almost always unobserved determinants of program placement.

The second problem is more common, and more worrying in this case. The error term e in the schooling regression probably contains variables that are not found in the survey, but might well influence participation in the program, i.e., they might be correlated with the error term ? in the participation equation. If that is the case then E(e| X, P) ? 0 and ordinary regression methods will still be biased when estimating regressions for schooling. Thus the key issue is the extent of the correlation between the error term in the equation for participation and that in the equation for schooling.

Exploring better ways to form a comparison group – propensity score matching

With further input from the Professor, the analyst learns there are better ways to form a comparison group. The objective is to compare schooling levels conditional on observed characteristics. If the sample groups are divided into groups of families with the same or similar values of X and you then compare the conditional means for PROSCOL and non-PROSCOL families. If schooling in the absence of the program is independent of participation, given X, then the comparison will give an unbiased estimate of PROSCOL's impact. This is sometimes called '*conditional independence*', and it is the key assumption made by all comparison-group methods.

Thus, a better way to select a comparison group, given the existing data, is to use as a control for each participant a non-participant with the same observed characteristics. This could, however, be very hard because the data set could have a lot of those variables. There may be nobody amongst the non-participants with exactly the same values of all the observed characteristics for any one of the PROSCOL participants.

A statistical approach, *propensity score matching*, provides techniques for simplifying the problem greatly. Instead of aiming to assure that the matched control for each participant has exactly the same value of X, the same result can be achieved by matching on the predicted value of P, given X, which is called the propensity score of X. Rosenbaum and Rubin (*Biometrika* 1983) show that if (in this case) schooling without PROSCOL is independent of participation given X then they are also independent of participation given the propensity score of X. Since the propensity score is just one number, it is far easier to control for it than X, which could be many variables. And yet propensity score matching is sufficient to eliminate the bias provided there is conditional independence given X.

In other words, if one first regresses P on X to get the predicted value of P for each possible value of X, which is then estimated for the whole sample. For each participant, one should find the non-participant with the closest value of this predicted probability. The difference in schooling is then the estimated gain from the program for that participant.

One can then take the mean of all those differences to estimate the impact. Or take the mean for different income groups. This, however, requires caution in how the model of participation is estimated. A linear model could give irregular predicted probabilities, above one, or negative. It is better to use the LOGIT command in the statistical package. This assumes that the error term ? in the participation equation has a logistic distribution, and estimates the parameters consistent with that assumption by maximum likelihood methods. This is based on the principals of the maximum likelihood estimation of binary response models.

Another issue to be aware of is that some of the non-participants may have to be excluded as potential matches right from the start. In fact there are some recent results in the literature in econometrics indicating that failure to compare participants and controls at common values of matching variables is a major source of bias in evaluations. (See Heckman et al., 1998).

The intuition is that one wants the comparison group to be as similar as possible to the treatment group in terms of the observables, as summarized by the propensity score. We might find that some of the non-participant sample has a lower propensity score than any of those in the treatment sample. This is sometimes called *'lack of common support'*. In forming the comparison group, one should eliminate those observations from the set of non-participants to assure that you are only comparing gains over the same range of propensity scores. One should also exclude those non-participants for whom the probability of participating is zero. It is advisable to trim a small proportion of the sample, say 2 percent, from the top and bottom of the non-participant distribution in terms of the propensity scores. Once the participants have been identified and non-participants over a common matching region, it is recommended to take an average of (say) the five or so nearest neighbors in terms of the absolute difference in propensity scores.

Box 3.1: Steps in propensity score matching

The aim of matching is to find the closest comparison group from a sample of non-participants to the sample of program participants. "Closest" is measured in terms of observable characteristics. If there are only one or two such characteristics then matching should be easy. But typically there are many potential characteristics. The main steps in matching based on propensity scores are as follows:

Step 1: You need a representative sample survey of eligible non-participants as well as one for the participants. The larger the sample of eligible non-participants the better, to facilitate good matching. If the two samples come from different surveys, then they should be highly comparable surveys (same questionnaire, same interviewers or interviewer training, same survey period and so on).

Step 2: Pool the two samples and estimate a logit model of program participation as a function of all the variables in the data that are likely to determine participation.

Step 3: Create the predicted values of the probability of participation from the logit regression; these are called the "propensity scores". You will have a propensity score for every sampled participant and non-participant.

Step 4: Some of the non-participant sample may have to be excluded at the outset because they have a propensity score which is outside the range (typically too low) found for the treatment sample. The range of propensity scores estimated for the treatment group should correspond closely to that for the retained sub-sample of non-participants. You may also want to restrict potential matches in other ways, depending on the setting. For example, you may want to only allow matches within the same geographic area to help assure that the matches come from the same economic environment.

Step 5: For each individual in the treatment sample, you now want to find the observation in the non-participant sample that has the closest propensity score, as measured by the absolute difference in scores. This is called the "nearest neighbor". You can find the five (say) nearest neighbors.

Step 6: Calculate the mean value of the outcome indicator (or each of the indicators if there is more than one) for the five nearest neighbors. The difference between that mean and the actual value for the treated observation is the estimate of the gain due to the program for that observation.

Step 7: Calculate the mean of these individual gains to obtain the average overall gain. This can be stratified by some variable of interest such as incomes in the non-participant sample.

This is the simplest form of propensity score matching. Complications can arise in practice. For example, if there is over-sampling of participants then you can use choice-based sampling methods to correct for this (Manski and Lerman, 1978); alternatively you can use the odds ratio ($p/(1-p)$, where p is the propensity score) for matching. Instead of relying on the nearest neighbor you can instead use all the non-participants as potential matches but weight them differently, according to how close they are (Heckman et al., 1998).

Next, all the all the variables in the data set that are, or could proxy for, the poverty indicators that were used by MSD in selecting PROSCOL participants should be included. Again, X should include the variables in Z. This, however, brings out a weakness of propensity score matching . With matching, a different X will yield a different estimate of impact. With randomization, the ideal experiment, the results do not depend on what *X* you choose. Nor does randomization require that one specifies a model for participation, whether a logit or something else. Box 3.1 summarizes the steps for doing propensity score matching.

Learning about biases due to unobservables

Even after forming the comparison group, the analyst cannot be sure that this will give a much better estimate of the programs' impact. The methods described above will only eliminate the bias if there is conditional independence, such that the unobservable determinants of schooling—not included in the set of control variables *X*—are uncorrelated with program placement. There are two distinct sources of bias, that due to differences in observables and that due to differences in unobservables; the latter is often called '*selection bias*'. Box 3.2 elaborates on this difference.

Box 3.2: Sources of bias in naïve estimates of PROSCOL's impact

The bias described by the Statistician is the expected difference in schooling without PROSCOL between families selected for the program and those not chosen. This can be broken down into two sources of bias:

- Bias due to differences in observable characteristics. This can come about in two ways. Firstly there may not be common support. The "support" is the set of values of the control variables for which outcomes and program participation are observed. If the support is different between the treatment sample and the comparison group then this will bias the results. In effect, one is not comparing like with like. Secondly, even with common support, the distribution of observable characteristics may be different within the region of common support; in effect the comparison group data is miss-weighted. Careful selection of the comparison group can eliminate this source of bias.
- Bias due to differences in unobservables. The term "selection bias" is sometimes confined solely to this component (though some authors use that term for the total bias in a non-experimental evaluation). This source of bias arises when, for given values of *X*, there is a systematic relationship between program participation and outcomes in the absence of the program. In other words, there are unobserved variables that jointly influence schooling and program participation conditional on the observed variables in the data.

There is nothing to guarantee that these two sources of bias will work in the same direction. So eliminating either one of them on its own does not mean that the total bias is reduced in absolute value. That is an empirical question. In one of the few studies to address this question, the true impact, as measured by a well-designed experiment, was compared to various non-experimental estimates (Heckman et al., 1998). The bias in the naïve estimate was huge, but careful matching of the comparison group based on observables greatly reduced the bias.

Going back to the Professor's last equation shows that conditional independence will hold if *P* is exogenous, for then $E(e_i | X_i, P_i) = 0$. However, endogenous program placement due to purposive targeting based on unobservables will still leave a bias. This is sometimes called '*selection on unobservables*'. Thus the conditions required for justifying the method raised earlier by the economist (section x) are no less restrictive than those needed to justify a version of the first method based on comparing PROSCOL families with non-PROSCOL families for households with similar values of *X*. Both rest on believing that these unobservables are not jointly influencing schooling and program participation, conditional on *X*.

Intuitively, one might think that careful matching reduces the bias, but that is no necessarily so. Matching eliminates part of the bias in your first naïve estimate of PROSCOL's impact. That leaves the bias due to any troublesome unobservables. However, these two sources of bias could be offsetting, one positive the other negative. Heckman et al. (1998) make this point. So the matching estimate could well have more bias than the naïve estimate. One cannot know on *a priori* grounds how much better off one is with even a well chosen comparison group, which is an empirical question.

Reviewing what could have been done with a baseline survey - double difference estimates

The analyst next inquires whether there would be another method besides randomization that is robust to these troublesome unobservables. This would require '*baseline data*' for both the participants and non-participants, collected before PROSCOL started. The idea is that data are collected on outcomes and their determinants both before and after the program is introduced, as well as data for an untreated comparison group as well as the treatment group. It is then possible to just subtract the difference between the schooling of participants and the comparison group before the program is introduced from the difference after the program. This is called the '*double difference*' estimate, or '*difference in differences*'. This will deal with the troublesome unobserved variables provided they do not vary over time.

This can be explained by adding subscripts to the earlier equation so that the schooling after the program is introduced:

$$S_{ia} = a + bP_i + cX_{ia} + \mathrm{e}_{ia}$$

Before the program, in the baseline survey, school attainment is instead:

$$S_{ib} = a + cX_{ib} + \mathrm{e}_{ib}$$

(Of course *P*=0 before the program is introduced.) The error terms include an additive time invariant effect, so we can write them as:

$$\mathrm{e}_{it} = h_i + m_{it} \quad \text{(for } t=a,b\text{)}$$

where h_i is the time invariant effect, which is allowed to be correlated with P_i, and m_{it} is an innovation error, which is not correlated with P_i (or X_i).

The essential idea here is to use the baseline data to reveal those problematic unobservables. Notice that since the baseline survey is for the same households as we have now, the *i*'th household in the equation for S_{ia} is the same household as the *i*'th in the equation for S_{ib}. We can then take the difference between the 'after' equation and the 'before' equation:

$$S_{ia} - S_{ib} = bP_i + c(X_{ia} - X_{ib}) + m_{ia} - m_{ib}$$

It is now possible to regress the change in schooling on program participation and the changes in *X*. OLS will give you an unbiased estimate of the program's impact. The unobservables – the ones correlated with program participation – have been eliminated.

Given this, if the program placement was based only on variables, both observed and unobserved, that were known at the time of the baseline survey then it would be reasonable to assume that the *h*'s do not change between the two surveys. This would hold as long as the problematic unobservables are time invariant, the changes in schooling over time for the comparison group will reveal what would have happened to the treatment group without the program.

This would require knowing the program well, and being able to time the evaluation surveys so as to coordinate with the program. Otherwise there are bound to be unobserved changes after the baseline survey that influence who gets the program. This would create 0's that changed between the two surveys.

This last equation can be interpreted as meaning that the child and household characteristics in *X* are irrelevant to the change in schooling if those characteristics do not change over time. But the gain in schooling may depend on parents' education (and not just any change in their education) and possibly on where the household lives, as this will determine the access to schools. There can also be situations in which the changes over time in the outcome indicator are influenced by the initial conditions. Then one will also want to control for differences in initial conditions. This can be done by simply adding X_a and X_b in the regression separately, so that the regression takes the form:

$$S_{ia} - S_{ib} = bP_i + c_aX_{ia} + c_bX_{ib} + m_{ia} - m_{ib}$$

Even if some (or all) variables in *X* do not vary over time one can still allow *X* to affect the changes over time in schooling.

The propensity-score matching method discussed above can help assure that the comparison group is similar to the treatment group before doing the double difference. In an interesting study of an American employment program, it was found that failure to assure that comparisons were made in a region of common support was a major source of bias in the double difference estimate when compared to a randomized control group. Within the region of common support, however, the bias conditional on *X* did not vary much over time. Thus taking the double difference makes sense, after the matching is done (see Heckman et al., in Econometrica 1998).

In practice, following up on households in surveys can, however, be difficult. When doing the follow-up survey, it may not be easy to find all those households who were originally included in the baseline survey. Some people in the baseline survey may not want to be interviewed again, or they have moved to an unknown location.

If drop outs from the sample are purely random then the follow up survey will still be representative of the same population in the baseline survey. However, if there is some systematic tendency for people with certain characteristics to drop out of the sample then there will be a problem. This is called '*attrition bias'*. For example, PROSCOL might help some poor families move into better housing. And even when participant selection was solely based on information available at or around the baseline date (the time-invariant effect 0_i), selected participants may well drop out voluntarily on the basis of changes after that date. Such attrition from the treatment group will clearly bias a double-difference estimate of the program's impact. Box 3.3 outlines the steps to form a double-difference estimate.

Box 3.3: Doing a double difference

The "*double difference*" method entails comparing a treatment group with a comparison group (as might ideally be determined by the matching method in Box 3.2) both before and after the intervention. The main steps are as follows:

Step 1: You need a "*baseline*" survey before the intervention is in place, and the survey must cover both non-participants and participants. If you do not know who will participate, you have to make an informed guess. Talk to the program administrators.

Step 2: You then need one or more follow-up surveys, after the program is put in place. These should be highly comparable to the baseline surveys (in terms of the questionnaire, the interviewing, etc). Ideally the follow-up surveys should be of the same sampled observations as the baseline survey. If this is not possible then they should be the same geographic clusters, or strata in terms of some other variable.

Step 3: Calculate the mean difference between the 'after' and 'before' values of the outcome indicator for each of the treatment and comparison groups.

Step 4: Calculate the difference between these two mean differences. That is your estimate of the impact of the program.

This is the simplest version of double-difference. You may also want to control for differences in exogenous initial conditions, or changes in exogenous variables, possibly allowing for interaction effects with the program (so that the gain from the intervention is some function of observable variables). A suitable regression model can allow these variations.

Using instrumental variables

Given that there is no baseline survey of the same households to do the double difference method, the Professor recommends another methodology to get an estimate that is robust to the troublesome unobservables – an *instrumental variable*.

An instrumental variable (IV) is the classic solution for the problem of an endogenous regressor. An instrumental variable is an observable source of exogenous variation in program participation. In other words, it is correlated with *P* but is not already in the regression for schooling, and is not correlated with the error term in the schooling equation, e. So one must have to have at least one variable in *Z* that is not in *X*, and is not correlated with e. Then the Instrumental Variables Estimate of the program's impact is obtained by replacing *P* by its predicted value conditional on *Z*. Since this predicted value depends solely on *Z* (which is exogenous) and *Z* is uncorrelated with e, it is now reasonable to apply ordinary least squares to this new regression.

Since the predicted values depend only on the exogenous variation due to the instrumental variable, and the other exogenous variables, the unobservables are no longer troublesome, since they will be uncorrelated with the error term in the schooling regression. This also suggests another, more efficient, way to deal with the problem. Remembering that the source of bias in the earlier estimate of the program's impact was the correlation between the error term in the schooling equation and that in the participation equation. This is what creates the correlation between participation and the error term in the schooling equation. Thus a natural way to get rid of the problem when one has an instrumental variable is to add the residuals from the first stage equation for participation to the equation for schooling, but still keeping actual participation in the schooling regression. However, since we have now added to the schooling regression the estimated value of the error term from the participation equation, it is possible to treat participation as exogenous and run OLS. This only works if there is a valid instrument. If not, the regression will not estimate, since the participation residual will be perfectly predictable from actual participation and *X*, in a linear model.

An instrumental variable can also help if there may be appreciable measurement error in the program participation data, another possible source of bias. Measurement error means that there is the possibility that program participation varies more than it actually does. This overestimation in the variance of *P* leads naturally to an underestimation of its coefficient *b*. This is called 'attenuation bias', because this bias attenuates the estimated regression coefficient.

While an instrumental variable can be extremely useful, in practice, caution is necessary. When the actual participation is just replaced with its predicted value and OLS is run, this will not give the correct standard errors since the computer will not know that previously estimated parameters to obtain the predicted values had to be used. A correction to the OLS standard errors is required, though there are statistical packages that allow one to do this easily, at least for linear models.

If there was a dependent variable, however, that could only take two possible values, at school or not at school say, then one should use nonlinear binary response model, such as Logit or Probit. The principle of testing for exogeneity of program participation is similar in this case. There is a paper by Rivers and Vuong (1988) that discusses the problem for such models; Blundell and Smith (1993) provide a useful overview of various nonlinear models in which there is an endogenous regressor.

Testing the methodologies

When the analyst begins to think about identifying an instrumental variable she realizes that this is not a straightforward process. Every possibility she has come up with could also be put in with the variables in *X*. The problem is finding a valid *"exclusion restriction"*, which justifies putting some variable in the equation for participation, but not in the equation for schooling.

The analyst decides to try the "*propensity score matching method*". The logit model of participation looks quite sensible, and suggests that PROSCOL is well targeted. Virtually all of the variables that one would expect to be associated with poverty have positive, and significant, coefficients. The analyst then does the propensity score matching. On comparing the mean school enrollment rates, the results show that children of the matched comparison group had an enrollment rate of 60 percent, as compared to the figure of 80 percent for PROSCOL families.

To account for the issue of foregone income, the analyst draws on an existing survey of child labor which asked about earnings. (In this developing country, there is an official ban on children working before they are 16 years of age, but the government has a hard time enforcing it; nonetheless, child wages are a sensitive issue.) From this survey, the earnings that a child would have had if they had not gone to school can be determined.

It is then possible to subtract from PROSCOL's cash payment to participants the amount of foregone income, and thus work out the net income transfer. Subtracting this net transfer from total income, it is possible to work out where the PROSCOL participants come from in the distribution of pre-intervention income. They are not quite as poor as first thought (ignoring foregone income) but they are still poor; for example, two-thirds of them are below countries' official poverty line.

Having calculated the net income gain to all participants, it is now possible to calculate the poverty rate with and without PROSCOL. The '*post-intervention*" poverty rate (with the program) is, simply stated, the proportion of the population living in households with an income per person below the poverty line, where "*income*" is the observed income (including the gross transfer receipts from PROSCOL). This can be calculated directly from the household survey. By subtracting the net income gain (cash transfer from PROSCOL minus foregone income from kids' work) attributed to PROSCOL from all the observed incomes, the results show a new distribution of pre-intervention incomes. The poverty rate without the program is then the proportion of people living in poor households, based on this new distribution. The analyst finds that the observed poverty rate in Northwest of 32 percent would have been 36 percent if PROSCOL had not existed. The program allows 4 percent of the population to escape poverty now. The schooling gains mean that there will also be both pecuniary and non-pecuniary gains to the poor in the future. In the process of measuring poverty, the analyst remembers learning that the proportion of people below the poverty line is only a basic measure, as it tells you nothing about changes below the line (see Box 3.4). She then calculates both the poverty gap index and the squared poverty gap index, with the results suggesting that these have also fallen as a result of PROSCOL.

Box 3.4: Poverty measures

The simplest and most common poverty measure is the *headcount index*. In this case, it is the proportion of the population living in households with income per person below the poverty line. (In other countries, it is a consumption-based measure, which has some advantages; for discussion and references see Ravallion, 1994.)

The headcount index does not tell us anything about income distribution below the poverty line: a poor person may be worse off but the headcount index will not change; not will it reflect gains amongst the poor, unless they cross the poverty line.

A widely used alternative to the headcount index is the *poverty gap index* (PG). The poverty gap for each household is the difference between the poverty line and the household's income; for those above the poverty line the gap is zero. When the poverty gap is normalized by the poverty line, and one calculates its mean over all households (whether poor or not), one obtains the poverty gap index.

The poverty gap index will tell you how much impact the program has had on the depth of poverty, but it will not reflect any changes in distribution amongst the poor due to the program. For example, if the program entails a small gain to a poor person who is above the mean income of the poor, at the expense of an equal loss to someone below that mean, then PG will not change.

There are various *"distribution-sensitive"* measures that will reflect such changes in distribution amongst the poor. One such measure is the *squared poverty gap* (Foster et al., 1984). This is calculated the same way as PG except that the individual poverty gaps as a proportion of the poverty line are squared before taking the mean (again over both poor and non-poor.) Another example of a distribution-sensitive poverty measure is the Watts index. This is the mean of the log of the ratio of the poverty line to income, where that ratio is set to one for the non-poor. Atkinson (1987) describes other examples in the literature.

In this calculation, the analyst also recognizes that there is some uncertainty about the country's poverty line. To test the results, she repeats the calculation over a wide range of poverty lines finding that at a poverty line for which 50% of the population are poor based on the observed post-intervention incomes, the proportion would have been 52% without PROSCOL. At a poverty line which 15% fail to reach with the program, the proportion would have been 19% without it. By repeating these calculations over the whole range of incomes, the entire "poverty incidence curves" have been traced, with and without the program. This is also called the "*cumulative distribution function*" (see Box 3.5.)

Box 3.5: Comparing poverty with and without the program

Using the methods described in the main text and earlier boxes one obtains an estimate of the gain to each household. In the simplest evaluations this is just one number. But it is better to allow it to vary with household characteristics. One can then summarize this information in the form of poverty incidence curves (PICs), with and without the program.

Step 1: The post-intervention income (or other welfare indicator) for each household in the whole sample (comprising both participants and non-participants) should already exist; this is data. You also know how many people are in each household. And, of course, you know the total number of people in the sample (N; or this might be the estimated population size, if inverse sampling rates have been used to "expend up" each sample observation).

Step 2: You can plot this information in the form of a PIC. This gives (on the vertical axis) the percentage of the population living in households with an income less than or equal to that value on the horizontal axis. To make this graph, you can start with the poorest household, mark its income on the horizontal axis, and then count up on the vertical axis by 100 times the number of people in that household divided by N. The next point is the proportion living in the two poorest households, and so on. This gives the post-intervention PIC.

Step 3: Now calculate the distribution of income pre-intervention. To get this you subtract the estimated gain for each household from its post-intervention income. You then have a list of post-intervention incomes, one for each sampled household. Then repeat Step 2. You will then have the pre-intervention PIC.

If we think of any given income level on the horizontal axis as a "poverty line" then the difference between the two PICs at that point gives the impact on the headcount index for that poverty line (Box 3.4). Alternatively, looking horizontally gives you the income gain at that percentile. If none of the gains are negative then the post-intervention PIC must lie below the pre-intervention on. Poverty will have fallen no matter what poverty line is used. Indeed, this also holds for a very broad class of poverty measures; see Atkinson (1987). If some gains are negative, then the PICs will intersect. The poverty comparison is then ambiguous; the answer will depend on which poverty lines and which poverty measures one uses. (For further discussion see Ravallion, 1994.) You might then use a priori restrictions on the range of admissible poverty lines. For example, you may be confident that the poverty line does not exceed some maximum value, and if the intersection occurs above that value then the poverty comparison is unambiguous. If the intersection point (and there may be more than one) is below the maximum admissible poverty line then a robust poverty comparison is only possible for a restricted set of poverty measures. To check how restricted the set needs to be, you can calculate the poverty depth curves (PDCs). These are obtained by simply forming the cumulative sum up to each point on the PIC. (So the second point on the PDC is the first point on the PIC plus the second point, and so on.)

If the PDCs do not intersect then the program's impact on poverty is unambiguous as long as one restricts attention to the poverty gap index or any of the distribution sensitive poverty measures described in Box 3.4. If the PDCs intersect then you can calculate the "poverty severity curves" with and without the program, by forming the cumulative sums under the PDCs. If these do not intersect over the range of admissible poverty lines then the impact on any of the distribution-sensitive poverty measures in Box 3.4 is unambiguous.

Incorporating input from the field

In the implementation of every program, there is insight from beneficiaries and program administrators which may or may not be reflected in program data. For example in this case, the perception of those working in the field is that the majority of PROSCOL families are poor and that the program indeed provides assistance. When the analyst discusses this with a sociologist working with the program, she learns of some uncertainty in the reality of forgone income, and the issue of work. The sociologist discusses that in the field, one observes many children from poor families who work as well as go to schools, and that some of the younger children not at school, don't seem to be working. The analyst realizes that this requires some checking - on whether there is any difference in the amount of child labor done by PROSCOL children versus a matched comparison group. This data, however, is not available in the household survey though it would be possible to present the results with and without the deduction for foregone income.

The Sociologist also has noticed that for a poor family to get on PROSCOL it matters a lot which school-board area (SBA) the family lives in. All SBAs get a PROSCOL allocation from the center, even SBAs that have very few poor families. If one is poor but living in a well-to-do SBA, they are more likely to get help from PROSCOL than if you live in a poor SBA. What really matters then, is relative poverty—relative to others in the area you live—which matters much more than the absolute level of living.

This allocation would influence participation in PROSCOL, but one would not expect it to matter to school attendance, which would depend more on one's absolute level of living, family circumstances, and characteristics of the school. Thus the PROSCOL budget allocation across SBA's can be used as instrumental variables to remove the bias in the estimates of program impact.

There is information on which school-board area each household belongs to in the household survey, the rules used by the center in allocating PROSCOL funds across SBAs, and how much the center has allocated to each SBA. Allocations are based on the number of school age children, with an "adjustment factor" for how poor the SBA is thought to be. However, the rule is somewhat vague.

The analyst attempts to take these points into account, and re-runs the regression for schooling, but replacing the actual PROSCOL participation by its predicted value (the propensity score) from the regression for participation, which now includes the budget allocation to the SBS. It helps to already have as many school characteristics as possible in the regression for attendance. Although school characteristics do not appear to matter officially to how PROSCOL resources are allocated, any omitted school characteristics that jointly influence PROSCOL allocations by SBA and individual schooling outcomes will leave a bias in her instrumental variable estimates. While there is always the possibility of bias, with plenty of geographic control variables, this method should at least offer a credible comparator to the matching estimate.

From the results it is determined that the budget allocation to the SBA indeed has a significant positive coefficient in the logit regression for PROSCOL participation. Now (predicted) PROSCOL participation is significant in a regression for school enrolment, in which all the same variables from the logit regression are included, except the SBA budget allocation. The coefficient implies that the enrollment rate is 15 percentage points higher for PROSCOL participants than would have otherwise been the case. A regression is also run for years of schooling, for boys and girls separately. For either boys or girls of 18 years, the results indicate that they would have dropped out of school almost two years earlier if it had not been for PROSCOL. This regression, however, raises questions – if the right standard errors are being used, and if using linear models should be used.

Planning for future work

Finally, the analyst is ready to report the results of the evaluations. They show that PROSCOL is doing quite well and as a result the policy makers show interest in expanding the program. From the process the analyst has gone through in carrying out the evaluation, she has a few important observations:

- impact evaluation can be much more difficult than first anticipated;
- it is possible to come up with a worryingly wide range of estimates, depending on the specifics of the methodology used;
- it is good to use alternative methods in the frequent situations of less than ideal data; though each method has pitfalls; and
- one has to be eclectic about data.

In addition to the lessons the analyst has learned, she has a few key recommendations for future evaluation work of PROSCOL. First, it would be desirable to randomly exclude some eligible PROSCOL families in the rest of the country and then do a follow up survey of both the actual participants and those randomly excluded from participating. This would give a more precise estimate of the benefits. It would, however, be politically sensitive to exclude some. Yet if the program does not have enough resources to cover the whole country in one go, and the program will have to make choices about who gets it first. It would indeed be preferable to make that choice randomly, amongst eligible participants. Alternatively, it would be possible to pick the schools or the school board areas randomly, in the first wave. This would surely make the choice of school or school board area a good instrumental variable for individual program placement.

Second, if this is not feasible, it is advisable to carryout a baseline survey of areas in which there are likely to be high concentrations of PROSCOL participants before the program starts in the South. This could be done at the same time as the next round of the national survey which was used for evaluating the PROSCOL program. It would also be good to add a few questions to the survey, such as whether the children do any paid work.

And third, it would be useful to include qualitative work, to help form hypotheses to be tested and assess the plausibility of key assumptions made in the quantitative analysis.

Chapter 4: Drawing on 'Good Practice' Impact Evaluations [15]

The previous chapters have presented the key methods, issues and challenges that can arise in evaluating project impact. In reviewing the case studies listed in Table 4.1 many illustrative examples emerge from interesting approaches in the design, use of data, choice and application of analytical methods used, and in-country capacity building. These examples are highlighted below as well as a discussion of the costs of evaluations and the political economy issues which may arise in implementation.

The 15 case studies included in the review were chosen from a range of evaluations carried out by the World Bank, other donor agencies, research institutions, and private consulting firms. They were selected as a sample of 'good practice' for their methodological rigor, attempting to reflect a range of examples from different sectors and regions. While each impact evaluation has its' strengths and weaknesses, the lessons learned from these experiences should help the project manager or policy analyst intending to design and implement future work.

Early and careful planning of the evaluation design

Adequate preparation during the beginning stages of project identification will ensure that the right information is collected, and that the findings can be used for mid-course adjustments of project components. With early and careful planning it is possible to incorporate all the elements which contribute to a rigorous impact evaluation such as a baseline survey with a randomized control group, and qualitative data on the processes which may affect impact.

Uganda Nutrition and Child Development Project.[16.] This evaluation, though still not yet under implementation, provides an excellent example of early and careful planning. The project itself focuses on strengthening the ability of parents and communities to care for children by providing them with knowledge on better childcare practices and by enhancing opportunities to increase income. It is community-based, and implemented by a network of Non-Government Organizations NGOs. The evaluation component, which was integrated into the project cycle from day one, approaches the 'ideal' in terms of evaluation design. First, it generates baseline and follow-up survey data, along with a randomized control group, so that the program's impact on beneficiary outcomes can be rigorously assessed. Second, it enhances this quantitative component with a participatory (qualitative) monitoring and evaluation (M&E) process.

[15] This chapter draws on the best practice case studies in Annex I and overview pieces prepared by Gillette Hall and Julia Lane, and the World Bank Poverty Group work on Impact Evaluation prepared by, Kene Ezemenari, Gloria Rubio and Anders Rudqvist, and Khalanidhi. Subbarao.

16 Sources: World Bank, 1998a; Garcia, Alderman and Rudqvist, 1999; and World Bank, 1999g

Table 4.1: Summary table of 'Good Practice' Impact Evaluations

Project:	Country	Database Type	Unit of analysis	Outcome Measures	Econometric Approach					Qualitative	Strengths
					Randomization	Matching	Reflexive Comparisons	Double Difference	Instrumental Variables		
Education											
Radio Nicaragua	Nicaragua	Baseline and post intervention survey	Students and classrooms	Test scores	Yes	No	Yes	No	No	No	Questionnaire design
School Autonomy Reform	Nicaragua	Panel survey and qualitative assessments	Students, parents, teachers, directors	Test scores, degree of local decision-making	No	Yes	Yes	No	No	Yes	Qualitative-Quantitative Mix
Textbooks	Kenya	Baseline and post intervention survey	Students, classrooms, teachers	Test scores	Yes	No	Yes	Yes	Yes	No	Analysis of confounding factors
Dropping out	Philippines	Baseline and post intervention survey	Students, classrooms, teachers	Test scores and dropout status	Yes	No	Yes	Yes	Yes	No	Cost/benefit analysis; capacity building
Labor Programs											
Trabajar	Argentina	Household survey, census, administrative records, social assessments	Workers, households	Income, targeting, costs	No	Yes	No	No	Yes	Yes	Judicious use of existing data sources, innovative analytic techniques
Probecat	Mexico	Retrospective and labor force surveys	Workers	Earnings and employment outcomes	No	Yes	No	No	No	No	Matching technique
Active Labor Programs	Czech Republic	Retrospective mail surveys.	Workers	Earnings and employment outcomes	No	Yes	No	No	No	No	Matching technique
Finance											
Micro Finance	Bangladesh	Post intervention survey plus administrative records	Households	Consumption and education	Yes	Yes	No	Yes	No	No	Analysis of confounding factors

Credit with Education	Ghana	Baseline and post intervention survey	Mother/child pairs	Income, health and empowerment	Yes	Yes	Yes	No	No	Yes	Use of qualitative and quantitative information
Health Financing	Niger	Baseline and post intervention survey plus administrative records	Households and health centers	Cost recovery and access	No	Yes (on districts)	Yes	No	No	No	Use of administrative data
Food and Nutrition											
Food for Education	Bangladesh	Household expenditure survey	Households and communities	School attendance	No	No	No	No	Yes	No	Imaginative use of instruments to address selection problem with standard data
Health, Education and Nutrition	Mexico	Baseline and postintervention surveys	Households	Health, education and nutrition outcomes	Yes	Yes	Yes	Not known	Not Known	No	Clear conceptualization; analysis of confounding factors
Infrastructure											
Social Investment Fund	Bolivia	Baseline and follow-up surveys	Households, projects	Education and health indicators	Yes	Yes	Yes	Yes	Yes	No	Range of evaluation methodologies applied
Rural Roads	Viet Nam	Baseline and follow-up surveys	Households, communities	Welfare indicators at household and commune levels	No	Yes	Yes	Yes	Yes	No	Measures welfare outcomes
Agriculture											
National Extension Project	Kenya	Panel data, beneficiary assessments	Households, farms	Farm productivity and efficiency	No	No	Yes	No	No	No	Policy-relevance of results

On the quantitative side, the project was designed to allow for an experimental study design, in which parishes will be randomly assigned into treatment and control groups. Health cards will then be used to record data on the child's weight in treatment and control parishes. In addition, the baseline household survey will be conducted before services are delivered to the communities, as well as a follow up survey of the same households two years later. A rapid review of these data is expected to inform the decision to scale up some components of the intervention during the mid-term review of the project. A deeper analysis of the data at the end of the project will guide the design of the second phase of the project.

Ghana Credit with Education Project. The evaluation of this project was very complex, with many intermediate steps. The project combines elements of a group lending scheme with education on the basics of health, nutrition, birth timing and spacing and small business skills. The evaluation generally focuses on assessing the nutritional status of children, women's economic capacity to invest in food and health care, women's knowledge and adoption of breast feeding, and weaning. It begins with a very clear conceptual framework which is illustrated below. This schematic clearly delineates the inputs, intermediate benefits and longer term outcomes in a way that both facilitates the development of several models and their interpretation. By carefully planning the evaluation and working with a schematic at an early stage, it was possible to clarify many points in a relatively complex design (see Annex 1.6)

Approaches to evaluation when there is no baseline

In practice, many evaluations do not have adequate data. Evaluations are added after it is possible to do a baseline survey or in the absence of comparison groups. Some examples of this are the Bangladesh Food for Education, Mexico Probecat, Czech Active Labor Programs, and Argentina Trabajar evaluations. Without a baseline, the controls must be constructed using the matching methods discussed in the previous chapters. This can, however, be quite tricky. The propensity score matching technique used in the Argentina Trabajar Project to construct a control group with cross-sectional data on program participants and non-participants provides a good example.

The Trabajar II Project in Argentina. This project was focused on providing employment at low wages in small social and economic infrastructure sub-projects selected by community groups. The impact evaluation of the program was designed to assess whether the incomes of program participants were higher than they would have been had the program not been in place. The most commonly used methods to estimate household income without the intervention were not feasible in the case of the TRABAJAR program: no randomization had taken place to construct a control group to which to compare the income of project beneficiaries; and no baseline survey was available, ruling out the possibility of conducting a before and after evaluation.

The Trabajar evaluation instead used existing data to construct a comparison group by matching program participants to non-participants from the national population over a set of socioeconomic variables such as schooling, gender, housing, subjective perceptions of welfare, membership in political parties and neighborhood associations, using a technique called 'propensity scoring'. The study demonstrates resourceful use of existing national household survey data (the *Encuesta de Desarrollo Social – EDS*) in generating the comparison group, combined with a smaller survey of Trabajar participants conducted specifically for the purposes of the evaluation. The smaller survey was carefully designed so that it used the same questionnaire as the EDS, the same interview teams, and was conducted at approximately the same time in order to successfully conduct the matching exercise. This technique was possible in the Trabajar case because a national household survey was being canvassed and the evaluators could take advantage of this survey to oversample Trabajar participants. The same interview teams were used for both the national and project surveys, resulting in efficiency gains in data collection (see Annex 1.1).

Czech Labor Market Programs Evaluation. This evaluation attempted to cover five active labor programs to: (i) determine if participants in the different programs were more successful in re-entering the labor market than were non-participants, and whether this varied across subgroups and with labor market conditions; and (ii) determine the cost-effectiveness of each program and make suggestions for improvements. The evaluation used a matching technique as there was no baseline data collected. The evaluators surveyed participants, and then chose a random sample of non-participants. Since the non-participants were systematically older and less educated, the evaluators needed to construct a reasonable comparison group for each program. This was done by taking each participant in turn and comparing them to each individual in the non-participant pool on the basis of seven characteristics: age, gender, education, number of months employed prior to registration, town size, marital status, and last employment type. The closest match was then put into the comparison group. Although this approach is straightforward, there is the potential for selection bias – that the non-participant group is systematically different from the participant group on the basis of unobservables (Annex 1.5).

Dealing with constraints to developing good controls

At times, randomization or experimental controls are possible, but not politically feasible. In this case, the randomization can be carried out by taking advantage of any plans to pilot the project in certain restricted areas. Areas in which the project will be piloted can initially be randomly selected, with future potential project areas as controls. Over time, additional communities can be randomly included into the project. Three examples illustrate how to handle a situation where randomization was politically or otherwise infeasible. In Vietnam, a Rural Transport project will be evaluated with limited information and no randomization. The Honduras Social Investment Fund provides an example of how to construct a control group in demand-driven projects, using an ex-post matched comparison based on a single cross-section data. Evaluating demand driven can be particularly difficult given that it is not known what projects or communities will participate in the project ahead of time. And third, the evaluation of the Bolivian Social

Investment Fund in the Chaco region provides a good example of how to incorporate randomization in demand-driven projects in a way that allows targeting.

The Vietnam Rural Roads Project. This project aims at reducing poverty in rural areas by improving access to rural communities and linking them to the district and provincial road networks. The design of the impact evaluation centers on baseline and follow-up survey data collected for a sample of project and comparison group communities identified through matched comparison techniques. Baseline and post-intervention information on indicators such as commune level agricultural production yields, income source diversification, employment opportunities, availability of goods, services and facilities, and asset wealth and distribution will be collected from a random sample of project (treatment) and non-project (comparison) communes. These data will be used to compare the change in outcomes before and after the intervention between project and non-project communes using 'double differencing'.

Ideally, treatment and comparison communes should be equivalent in all their observed and unobserved characteristics; with the only difference between them being that treatment communes benefit from the intervention while comparison communes do not. Since random assignment to treatment and comparison groups had not taken place, and the requisite data to make informed choices on appropriate controls were not available at the time of sample selection, random samples of project communes and non-project communes were drawn. Specifically, project communes were selected from a list of all communes with proposed projects in each province. Next, comparison communes were selected from a list of all remaining communes without proposed projects, but in the same districts as treatment communes. Using information collected for the evaluation, propensity-score matching techniques will then be used to ensure that selected non-project communes are appropriate comparison groups. Any controls with unusual attributes relative to the treatment communes will be removed from the sample. (Annex 1.15)

Honduran Social Investment Fund.[17] The Honduran Social Investment Fund (FHIS) aims to improve the living conditions for marginal social groups by financing small scale social and economic infrastructure subprojects. The FHIS is a demand-driven institution that responds to initiatives from municipalities, government ministries, NGOs and community groups by providing financing for investments in infrastructure, equipment and training. The impact evaluation of the FHIS uses matched comparison techniques, drawing the treatment group sample randomly from a list of communities in which FHIS projects have been in operation for at least one year. The comparison group, on the other hand, was selected from a list of 'pipeline' projects – those that have been requested and approved, but where the FHIS investment has not yet taken place. In theory, comparison group communities are automatically matched to project communities according to the self-selection process and FHIS project approval criteria. A household survey was then conducted in both treatment and comparison group communities, complemented by a qualitative evaluation component (focus groups and interviews with

[17] Source: The World Bank, 1998

key informants) conducted in a subset of treatment communities. This initial evaluation is a first step towards establishing an ongoing M&E system that will be eventually integrated into FHIS operations. In particular, the data collected from communities with pipeline projects will become a useful baseline from which to track future changes in impact indicators, after FHIS investment takes place.

Educational Investments in the Chaco Region of Bolivia. Education projects financed by the Bolivian Social Investment Fund (SIF) are aimed at upgrading physical facilities and training teachers in rural public school. Delays in the implementation of the project in the Chaco Region and limited funds for school upgrading provided an opportunity to use an experimental evaluation design while also ensuring that the neediest schools benefit from the project. Schools in the Chaco Region were ranked according to a school quality index based on the sum of five school infrastructure and equipment indicators: electric lights, sewerage, water source; desks per student and square meters of space per student. Only schools below a particular cutoff value were eligible for a SIF intervention. Among eligible facilities, the worst-off schools were automatically selected to benefit from investments financed by SIF. The next highest priority group contained 120 schools, but funds were available to upgrade only less than half of them. Thus, eligible schools in this second priority group were randomly assigned to treatment or comparison groups, providing the conditions for an experimental evaluation design (Annex 1.4).

Combining methods

For most evaluations, more than one technique is required to achieve robust results which address several evaluation questions. Each question may necessitate different techniques, even within one project design. Three examples illustrate how several techniques were combined in one evaluation; the Bolivia Social Fund, the Trabajar Evaluation in Argentina, and the Nicaragua School Reform.

The Bolivia Social Fund. Social funds generally include several different types of subprojects and thus designing an evaluation can involve several approaches. In the Bolivia Social fund, the pattern of project implementation dictated evaluation methods. In the case of education, schools that were to receive the intervention had already been identified therefore randomization could not be used. Instead, matching methods were adopted. In the case of health projects, reflexive methods were used because the intervention was to be implemented in all health centers in the region (see Annex 1.4).

Using a Broad Mix of Evaluation Components, Argentina Trabajar II. The Trabajar evaluation includes an array of components designed to assess how well the program is achieving its policy objectives. The first component draws on household survey data to assess the income gains to Trabajar participants. The second component monitors the program’s funding allocation (targeting), tracking changes over time as a result of reform. This component generates twice-yearly feedback used to refine program targeting. Additional evaluation components include a cost-benefit analysis of infrastructure projects, and social assessments designed to provide community feedback on project implementation. Each of these components has been conducted twice. Three future components are planned. The matched-comparison research technique will be applied again to assess the impact of Trabajar program participation on labor market activity. Infrastructure project quality will be reassessed, this time for projects that have been completed for at least one year to evaluate durability, maintenance and utilization rates. Finally, a qualitative research component will investigate program operations and procedures by interviewing staff members in agencies that sponsor projects as well as program beneficiaries.

The evaluation results provide clear direction to policy reform. The first evaluation component reveals that the Trabajar program is highly successful at targeting the poor - self-selection of participants by offering low wages is a strategy that works in Argentina, and participants do experience income gains as a results of participation. The second component finds that the geographic allocation of program funding has improved over time - the program is now more successful at directing funds to poor areas - however the on-going evaluation process indicates that performance varies and is persistently weak in a few provinces, to which further policy attention is currently being directed. Disappointing evaluation results on infrastructure project quality have generated tremendous efforts by the project team at improving performance in this area through policy reform – insisting of more site visits for evaluation and supervision, penalizing agencies with poor performance at project completion, and strengthening the evaluation manual. And finally, the social assessments uncovered a need for better technical assistance to NGOs and rural municipalities during project preparation and implementation, as well as greater publicity and transparency of information about the Trabajar program (Annex 1.1).

Nicaragua's School Autonomy Reform. In 1993, the Nicaraguan Government took decisive steps to implement a major decentralization initiative in the education sector granting management and budgetary autonomy to selected primary and secondary schools. The goal of the reforms is to enhance student learning – as school management becomes more democratic and participatory, local school management and spending patterns can be allocated towards efforts that directly improve pedagogy and boost student achievement. The impact of this reform has been evaluated using a combination of quantitative and qualitative techniques to asses the outcome as well as the process of decentralization. The purpose of the qualitative component is to illuminate whether or not the intended management and financing reforms are actually taking place in schools, and why or why not. The quantitative component fleshes out these results by answering the question "do changes in school management and financing actually produce better learning outcomes for children?" The qualitative results show that successful implementation of the reforms depends largely on school context and environment (i.e. poverty level of the community), while the quantitative results indicate that increased decision-making by schools is in fact significantly associated with improved student performance.

Different but complementary methodologies and data sources were used to combine both approaches. On the one hand, the quantitative evaluation followed a quasi-experimental design in which test scores from a sample of students in autonomous schools (treatment group) are compared to results from a matched sample of non-autonomous public schools and private schools (comparison group). Data for this component of the evaluation were collected from a panel of two matched school-household surveys and student achievement test scores. The qualitative evaluation design, on the other hand, consisted of a series of key informant interviews and focus groups discussions with different school-based staff and parents in a sub-sample of the autonomous and traditional schools included in the quantitative study.

Using both qualitative and quantitative research techniques generated a valuable combination of useful, policy relevant results. The quantitative work provided a broad, statistically valid overview of school conditions and outcomes; the qualitative work enhanced these results with insight into why some expected outcomes of the reform program had been successful while others had failed, and hence help guide policy adjustments. Furthermore, because it is more intuitive, the qualitative work was more accessible and therefore interesting to Ministry staff, which in turn facilitated rapid capacity-building and credibility for the evaluation process within the Ministry (Annex 1.11).

Exploiting existing data sources

Existing data sources such as a national household survey, census, program administrative record or municipal data can often provide valuable input to evaluation efforts. Drawing on existing sources reduces the need for costly data collection for the sole purpose of evaluation, as illustrated in case of the Viet Nam Rural Roads evaluation. Furthermore, while existing data may not contain all of the information one would ideally collect for purposes of the evaluation, innovative evaluation techniques can often compensate for missing data, as shown in the Kenya National Agricultural Extension Project.

The Vietnam Rural Roads Project. The data used in the this evaluation draws on an effective mix of existing national and local data sources with surveys conducted specifically for the purposes of the evaluation. The evaluation household survey is efficiently designed to replicate a number of questions in the Viet Nam Living Standards Survey (VNLSS), so that drawing on information common to both surveys, regression techniques can be used to estimate each household’s position in the national distribution of welfare.

The evaluation draws extensively on commune-level data collected annually by the communes covering demographics, land use, and production activities. This data source is augmented with a **commune-level survey** conducted for the purposes of the evaluation. Two additional databases were set up using existing information. An extensive **province-level database** was established to help understand the selection of the provinces into the project. This database covers all of Viet Nam’s provinces and has data on a wide number of socio-economic variables. Finally, a **project-level database** for each of the project areas surveyed was also constructed, in order to control for the magnitude of the project and the method of implementation in assessing project impact (Annex 1.15).

The Kenya National Extension Project (NEP). The performance of the Kenya's National Extension Project (NEP) has been controversial, and is part of the larger debate on the cost-effectiveness of the Training & Visit (T&V) approach to agricultural extension services. In the Kenyan context, the debate has been elevated by on the one hand, very high estimated returns to T&V reported in one study (Bindlish and Evenson, 1993), 1997) and on the other, the lack of convincing visible results – including the poor performance of Kenyan agriculture in recent years.

The disagreement over the performance of NEP has persisted pending the results of this evaluation, which was designed to take a rigorous, empirical approach to assessing the program's institutional development and impact on agricultural performance. The evaluation uses a mix of qualitative and quantitative methods to ask highly policy relevant questions, and reveals serious weaknesses in the program: i) The institutional development of NEP has been limited, and after 15 years there is little improvement in the effectiveness of it's services; ii) the quality and quantity of service provision is poor; iii) extension services have only a small positive impact on farm efficiency, and none on farm productivity.

The evaluation is able to draw an array of concrete policy conclusions from these results, many of which are relevant to the design of future agricultural extension projects. First, the evaluation reveals a need to enhance T&V targeting, focusing on areas and groups where the impact is likely to be greatest. Furthermore, advice needs to be carefully tailored to meet farmer demands, taking into account variations in local technological and economic conditions. Successfully achieving this level of service targeting, in turn, calls for appropriate flows of timely and reliable information, hence a program monitoring and evaluation system (M&E) generating a constant flow of feedback from beneficiaries on service content. In order to raise program efficiency, a leaner and less-intense T&V presence with wider coverage is likely to be more cost-effective. The program's blanket approach to service delivery, using a single methodology (farm visits) to deliver standard messages, also limits program efficiency. **Institutional reform** is likely to enhance the effectiveness of service delivery. Decentralization of program design, including participatory mechanisms that give voice to the farmer (such as cost-sharing, farmer organizations, etc.) should become an integral part of the delivery mechanism. Finally, **cost-recovery**, even if only partial, would provide appropriate incentives, address issues of accountability and quality control, make the service more demand-driven and responsive, and provide some budgetary improvement (Annex 1.8).

Costs and Financing

There are no doubt, many costs involved in carrying out an impact evaluation which explains why some countries are reluctant to finance such studies. These costs include data collection, and the value of staff time for all the members of the evaluation team. Financing for an impact evaluation can come from within a project, other Government resources, a research grant, or an outside donor. Information for a sample of World Bank Evaluations shows that while for many countries, the country itself assumed the majority of the evaluation costs, the successful implementation of an impact evaluation required substantial outside resources beyond those provided for in a project's loan or credit. These resources came from a combination of the following sources: (i) a World Bank loan or credit financing for the data collection and processing; (ii) the Government through the salaries paid to local staff assigned to the evaluation effort[18] (iii) a World Bank research grants and bilateral donor grants which financed technical assistance from consultants with specific expertise required for the evaluation; and (iv) the World Bank overhead budget through the staff time provided to guide the impact evaluation and often actively participate in the analytical work.

While few impact evaluations document the cost of carrying out the work, Table 4.2 provides cost estimates for a sample of impact evaluations with World Bank involvement. These cost estimates do not, however, include the value of the staff time contributed by client country counterparts (which can be significant) because this information was unavailable. As a benchmark, in the ten cases below, it was not unusual to have up to five staff assigned to the evaluation effort for several years, a level of effort substantial enough to substantially raise the cost of the evaluation in many of the cases.

The average estimated cost for the impact evaluation was $433,000. This reflects a range from $263,000 for the evaluation of a vocational skills training program for unemployed youth in Trinidad and Tobago to $878,000 for the evaluation of the Bolivian Social Investment Fund. Spending on the impact evaluations for the projects below reflect, on average, 0.6% of the total cost of the project (which sometimes included financing from several donors), or 1.3% of the cost of the IBRD loan or IDA credit . The most expensive components of the evaluations listed below were data collection and consultants, both local and international. In many of the cases travel costs included local staff travel to meet with World Bank staff and researchers for strategy sessions and training, as capacity building for client country staff was a key objective. The two examples below for the impact evaluations of projects in Trinidad and Tobago and Bolivia illustrate some of the variation that can arise in program costs.

The vocational skills training program evaluation in Trinidad and Tobago took advantage of a national income and employment survey to oversample program graduates and create a comparison group from a subset of the national sample. In addition, the evaluation team helped design and use available administrative data kept from records of program applicants, so pre-intervention data were available and no enumeration was required to locate program graduates. The total sample size for each of the three tracer studies was approximately 2500 young people, counting both the treatment and comparison groups. There was only one short questionnaire administered in the survey and the questionnaire was given only to the program graduates. Trinidad and Tobago is a

[18] As is explained in Table 1, these staff costs have not been included in the calculation of the evaluation costs conducted for the cases reviewed in this brief because of data limitations.

small country, communities are relatively easy to access by road, and English is the common language in the country and among program graduates.

Table 4.2. Summary of Estimated Costs from several World Bank Impact Evaluations

Project	**Estimated Cost Of Evaluation**[1]	**Cost as % of Total Project Cost**[2]	**Cost as % of IBRD Loan or IDA credit**[2]	**Breakdown of Evaluation Costs**			
				Travel[3]	*World Bank Staff*	*Consultants*	*Data Collection*
Nicaragua School-Based Management	$495,000	1.26%	1.5%	8.1%	18.1%	39.0%	34.8%
El Salvador School-Based Management	$443,000	0.60%	1.3%	7.7%	7.4%	25.8%	59.2%
Colombia Voucher Program	$266,000	0.20%	0.3%	9.4%	9.8%	21.8%	59.0%
Honduras Social Fund	$263,000	0.23%	0.9%	3.0%	11.5%	53.2%	32.3%
Nicaragua Social Fund	$449,000	0.30%	0.8%	4.9%	33.0%	7.8%	55.7%
Bolivia Social Fund	$878,000	0.50%	1.4%	3.4%	14.6%	12.9%	69.1%
Trinidad and Tobago Youth Training	$238,000	0.80%	1.2%	7.6%	11.5%	17.9%	63.1%
AVERAGE	$433,000	0.56%	1.0%	6.3%	15.1%	25.5%	53.3%

[1]This cost does not include the cost of local counterpart teams not financed from the loan/credit. The figures refer to the time period under which the projects in the evaluation sample were selected, not total financing ever provided by the Bank and others to those institutions.

[2]These costs as a percentage of the Loan/Credit or Project are presented as a reference only. In many cases the actual financing for the evaluation was obtained from sources outside of the project financing.

[3]The travel cost estimates include mission travel for World Bank staff and international consultants to the client countries, as well as travel from client country counterparts, particularly to participate in strategy sessions and analytical workshops with international consultants and World Bank staff.

The Bolivia Social Fund evaluation used its own baseline and follow-up surveys of treatment and comparison groups to evaluate interventions in health, education, water and sanitation. There were no national surveys available from which to conduct analyses or carry out oversampling, placing the entire burden of data collection on the evaluation. The sample of treatment and comparison groups consisted of close to 7,000 households and 300 facilities interviewed in both the 1993 baseline survey and 1998 follow-up survey. In Bolivia, the data collection instruments for the impact evaluation consisted of portable laboratories for conducting laboratory-based water quality tests, achievement tests and 8 questionnaires for informants from households and facilities.[19] To assess

[19] The 8 questionnaires in the Bolivian Social Investment Fund impact evaluation consisted of: two household questionnaires (one for the principal informant and one for women of childbearing age), a community questionnaire, four different health center questionnaires for the different types of health

targeting, the evaluation included a consumption-based measure of poverty which required the collection of detailed consumption data from households as well as regional price data from communities. The fieldwork was conducted in rural areas where the majority of the Social Investment Fund projects are located and included a random sample of rural households which were often accessible only by foot or horseback. Finally, the questionnaires had to be developed and administered in Spanish, Quechua and Aymara.

Political Economy Issues

There are several issues of political economy which affect not only whether or not an evaluation is carried out, but also how it is implemented. The decision to proceed with an evaluation is very much contingent on strong political support. Many Governments do not see the value of evaluating projects and thus do not want to invest resources in this. Additionally, there may be reluctance to allow an independent evaluation which may find results contrary to Government policy, particularly in authoritarian or close regimes. More open Governments may, however, view evaluations and the dissemination of the findings as an important part of the democratic process.

Evaluations are also sensitive to political change. Three of the ten impact listed in Table 4.2 were canceled because of political economy issues. Turnover in regimes or key posts within a counterpart government office and shifts in policy strategies can affect not only the evaluation effort, but more fundamentally the implementation of the program being evaluated. One example of this type of risk comes from the experience of a team working on the design and impact evaluation of a school-based management pilot in Peru as part of a World Bank financed primary education project. The team composed of Ministry of Education officials, World Bank staff, international and local consultants had worked for over a year developing the school-based management models to be piloted, establishing an experimental design, designing survey instruments and achievement tests, and collecting baseline data on school characteristics and student achievement. Just prior to the pilot's introduction in the randomly-selected schools, high level government officials canceled the school-based management experiment in a reaction to perceived political fallout from the pilot. A similar reform was introduced several years later, but without the benefit of a pilot test or an evaluation.

In Venezuela, an evaluation of a maternal and infant health and nutrition program was redesigned three times with three different client counterparts as the government shifted responsibility for the evaluation from one agency to another. Each change was accompanied by a contract renegotiation with the private sector firm that had been identified to carry out the data collection and the majority of the analysis for the evaluation. When the legitimacy of the third government counterpart began to be questioned, the firm nullified the contract and the evaluation was abandoned. These

centers ranging from small community clinics to hospitals, and a school questionnaire for the school director.

incidents occurred during a period of political flux characterized by numerous cabinet reshufflings that ended with the collapse of the elected government serving as a counterpart for the project, so the evaluation was hardly alone in suffering from the repercussions of political instability. Nonetheless, in both the Peruvian and Venezuelan cases, it is sobering to reflect upon the amount of resources devoted to an effort that was never brought to fruition. A less dramatic example of the effect of political change on evaluation strategies comes from El Salvador where the recognized success of a reform introduced in rural schools prompted the government to introduce a similar education reform in all of the urban schools at once, instead of randomly phasing in schools in over time, as originally planned. This decision eliminated the possibility of using an experimental design and left it using a less-robust reflexive comparison as the only viable evaluation design option.

Bibliography

Atkinson, Anthony, 1987, “On the Measurement of Poverty”, Econometrica, 55: 749-64.

Bamberger, Michael, 1999. *Integrating Quanititative and Qualitative Methods in Development Research,* World Bank, forthcoming.

Barnes, Carolyn. “Assets and the Impact of Microenterprise Finance Programs.” *USAID AIMS Project Brief* 6 (1996).

Barnes, Carolyn, and Erica Keogh. “An assessment of the Impact of Zambuko’s Microenterprise Program in Zimbabwe: Baseline Findings.” *USAID AIMS Project Brief* 23 (1999).

Benus, Jacob, Neelima Grover, and Recep Varcin. *Turkey: Impact of Active Labor Programs.* Bethesda, MD: Abt Associates, 1998.

Benus, Jacob, Neelima Grover, Jiri Berkovsky, and Jan Rehak. *Czech Republic: Impact of Active Labor Programs.* Bethesda, MD: Abt Associates, 1998.

Besharov, Douglas J., Peter Germanis, and Peter H. Rossi. *Evaluating Welfare Reform: A Guide for Scholars and Practitioners*. College Park, MD: The University of Maryland, 1997.

Bloom, Howard S., Larry L. Orr, George Cave, Stephen H. Bell, Fred Doolittle, and Winston Lin. *The National JTPA Study: Overview, Impacts, Benefits, and Costs of Title II-A*. Bethesda, MD: Abt Associates, 1994.

Blundell, Richard W. and R.J. Smith, 1993, “Simultaneous Microeconometric Models with Censoring or Qualitative Dependent Variables”, in G.S. Maddala, C.R. Rao and H.D. Vinod (eds) Handbook of Statistics Volume 11 Amsterdam: North Holland.

Bourguignon, Francois, Jaime de Melo, and Akiko Suwa, 1991 “Distributional Effects of Adjustment Policies: Simulations for Archetype Economies in Africa and Latin America,” World Bank Economic Review, Vol. 5, No. 2, pp. 339-366.

Burtless, Gary. “The Case for Randomized Field Trials in Economic and Policy Research.” *Journal of Economic Perspectives* 9 (Spring 1995): 63-84.

Card, David, and Philip K. Robins. “Do Financial Incentives Encourage Welfare Recipients To Work? Evidence From A Randomized Evaluation of the Self-Sufficiency Project. *National Bureau of Economic Research* Paper 5701, August 1996.

Carvalho, Soniya, and Howard White. “Indicators for Monitoring Poverty Reduction.” *World Bank Discussion Papers* 254 (1994).

Chen, Martha Alter, and Donald Snodgrass. “An Assessment of the Impact of Sewa Bank in India: Baseline Findings.” USAID AIMS Project, Mimeo, 1999.

Cohen, Monique, and Gary Gaile. “Highlights and Recommendations of the Second Virtual Meeting of the CGAP Working Group on Impact Assessment Methodologies.” USAID AIMS Project, Washington, DC: Management Systems International, 1998.

———. “Highlights and Recommendations of the Virtual Meeting of the CGAP Working Group on Impact Assessment Methodologies.” AIMS Project, Washington, DC: Management Systems International, 1997.

Dar, Amit, and Indermit S. Gill. “Evaluating Retraining Programs in OECD Countries: Lessons Learned.” *The World Bank Research Observer* 13 (February 1998): 79-101.

Dar, Amit, and Zafiris Tzannatos. “Active Labor Market Programs: A Review of Evidence from Evaluations.” *World Bank Social Protection Discussion Paper* 9901 (1999).

Dehejia, Rajeev H. and Sadek Wahba. “Causal Effects in Non-experimental Studies: Re-evaluating the Evaluation of Training Programs."” *NBER Working Paper Series* 6586 (1998). <http://www.nber.org/papers/w6586>

Dennis, Michael L. and Robert F. Boruch. “Randomized Experiments for Planning and Testing Projects in Developing Countries: Threshold Conditions.” *Evaluation Review* 13 (June 1989): 292-309.

Diagne, Aliou, and Manfred Zeller. “Determinants of Access to Credit and Its Impact on Income and Food Security in Malawi.” Manuscript submitted to IFPRI’s Publication Review Committee for consideration as an IFPRI Research Report (1998).

Diop, F, A Yazbeck and R. Bitran “The impact of alternative cost recovery schemes on access and equity in Niger” Health Policy and Planning, 10(3) 223-240

Donecker, Jane, and Michael Green. *Impact Assessment in Multilateral Development Institutions*. London: Department for International Development, 1998.

Dunn, Elizabeth. “Microfinance Clients in Lima, Peru: Baseline Report for AIMS Core Impact Assessment.” USAID AIMS Project, Mimeo, 1999.

Edgecomb, Elaine L., and Carter Garber. “Practitioner-Led Impact Assessment: A Test in Honduras.” *USAID AIMS*, 1998.

Ezemenari, Kene, Anders Ruqvist, and K. Subbarao. *Impact Evaluation: A Note on Concepts and Methods*. World Bank Poverty Reduction and Economic Management Network, Mimeo, 1999.

Foster, James, J. Greer, and Erik Thorbecke, 1984, “A Class of Decomposable Poverty Measures”, Econometrica, 52: 761-765.

Friedlander, Daniel, and Gayle Hamilton. *The Saturation Work Initiative Model in San Diego: A Five-Year Follow-up Study*. New York: MDRC, 1993.

Friedlander, Daniel, and Philip K. Robins. “Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods.” *American Economic Review* 85 (September 1995): 923-937.

Friedlander, Daniel, David H. Greenberg, and Philip K. Robins. “Evaluating Government Training Programs for the Economically Disadvantaged.” *Journal of Economic Literature* 35 (December 1997): 1809-1855.

Fuller, Bruce and Magdalena Rivarola. 1998. *Nicaragua’s Experiment to Decentralize Schools: Views of Parents, Teachers and Directors*. Working Paper Series on Impact Evaluation of Education Reforms, paper no. 5. The World Bank. Washington, DC.

Gaile, Gary L., and Jenifer Foster. “ Review of Methodological Approaches to the Study of the Impact of MicroEnterprise Credit Programs.” *USAID AIMS Brief* 2, 1996.

Glewwe, Paul, Michael Kremer, and Sylvie Moulin. “Textbooks and Test Scores: Evidence from a Prospective Evaluation in Kenya,” (October 28, 1998).

Goodman, Margaret, Samuel Morley, Gabriel Siri and Elaine Zuckerman. *Social Investment Funds in Latin America: Past Performance and Future Role*. Washington DC, March 1997.

Government of Denmark. *Methods for Evaluation of Poverty Oriented Aid Interventions*. Copenhagen, Denmark: Ministry of Foreign Affairs, 1995

Greene, W. H., *Econometric Analysis*, New Jersey, Prentice Hall Press, 1997.

Greenberg, David, and Mark Shroder. *The Digest of Social Experiments*, 2nd ed. Washington, DC: The Urban Institute Press, 1997.

Grossman, Jean Baldwin. “Evaluating Social Policies: Principles and U.S. Experience.” *The World Bank Research Observer* 9 (July 1994): 159-180.

Habicht, JP, CG Victoria, and JP Vaughan. “Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact.” *International Journal of Epidemiology* 28 (1999): 10-18.

Harrell, Adele, Martha Burt, Harry Hatry, Shelli Rossman, Jeffrey Roth, and William Sabol. *Evaluation Strategies for Human Service Programs: A Guide for Policymakers and Providers*. Washington, DC: The Urban Institute, 1996.

Heckman, J., H. Ichimura, J. Smith, and P. Todd, 1998, “Characterizing Selection Bias using Experimental Data”, Econometrica, 66: 1017-1099.

Heckman, James and Richard Robb, 1985, “Alternative Methods of Evaluating the Impact of Interventions: An Overview”, Journal of Econometrics, 30: 239-67.

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

Heckman, James, 1997, “Instrumental Variables. A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations”, Journal of Human Resources, 32(3): 441-461.

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd. “Characterizing Selection Bias Using Experiential Data.” NBER Working Paper Series 6699 (1998). <http://www.nber.org/papers/w6699>

Hicks, Norman. *Measuring the Poverty Impact of Projects in LAC*. World Bank LCSPR, 1998.

Holder, Harold D., Andrew J. Treno, Robert F. Saltz, and Joel W. Grube. “Summing Up: Recommendations and Experiences for Evaluation of Community-Level Prevention Programs.” *Evaluation Review* 21 (April 1997): 268-278.

Hollister, Robinson G., and Jennifer Hill. *Problems in the Evaluation of Community-Wide Initiatives*. New York: Russell Sage Foundation, 1995.

Hulme, David. “Impact Assessment Methodologies for Microfinance: Theory, Experience and Better Practice.” Manchester: Institute for Development Policy and Management, University of Manchester, 1997.

International Food Policy Research Institute. *Programa Nacional de Educación, Salud, y Alimentación (PROGRESA): A Proposal for Evaluation* [with *Technical Appendix*]. Washington, DC: IFPRI, 1998.

Jalan, Jyotsna and Martin Ravallion. “ Income Gains from Workfare: Estimates for Argentina’s Trabajar Program Using Matching Methods.” Washington DC: Development Research Group, World Bank, 1998

———. “Transfer Benefits from Workfare: A Matching Estimate for Argentina.” Washington DC: Development Research Group, World Bank, 1998.

Jamison, Dean T., Barbara Serle, Klaus Galda and Stephen P. Heyneman "Improving Elementary Mathematics Education in Nicaragua: An Experimental Study of the Impact of Textbooks and Radio on Achievement" Journal of Educational Psychology, 73(4), August 1981556-567

Karoly, Lynn A., Peter W. Greenwood, Susan S. Everingham, Jill Houbé, M. Rebecca Kilburn, C. Peter Rydell, Matthew Sanders, and James Chiesa. *Investing in Our Children: What We Know and Don't Know About the Costs and Benefits of Early Childhood Interventions*. Santa Monica, CA: Rand, 1998.

Kemple, James J., Fred Doolittle, and John W. Wallace. *The National JTPA Study: Site Characteristics and Participation Patterns*. New York: MDRC, 1993.

Khandker, Shahidur R. 1998. *Fighting Poverty with Microcredit: Experience in Bangladesh.* New York: Oxford University Press for the World Bank.

Killick, Tony, IMF Programmes in Developing Countries, Design and Impact, 1995, London, England.

King, Elizabeth, and Berk Ozler. 1998. *What's Decentralization Got To Do With Learning? The Case of Nicaragua's School Autonomy Reform*. Working Paper Series on Impact Evaluation of Education Reforms, paper no. 9. The World Bank. Washington, DC.

King, Elizabeth, Berk Ozler and Laura Rawlings. 1999. *Nicaragua's School Autonomy Reform: Fact or Fiction?* The World Bank. Washington, DC.

Levinson, James F., Beatrice Lorge Rogers, Kristin M. Hicks, Thomas Schaetzel, Lisa Troy, and Collette Young. *Monitoring and Evaluation: A Guidebook for Nutrition Project Managers in Developing Countries*. Medford, MA: International Food and Nutrition Center, Tufts University School of Nutrition Science and Policy, March 1998.

Manski, Charles and Irwin Garfinkel (eds), 1992, Evaluating Welfare and Training Programs, Cambridge, Mass: Harvard University Press.

Manski, Charles and Steven Lerman, 1977, "The Estimation of Choice Probabilities from Choice-Based Samples", Econometrica, 45: 1977-88.

Meyer, Bruce D., 1995, "Natural and Quasi-Experiments in Economics", Journal of Business and Economic Statistics, April.

Miles, Matthew B. and A. Michael Huberman, 1994, *Qualitative Data Analysis*, Sage Publications, London.

MkNelly, Barbara and Karen Lippold. "Practitioner-led Impact assessment: A Test In Mali." *USAID AIMS Brief* 21 (1998).

MkNelly, Barbara and Christopher Dunford (in collaboration with the Program in International Nutrition, University of California, Davis)"Impact of Credit with Education on Mothers' and their Young Children's Nutrition: Lower Pra Rural Bank Credit with Education Program in Ghana" Freedom from Hunger Research Paper No. 4, March 1998

Moffitt, Robert, 1991, "Program Evaluation with Nonexperimental Data", Evaluation Review, 15(3): 291-314.

Mohr, Lawrence B. "The Qualitative Method of Impact Analysis." *Evaluation Review* (forthcoming, 1999).

———. *Impact Analysis for Program Evaluation,* 2nd ed. Thousand Oaks, CA: Sage Publications, 1995).

Morduch, Jonathan. "The Microfinance Promise." *Journal of Economic Literature* (forthcoming, 1999).

———. "Does Microfinance Really Help the Poor? New Evidence from Flagship Programs in Bangladesh." Harvard University Institute for International Development, Mimeo, June 1998.

———. "The Microfinance Schism." Harvard Institute for International Development, *Development Discussion Paper* 626 (February 1998).

Newman, John. "Impact Evaluation Study: Bolivian Social Investment Fund (Part of SIF 2000 Study)." (November 12, 1998).

Newman, John, Laura Rawlings, and Paul Gertler. "Using Randomized Control Designs in Evaluating Social Sector Programs in Developing Countries." *The World Bank Research Observer* 9 (July 1994): 181-202.

Operations Evaluation Department. "An Overview of Monitoring and Evaluation in the World Bank." (June 30, 1994).

Poppele, J. S. Summarto and L. Pritchett, 1999. "Social Impacts of the Indonesia Crisis: New Data and Policy Implications", SMERU, World Bank Processed.

Pradhan, Menno, Laura Rawlings, and Geert Ridder. 1998. *The Bolivian Social Investment Fund: An Analysis of Baseline Data for Impact Evaluation*. World Bank Economic Review, 12(3). Pp. 457-82.

Ravallion, Martin. 1999. Monitoring Targeting Performance when Decentralized Allocations to the Poor are Unobserved. The World Bank. Washington, DC. Processed.

Ravallion and Wodon, 1998, Evaluating a Targeted Social Program when Placement is Decentralized, Policy Research Working Paper, 1945, World Bank.

Ravallion, Martin, 1994, Poverty Comparisons, Fundamentals in Pure and Applied Economics Volume 56, Harwood Academic Publishers.

Ravallion, Martin, Dominique van de Walle, and Madhur Gautam. "Testing a social safety net." *Journal of Public Economics* 57 (1995): 175-199.

Rawlings, Laura. Forthcoming. *Assessing Educational Management and Quality in Nicaragua*, in [*book title*]. The World Bank. Washington, DC.

Rebien, Claus C. "Development Assistance Evaluation and the Foundations of Program Evaluation." *Evaluation Review* 21 (August 1997): 438-460.

Revenga, Ana, Michelle Riboud and Hong Tan "The Impact of Mexico's Retraining Program on Employment and Wages" World Bank Economic Review, 8(2), 1994 p 247-277.

Rivers, Douglas and Quang H. Vuong, 1988, "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models", Journal of Econometrics, 39: 347-366.

Robinson, Sherman, 1991, "Macroeconomics, Financial Variables, and Computable General Equilibrium Models", World Development, Volume 19, No. 11, pp 1509-1525.

Rosenbaum, P. and D. Rubin, 1983, "The Central Role of the Propensity Score in Observational Studies for Causal Effects", Biometrika, 70: 41-55.

Rosenbaum, P. and D. Rubin, 1985, "Constructing a Control Group using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," American Statistician, 39: 35-39.

Rossman, Gretchen B. and Bruce L. Wilson. "Numbers And Words: Combining Quantitative and Qualitative Methods in a Single Large-Scale Evaluation Study. *Evaluation Review* 9:3 (October 1985): 627-643.

Sahn, David, Paul Dorosh, and Stephen Younger. "Exchange Rate, Fiscal and Agricultural Policies in Africa: Does Adjustment Hurt the Poor?" in World Development, Vol. 24, No. 4, pp. 719-747, 1996

Sebstad. Jennifer and Gregory Chen. "Overview of Studies on the Impact Of MicroEnterprise Credit." *USAID AIMS,* June 1996.

Sharma, Manohar, and Gertrud Schrieder. "Impact of finance on food security and poverty alleviation – a review and synthesis of empirical evidence." Paper presented to Workshop on Innovations in Micro-Finance for the Rural Poor: Exchange of Knowledge and Implications for Policy, Ghana, November 1998.

Smith, William. "Group Based Assessment of Change: method and results." Hanoi, Vietnam: ActionAid, 1998.

Subbarao, K., Kene Ezemenari, John Randa, Gloria Rubio. *Impact Evaluation in FY98 Bank Projects: A Review*. World Bank Poverty Reduction and Economic Management Network, Mimeo, January 1999.

Tan, J.P, J. Lane and G. Lassibille " Schooling Outcomes in Philippine Elementary Schools: Evaluation of the Impact of Four Experiments" World Bank Economic Review, September 1999.

Taschereau, Suzanne. *Evaluating the Impact of Training and Institutional Development Programs, A Collaborative Approach*. Economic Development Institute of the World Bank, January 1998.

USAID. "Assessing The Impacts of MicroEnterprise Interventions: A Framework for Analysis." Microenterprise Development Brief, June 1995.

———. "Study of Low Cost Housing." New Delhi, India: USAID, June 1984.

van de Walle, Dominique. 1999. Assessing the Poverty Impact of Rural Road Projects. The World Bank. Processed.

Weiss, Carol H. 1998. *Evaluation*. Prentice Hall, New Jersey

Wilson, Sandra Jo. "A Multi-Method Proposal to Evaluate the Impact of CGIAR-Instigated Activities on Poverty Reduction." Paper presented to the Annual Conference of the American Evaluation Association, November 1998.

Wouters, A. "Improving quality through cost recovery in Niger" Health Policy and Planning 10(3) 257-270

World Bank, Operations and Evaluation Department. *1998 Annual Review of Development Effectiveness*. Washington, DC: The World Bank, 1998.

———, Poverty Reduction and Economic Management, Latin America and Caribbean Region. *Measuring the Poverty Impact of Projects in LAC*. Washington DC: The World Bank, 1998.

——— , Poverty and Human Resources Division, Policy Research Department. *Impact Evaluation of Education Projects Involving Decentralization and Privatization*. Working Paper. Washington DC: The World Bank, 1994.

——— , Learning and Leadership Center, and Operations Policy Department, *Handbook on Economic Analysis of Investment Operations,* Washington DC: The World Bank, 1994.

Zeller, Manfred, Akhter Ahmed, Suresh Babu, Sumiter Broca, Aliou Diagne, and Manohar Sharma. "Security of the Poor: Methodologies for a Multicountry Research Project." IFPRI *FCND Discussion Paper* 11 (1996).