Impact Evaluation of Social Programs: A Policy Perspective

John Blomquist

Revised Draft, September 2003

Comments Welcome

The findings, interpretations, and conclusions expressed in this paper are entirely those of the author(s) and should not be attributed in any manner to the World Bank, to its affiliated organizations or to members of its Board of Executive Directors or the countries they represent.

Abstract

There is increasing recognition among governments and donor organizations that rigorous evaluations of public interventions should feature in the social policy decisionmaking process. Yet there is frequently a gap between the desire for information on the effectiveness of programs and an understanding of the potential and the limitations of evaluation tools. What questions can evaluations answer? What administrative structures are required to implement them? What are the political and social factors surrounding the acceptance of evaluations by target groups and the public? How much do evaluations cost? How long do they take to complete? This paper addresses these questions by drawing from the experiences of recent evaluations of social safety net programs conducted in both developing and developed countries.

While the focus is on safety net interventions, conclusions are applicable to broader social policy. In general, evidence suggests that formal impact evaluations are a valuable policy tool, but must be carefully designed and planned in advance of implementation, and should be used in conjunction with other performance management systems. If designed and implemented properly, evaluations can provide unique information critical to the formulation of sound social policy

Table of Contents

Impact Evaluation of Social Programs: A Policy Perspective

John Blomquist, Senior Economist
Social Protection Human Development Network

I. Introduction

There is increasing recognition among many governments and donor organizations that rigorous evaluations of public interventions should feature in the social policy decisionmaking process. As pressures worldwide mount to reduce the size of governments and expand private sector and nongovernmental involvement in social services, it becomes increasingly important to justify public spending and ensure that the funded interventions are achieving intended objectives. Countries from Chile to Indonesia to Sweden have embraced evaluation as a crucial element of good public sector management. The international community has also turned to more systematic evaluation of its own programs in an effort to make aid and assistance more effective.

The single most critical policy question pertaining to a public program is whether in a cost-effective manner it truly helps those who participate in it. This and related questions are addressed by a special class of evaluation known as program impact evaluation. Impact evaluations can provide information on whether a program measurably benefits participants, determine if it is cost-effective relative to other options, and yield insights into why a program may not deliver as intended. Collectively, impact evaluations provide the best evidence on which programs and policies are likely to help a society achieve its social goals.

Yet many policy stakeholders, including development organizations, government officials and program proponents in both developed and developing countries, exhibit a reluctance to undertake formal evaluation of social programs. A study by Rubio and Subbarao (2001) found that among a sample of social protection projects supported by the World Bank in 1999, just over 20 percent had well-developed evaluation plans, and only half possessed an information base suitable for evaluation with most having incomplete or no plans to evaluate impacts. There are numerous examples of impact evaluations that have been planned by governments, only to be shelved or cancelled for political or cost considerations or a change in administration.

There are two main reasons for this reluctance. Broadly, the reasons have to do, first, with perceived limitations of the art of evaluation and, secondly, with the political economy of the public policy environment. More specifically, they involve: (i)

Confusion and misunderstanding regarding what impact evaluations can deliver. Results are not always available on a timely enough basis for policymakers and they can appear ambiguous and difficult to translate into policy actions; and (ii) Political concerns over the conduct of a formal evaluation and the possible repercussions from the results. Evaluation is assumed to be very costly, particularly in relation to the scarce resources available for social programs. Negative findings have the potential to hinder social agendas and damage political careers. These concerns, justified or not, have conspired to limit the implementation of impact evaluations in many settings.

This paper will explore the political economy considerations surrounding impact evaluations, focusing on barriers to implementation, common misconceptions, and the potential to expand the use of evaluation. To adequately review the political aspects of evaluation, it is necessary to discuss the components and techniques in some detail. Therefore, a second purpose is to provide a non-technical primer on the impact evaluation of social programs. The intention is to concisely present key features and lessons that can be readily digested by those considering an impact evaluation. Resources for more indepth treatment of topics can be found in footnotes and in the bibliography.

The next section clarifies the meaning of program evaluation, the program policy questions that can be addressed, and the components of complete evaluations. Section three identifies the steps and methods used to conduct a good evaluation, with a non-technical summary of the most common evaluation methods. Section four examines key political economy issues and outlines elements needed to establish a culture conducive to incorporating impact evaluation into the policy process. Section five concludes.

II. What is Program Impact Evaluation?

There are many types of evaluations. Stating that a program or policy has been evaluated does not in itself suggest what analysis was performed, or even whether or how outcomes have been examined. According to the OECD, evaluations are broadly "analytical assessments addressing results of public policies, organizations or programs, that emphasize reliability and usefulness of findings." (OECD 1999). This definition encompasses many types of assessments, including policy-level evaluations, concurrent assessments, tracer studies, rapid appraisals and beneficiary assessments, indicator monitoring, and even public expenditure tracking surveys in the context of public sector management. Each of these can have an important role in a system of monitoring and evaluation. But impact evaluations have a special meaning and defined purpose. Over time, they have evolved into a fairly uniform set of related analyses.

Finding Impacts and Answering Program Questions

An impact evaluation is an assessment of the impacts on participants that can be attributed to direct participation in a program or intervention. It attempts to determine whether the program as implemented does what it is intended to do for participants, and it is this determination of true program "impacts" that distinguishes impact evaluation from

other assessments. In the context of development, impact evaluations have been defined as "systematic identification of the effects . . . on individuals, households, institutions, and the environment caused by a given development activity such as a program or project." (World Bank 2002, pg. 20).

The specific techniques of estimating impacts vary according to setting, as will be seen below. But the fundamental conceptual exercise is the same for all. Conceptually, impacts are determined by comparing the relevant outcomes of program participants with the outcomes those same individuals would have experienced in the absence of the program. Such an experiment is impossible, of course, and all methodologies center on ways to construct a plausible counterfactual comparison group.

It should be stressed that impact evaluation is quite distinct from program monitoring, despite the fact that "monitoring and evaluation (M&E)" are often lumped together in the development and public management literature. Simply put, ". . . evaluation is concerned with tracing causes to outcomes whereas monitoring is concerned with tracking the progress of implementation and processes." (Ezemanari et. al. 1999 pg. 1). Program monitoring involves setting performance indicators and reviewing administrative implementation through Management Information Systems and other means as the program is active. Program evaluation takes a retrospective, summative perspective in examining impacts after the program has been completed.[1] It may well use the data afforded through monitoring exercises, but it is a separate undertaking.

Impact evaluations would be valuable if they only addressed the question of participant impacts. But they generally address a range of policy-relevant questions through different analytic components. Impact evaluation can typically answer the following sorts of questions about a program:

- Does the program or intervention achieve the stated goals? Does it have unintended effects on participants?
- Are program impacts stronger for particular groups or subsets of participants?
- Is the program cost effective in relation to other options?
- What are likely reasons why the program is or isn't successful?
- How can the design or implementation be changed to improve performance?

Full impact evaluations assess the complete effects of the program as well as its operation. Necessarily, they have an analytic element and a speculative/interpretative element. The analytic element involves the calculation of pure program impacts given a suitable comparison or control group of nonparticipants. The interpretative element refers to the task of examining why the impacts (or lack of them) are what they are, and what the resulting program and policy implications may be. The two together constitute a good evaluation.

[1] It is not necessary for the program to formally end, only that a group of participants has finished its involvement with the program over some defined time period.

For example, if the program is a cash transfer program conditional on children attending school and getting regular medical care, an impact evaluation would seek to determine whether the program results in higher school attendance and graduation rates and fewer child sicknesses than is the case for the comparison group that doesn't participate in the program. Are families that participate less poor after receiving benefits than those who don't? Are there other benefits to program participation, such as improved nutrition of family members or reduced illicit behaviors? Are there particular age groups or types of families that benefit more than others? What aspects of program operations are likely to contribute to the success or failure of the intervention? Are different types of schools associated with varying participant impacts? Is the program more costly than alternative existing programs? Do the participant benefits justify the costs?

Note that evaluations cannot answer every question a policy maker might like to know in relation to a particular program. In particular, they cannot address many "what if" questions. What if the program were made national? What if the means-test were increased? These questions can be examined, if not definitively, using a variety of ex ante simulation and modeling methods that are generally not part of standard evaluations. Impact evaluations focus on assessing the existing program *as implemented*. However, a variety of different questions can be addressed with careful study design and advance planning.

Elements of Typical Impact Evaluations

The use and methodology of impact evaluation is not new. The common techniques have been used since the 1960's and 1970's, many of which were pioneered in the evaluation of US government public policy programs. In fact, a small private industry emerged from the 1970's to meet the demand of the US government, particularly regarding social safety net and employment and training programs. Full program evaluations have evolved to include several elements, or related study components.

Impact evaluations often consist of the following components:

- **Process study**. This analysis examines the operations and processes that make up the particular program under study. It is not an examination of impacts on participants.
- **Impact assessment**. This analysis examines impacts on participants, and requires survey data and econometric methods to isolate the effects. The techniques used vary from random assignment to simple reflexive assessments, and represent the heart of an impact evaluation.[2]
- **Cost-benefit assessment**. Calculates the costs of program operation and compares them with the benefits to determine its net value. Two versions can

[2] Some authors have commented that modern evaluation has become too routinized, relying too heavily on standard impact evaluation techniques with too little attention to the qualitative assessment of program implementation processes (see for example Manski, 1990).

be conducted. Cost-effectiveness analysis estimates inputs in monetary terms and outcomes in non-monetary quantitative terms. Cost-benefit analysis estimates both inputs and outputs in monetary terms.

Until recently, full evaluations with the above components were generally conducted in North America, Australia, and parts of Europe. Recently, though, the use has spread to much of Latin America , the Caribbean and East Asia, along with some Eastern European countries, particularly with respect to assessing labor market interventions.

III. How is an Impact Evaluation Conducted?

Conducting a successful impact evaluation requires advance planning. Insufficient planning can unnecessarily compromise an evaluation effort, weakening findings and their subsequent policy value. A good evaluation plan will: (i) establish evaluation objectives; (ii) determine appropriate evaluation methods; (iii) provide a data collection strategy and identify available sources; and (iv) establish a timeline for producing and disseminating findings. Each of the elements is discussed below, with attention to the pros and cons associated with the available choices.

Objectives

Evaluation objectives should consist of the desired policy questions to be answered balanced against the likely resource constraints, including time, money, and data availability. Determining objectives therefore requires a good understanding of the different elements of impact evaluation and their costs.

The principal objective is always to determine whether the program is helping beneficiaries. Therefore, the program objectives are central. Some examination of available data and measurable indicators should be conducted prior to determining objectives to avoid setting an unachievable goal. For example, if the program objective is to raise the self-esteem of disadvantaged youth, care must be taken to assess how this will be measured and translated into an evaluation objective.

A clear statement of objectives is essential prior to selecting an evaluator. The statement will include features of the program to be analyzed, the outcomes to be assessed, and the time period over which the analysis is expected. For example, stating as an object that evaluation will examine the impact of the program on participants' health, education, and consumption is open to wide interpretation. Restating the objectives in terms of grade school annual attendance, annual sick days, promotion to middle school, core disease immunization rates of children under the age of five years, and annual consumption of participant families is more precise. Such detail is important to avoid future confusions, and should be incorporated into a complete terms of reference.

The Evaluator

Choosing an appropriate entity to evaluate the program is crucial. For some types of evaluations, it may be appropriate to allow an internal government agency to conduct the study. However, for the special purpose and requirements of an impact evaluation, a specialized external evaluator is preferred.

There are two main factors to consider. First, to have any policy value, the evaluation must be objective and reasonably independent. Since government officials are involved in the design and administration of programs, using government staff in audit agencies or evaluations units within ministries to evaluate those programs may not be completely objective. There may be political pressure to bias the reporting of results in a particular way. Even if internal evaluators were unbiased, the legislature and the public might not accept the results. An external evaluator enhances public credibility. The second factor is that quantitative impact evaluation requires specialized skills and expertise. Detailed knowledge of sampling techniques, survey design and data collection, and impact determination methodology are needed to conduct a rigorous evaluation. In most countries, these skills, if they exist, are to be found in academia and consulting firms.

There are potential political and resource costs of not using insiders, however. An agency may be less willing to use evaluation results if it has not been involved in the planning or if staff feel they or their program are being judged from the outside. In addition, capacity building and knowledge transfer to agencies is weakened if a private firm is contracted. Finally, government administrators and agencies have the best understanding of the inner workings of programs, they frequently have a sense of whether there are impacts, and they are likely to know key administrative and other data sources.

On balance, the skilled external evaluator will be in the best position to deliver a quality impact evaluation. A range of acceptable choices and teaming arrangements is possible. If the scope of the evaluation is limited and uses existing data, local university professors might be sufficient or even specialized research staff from an international agency, if available. If a large-scale experimental design is envisioned, a competitive contract could be issued to an international firm specializing in impact evaluations. In all cases, the evaluation team should work closely with the relevant government agencies and program administrators. It is important that a sense of ownership be developed and that the evaluation is not just some funding requirement or another piece of irrelevant analysis.

Care should be taken to convey the objectives of the evaluation in as much detail as possible to avoid misunderstandings and unmet expectations on the part of either the evaluator, the client, or the program administrators. A terms of reference or request for proposal should be prepared, giving details of the expected data requirements, preferred methodology, sample sizes if possible, and a timeline for activities and reports.[3]

[3] Baker (2000) contains sample terms of references.

Evaluation proposals may also be solicited if there are a number of possible evaluators. It is always desirable to have the evaluator prepare an evaluation design report early in the process to refine the evaluation planning in light of actual conditions.

Quantitative Impact Estimation Methods

Methods to estimate impacts fall into two strategies, experimental or non-experimental. Within the non-experimental design are several different methods to identify the comparison group and to statistically adjust for differences between participants and comparisons. In addition, qualitative methods should be used to complement and enhance the quantitative techniques.

Experimental design. This is the preferred design because it eliminates, under weak conditions, bias in the estimates of impacts. Bias in impact estimation results from pre-existing or ex ante differences between the participants and the comparison group that can be confounded with the effects of program participation. See Box 1 for more on bias. In an experimental design, the outcomes of program participants are compared to those who are statistically just like the participants except they do not participate in the program. This is achieved by randomly assigning a set of individuals or households who are eligible and have volunteered to participate into two groups, a treatment group that participates in the program and a control group that is denied entry into the program.

Because members of both groups have volunteered to participate and are eligible, they do not differ from one another in terms of their motivations or abilities except by chance. Similarly, the groups do not differ on measurable socioeconomic characteristics because they have been randomly assigned, and remaining difference can be accounted for in the data. In this way, any observed differences over time between the groups will be due, on average, to participation in the program. Calculating program impacts is then a matter of subtracting the mean of the control group outcome (or outcomes) from the mean of the treatment group outcome(s).

Two conditions are needed to ensure that random assignment eliminates pre-existing differences between treatments and controls: a) the assignment procedure must be truly random and the process should not affect the program itself, and b) control group members should not have access to the program or to a close substitute which can affect outcomes prior to measurement.[4] These can fairly reasonably be met in many circumstances. however there can be situations where random assignment disrupts normal behavior, such as if control group members denied participation decide to enroll in other programs or take other steps in response to their knowledge about the program being evaluated, or they may refuse to participate in data collection in a systematic way (Heckman 1999). It is a matter of debate how commonly such conditions prevail; however it is generally agreed that randomization is the preferred evaluation methodology, all things considered (Barnow and King 2000, Orr 1999, LaLonde 1986). Experimental designs are quite common throughout North America, but have only

[4] Some violations of the first condition have been termed "randomization bias" and others are examples of the Hawthorne effect.

recently emerged as viable impact evaluation alternatives in much of Europe and Latin America (Smith, 2000).

While an experimental design is optimal, there can be operational reasons why it is not feasible. First, experimental designs are expensive and time-consuming. It frequently takes a year or more before results are available. Second, there are political and ethical issues associated with denying treatment to eligible needy individuals or households. This will be taken up again in section IV. If adequate advance planning has not occurred, or concession has not been secured from administrators, it may not be possible to implement the random assignment procedure.

Non-experimental designs. If random assignment is ruled out, it is still possible under certain conditions to estimate impacts reliably using non-experimental methods. These can be further divided into two groups, those that primarily address bias due to *observable* characteristics, including multivariate regression models and matched comparison methods, and those concerned with bias from *unobservables*, including reflexive comparison, double difference and instrumental variables methods (Smith 2000).[5]

Multivariate regression is used to account for possible differences between participants and the comparison group on measurable characteristics. The regression framework allows the analyst to focus on one parameter of interest, holding the effects of other variables parameterized in the model as constant or unchanging. The outcome of interest is regressed on an indicator of program participation and all measured personal and environmental characteristics that might affect the outcome. Here, the parameter of interest is the marginal effect on the outcome of participation in the program, netting out the effect of other characteristics. In principle, if all characteristics that affect the outcome could be measured and included in the regression, it would produce an unbiased estimate of the program impact.

The multivariate regression model forms the analytical framework used in most impact evaluations. But the model is never adequate by itself, without additional care taken in the selection of the comparison group or otherwise adjusting for selection bias. Regressions are used in conjunction with random assignment, with matched comparison designs, and nearly all other techniques.

A **matched comparison** is formed by selecting from a pool of nonparticipants individuals or households that are very similar to the participant group using measured characteristics. Matching is done by finding nonparticipants that are similar to participants on key characteristics that could influence program outcomes – educational status, economic background, occupational characteristics, etc. The methods require a national level household or individual data with a rich set of variables that could influence program participation. It is often difficult to know which variables to use in

[5] Different classifications of evaluation methods exist in the literature. Methods that rely on the artificial construction of the comparison group are sometimes labeled "quasi-experimental," while other purely econometric techniques are "nonexperimental." But this terminology is not maintained here.

matching, and to determine how "close" the match should be. A way to surmount these difficulties is through propensity score matching. All observed characteristics are used to predict the likelihood of participating in the program (the propensity score), and nonparticipants are selected for comparison based on how close their estimated propensity score is to each member of the participant group. The impact estimate is then the difference between the mean outcome of the participants and the comparison group mean, or more usually, a regression framework is used to control for other observable factors expected to affect the outcome.

There is considerable debate about the precision and usefulness of matching methods. Some researchers maintain that matching is not a good substitute for an experimental design, while others say that matching can be quite effective under the right circumstances (Orr 1999, Dehejia and Wahba 1999). For example, Heckman (1996) argues that with JTPA data selection on unobservables is empirically less important than other components of bias, and that matching can reduce overall bias satisfactorily. What is clear is that the ability of a matched comparison design to satisfactorily reduce bias depends on the selection of the matching characteristics and the subsample comprising the comparison group. The actual extent of bias is an empirical question that will vary from situation to situation.

Matching methods have been used extensively in safety net and job training evaluations conducted in both developed and developing countries dating from the 1970s. See for example Heckman, LaLonde and Smith (1999), Barnow (1987), and Gerfin and Lechner (2000). The main operational advantage of a matched comparison design is that it is often less expensive and can be executed more quickly than an experiment. It relies on existing data sources and does not require extensive advance planning. Matched comparison groups can be selected either before program implementation or afterwards. The principal disadvantage is that the methodology does not solve the problem of unobservable selection bias.

The **reflexive comparison** addresses bias from unobservable or unmeasurable factors. In this case, program participants are reflexively compared to themselves before exposure to the program. Since individuals are being used as their own comparison group, there cannot be any selection on unobservables such as ability or motivation. It does not solve the problem of isolating program effects from other factors influencing outcomes, however. Unless adequate variables are available to account for all changes in the outcome variable from sources other than the program, such as changes in economic conditions, changes in personal assets, or exposure to other programs, then a reflexive comparison will produce a biased estimate of impact. Such data is usually unavailable, and for this reason reflexive comparisons are not a preferred impact methodology.

The method is useful when it is impossible to establish an external comparison group. It is often used in assessing the effects of a broad policy or set of programs in which there is full participation by the target population (for example, assessing the effect of educational policy on graduation rates or school enrollment). The question of bias remains nevertheless. Implementing the reflexive comparison approach requires panel

data on both participants and comparisons, including at least one period prior to program participation covering all variables of interest.

The **double-difference** (or difference-in-difference, or fixed effects) method attempts to eliminate bias from unobservable characteristics when the comparison group is externally selected. The notion is that if the participants and the comparison group differ from one another primarily in ways that are not measurable, but that these factors do not change over time, these can be subtracted out from an estimate of program impact given panel data. The impact estimate is then the difference between the before-and-after program change in outcome for the participants (the first difference) and the before-and-after program change in outcome for the comparisons (second difference). Operationally, a regression is run with the change in outcomes as the dependent variable and a set of characteristics for each sample member and a dummy variable for program participation as the independent variables. The estimated program impact is the coefficient corresponding to the participation dummy variable. Under the assumption that the ex ante difference between the two groups is due to a fixed unobservable component (like ability) and this determines selection into the program, the double difference method provides a consistent estimate of the overall program impact.

Considerable evidence indicates that the assumption of a time-invariant unobserved fixed effect is not an appropriate assumption for many program settings. The assumption implies that there should be a fixed difference between participants and non-participant comparisons prior to program exposure. Participants in many social programs, particularly labor market and means-based safety nets transfer programs, exhibit a sharp decline in earnings or expenditures just prior to participation (Heckman, LaLonde and Smith 1999). The "pre-program dip" phenomenon is consistent with the view that individuals either decide on their own or are selected into a program based on transitory earnings. Non-participants often do not exhibit this dip, contradicting the view that there is simply a fixed unobservable difference in income/earnings levels. The reasonableness of the fixed effect notwithstanding, the double-difference method has been used extensively in impact evaluations.

Yet another approach to control for bias due to unobservables are the **instrumental variables** methods. Using the regression model, the evaluator would regress the outcome on the indicator of program participation and other variables that affect the outcome. If participants and comparisons differ due to unobservable characteristics and these are related to program participation, then the program indicator variable is correlated with the regression disturbance term, yielding a biased estimate of the coefficient – the program impact. A standard econometric correction for this bias substitutes an "instrument" or instrumental variable for the biased variable. The evaluator needs a variable or variables that are highly correlated with program participation but are not related to the outcome (that is, is not correlated with the disturbance term). This variable is substituted for the program participation variable in the impact regression.

There are two types of instrumental variables corrections used in the evaluation literature, the first much more common than the second. The first method relies on

predicting program participation and using this as the instrumental variable. All the variables that are expected to influence the decision to participate in the program are used in the prediction. The key is to include one or more variables that influence program participation but not the outcome. In practice, it is often difficult to find convincing instrumental variables. In cases where a program is available only in certain geographic regions, location has been used as an instrument on the grounds that the availability of a program should not influence outcomes. A second instrumental variables technique is the so-called Heckman "two-stage" estimator. In the first stage, the probability of participating in the program is estimated as before, and in the second stage the first stage results are used to statistically adjust the disturbance term in the outcome regression so that the impact estimate will be unbiased.

Instrumental variables approaches have been used relatively rarely in impact evaluations, but examples include the Argentina TRABAJAR evaluation, the PROBECAT evaluations conducted in the 1990s, the Bangladesh Food-for-Education evaluation and an evaluation of the US Comprehensive Employment and Training Act in the 1980s.[6]

The foregoing brief overview of quantitative impact estimation methods illustrates several points. First, the main issue to be concerned with is bias stemming from differences between the participants and nonparticipants. Second, addressing bias is not necessarily straightforward, and other than rigorously applied random assignment, there is no single best method in all circumstances. The best impact evaluations use several different methods or combinations of methods to determine impacts.

Analyzing Program Processes and Cost-Benefit Methods

Process studies are among the most important elements of an impact evaluation from the perspective of program administrators. A process study examines aspects of program operations or program environmental factors that may contribute or hinder successful implementation. It can help explain the linkages between program operations, activities, and outcomes (GAO 1998).

Process studies may use quantitative or qualitative techniques and data collection. Often, they include surveys of program administrators and focus group surveys of beneficiaries. They will examine the program operation in detail, from office setup and linkages, through eligibility determination and delivery of services. The study requires the full cooperation of administrators to gain access to staff and program information.

[6] See Ravallion, Galasso, Philipp (2001), Ravallion and Wodon (1998), Wodon and Minowa (1999) and Barnow (1987).

Box 1: To Experiment or Not: The Problem of Bias

Determining program impacts is a matter of accounting for as many of the personal, social, economic and other factors that influence the outcome of interest in order to isolate the effect of participation in the program itself. This is usually addressed by comparing the outcomes of the treatment group with those of a comparison group where the groups are similar to each other in all respects except program participation. The similarity of the two groups in the absence of the program is therefore crucial.

The existing differences between the comparison and treatment groups can result in a biased estimate of program impacts in two ways:

?? *Differences in observable characteristics*. If the treatment and comparison groups are very different from one another on measurable factors such as age, education, or economic status, then it becomes difficult to disentangle the effects of these variables from the participation in the program.

?? *Differences in unobservable characteristics*. There may be differences between the two groups which are not measurable but which are related to the participation in the program. For example, individuals who volunteer to participate may be more highly motivated or of higher ability than others, making them more likely to show positive outcomes even without the program. Resulting differences in the outcome of interest will be attributed to the program while they may be due to the unobservable differences between the groups. This is often called "selection bias."

The only way to eliminate both sources of bias is to randomly assign individuals or households who volunteer to participate into treatment and control groups. This experimental design assures that under weak conditions and with a large enough sample the two groups are statistically similar in terms of unobservables and observable characteristics.

Experimental evaluation designs are expensive, however, and require advance planning and cooperation from authorities. Careful nonrandom selection of the comparison group can significantly reduce the bias from observable characteristics, and adequate data can help reduce the selection bias under certain circumstances. But there is no way to ensure that selection bias has been eliminated, and no way to determine in advance how big a problem this will pose. There is therefore a tradeoff between the preferred methodology of experimental design and the less expensive and more timely application of comparison group strategies.

Sources: Adapted from Baker (2000), Ravallion (1999), and Orr (1999).

An important question regarding any program is whether the costs justify the benefits. Is the program worth the resources, or should alternatives be considered by policymakers? Is the program cost-effective in delivering services? In thorough, large-scale evaluations, cost-benefit assessments are conducted although they are frequently omitted in smaller evaluations without access to resources or adequate data.

There are two ways in which program costs and benefits are compared. Cost-effectiveness analysis estimates inputs in monetary terms and outcomes in non-monetary quantitative terms. By definition, it is a comparative exercise, examining the unit cost of one program versus others. Full cost-benefit analysis estimates both inputs and outputs in monetary terms. Cost-benefit exercises determine whether a program has net benefits to participants and to society. In cases where benefits cannot be quantified monetarily, as is

true with many social programs involving health and education outcomes, cost-effectiveness is used.

Typically, the cost of the program is estimated from administrative data on staff salaries, overhead and operating costs with some estimate of participants foregone earnings or opportunity cost of participating in the program. Benefits are taken from the impact assessment. In principle, calculations are straightforward. All the costs of program operation are subtracted from the benefits of participation. If the result is positive (or the ratio of cost to appropriate unit of benefit is deemed reasonable compared to alternatives), the program is cost effective. In practice, however, there are many operational difficulties, data constraints and measurement issues to be overcome. The principal challenge lies in identifying and measuring all of the costs associated with program participation, including program administration, service delivery, the opportunity cost of participant's time, and the costs of the evaluation itself. Methodological issues arise frequently in the calculations. For example, in assessing a public works program, it is necessary to value the work performed by participants in order to compare it to the next best available use of their efforts and time. This can be done in a variety of ways. Despite the challenges, cost analysis can be a very useful element of an evaluation.

Integrating Quantitative and Qualitative Methods

Qualitative methods can and should also be used to assess impacts. Quantitative techniques attempt to determine whether impacts occurred, but are limited in saying much about why they occurred. Qualitative methods can add rich detail and permit a grounded analysis of the underlying causes of outcomes. They are designed to understand program processes, external conditions, and individual behaviors to provide insight into how participants (program beneficiaries or administrators) perceive that project and how they are affected by it. The methods are open-ended, relying on semi-structured interviews in an individual or group setting and interviewer observations. Participatory techniques involve stakeholders in all stages of the project, from determining objectives to participating in elements of data collection and analysis. Rather than statistical analysis, triangulation is often used to verify data validity and reliability. Triangulation involves the comparison of data sources, collection methods, investigators, and theory to determine the degree of likely biases and the reliability of the information gathered.

Integrating quantitative and qualitative techniques so that each draws on the strengths of the other is one of the challenges and arts of evaluation. Qualitative information can be used to help determine the design of the quantitative analysis, including survey and sample design, and they can help gauge the operation of program process and suggest reforms to enhance effectiveness, among other roles.

Data

Data quality is the most important factor affecting the quality of an impact evaluation. All the econometrics in the world cannot compensate for poor data and unmeasured variables. It is therefore critical to consider the availability of data prior to the evaluation. In practice, it is usually the nature of the data (and available financial resources) that determine the methodology. There are several sources of data that should be investigated. In addition, calculations should be performed in advance to estimate sample size requirements, particularly where new survey efforts are expected.

Measurement. The outcomes of interest must be measured for evaluation to be possible, obviously. But this has two implications for data collection, first that the outcomes of interest must be quantifiable, and second that enough time is allowed to elapse before the final measurement for impacts to have occurred. The first issue is to determine what variables or characteristics need to be measured. This will depend on the type of program being evaluated as well as the methodology. For programs that are designed to improve the income or expenditure of participants or help them find jobs, measurement issues are clear. However, some programs have objectives that are challenging to translate into quantitative variables measurable in a limited time. Measuring the nutritional status of young children, for example, can be a complex undertaking, usually requiring a specialized survey. The time required for impacts to occur can be a tricky issue. If too long a period is deemed necessary, the value of the evaluation for short-term policy needs may be negated.

Existing data. Many countries use nationally representative surveys to track poverty and socioeconomic conditions. For example, more than 30 countries have a version of the Living Standards Measurement Survey (LSMS), and many have surveys for multiple years (Grosh and Glewwe 2000). This data can sometimes be used to provide baseline information and provide a sampling universe for a comparison group, making a separate baseline survey unnecessary. Use of national data can be crucial if the timing of the evaluation does not make advance selection of a comparison group possible. In the evaluation of the TRABAJAR program, it was not possible to construct a baseline sample since the program had already been operating prior to the evaluation. The national socioeconomic survey and propensity score matching techniques were used to statistically choose a comparison group with which to compare TRABAJAR participants. However, this was possible only because of the richness of the Argentine data and the flexibility to incorporate a special module on the program. The quality of the data will largely determine how well impact estimators perform. It has been argued that too much emphasis has been placed in the evaluation literature on selecting impact estimation methods and far too little on the role of data quality in reducing bias in non-experimental evaluations (Smith 2000, Heckman, LaLonde and Smith 1999).

Administrative data can also play an important role in evaluation. The purpose of administrative data is of course to monitor the administration of a program, but it often contains useful information on the outcomes and behavioral characteristics of participants that can be used. Further, administrative information on program costs is crucial to conduct cost-benefit analyses.

Administrative data is used quite extensively in evaluations conducted in developed countries. In the US, the evaluation of state waivers to the old Aid to Families with Dependent Children (AFDC) often used administrative information. For example, university researchers used administrative information on county caseloads to evaluate the California Work Pays demonstration in the early 1990s. The evaluation of Mexico's Progresa program also made heavy use of data on means-testing and program costs for the cost-benefit analysis (Coady 2000). These sources are less comprehensive in developing countries, but there is still considerable scope for the use of project MIS systems and other data.

Surveys. Frequently, new survey efforts are required for a sound impact evaluation. The need for surveys may come from the choice of methodology as noted above, or from the inadequacy of existing data sources. There is often very limited or no information available on the small set of individuals involved in a program, particularly if that program is targeted at a special population group or if the program is being evaluated at the small-scale pilot stage. In an experimental design, special baseline and followup surveys must be administered to the treatment and control groups, giving an opportunity to capture relevant variables. The need for special surveys can be viewed as a positive feature when evaluating programs in low-information environments. Because no alternative data sources exist, surveys must be developed which allow high quality information to be collected on the program in question and which permit the use of a respected random assignment strategy.

Sample size. Regardless of data source, it is necessary to establish appropriate sample sizes for the evaluation. There is a tradeoff between the size of the sample and the cost of data collection, which can strongly influence evaluation design in cases where new survey efforts are required. The sample needs to be of sufficient size to ensure reasonable precision of the resulting impact estimates but not so large that data collection becomes too costly.

Minimal sample size is typically determined through the use of statistical power calculations done during evaluation planning. These calculations relate the statistical precision of an estimate to the sample size and the variance of the outcome variable. Assuming a particular level of precision and the variance of the outcome, the necessary sample size can be estimated. The variance is usually taken from other studies or previous experience. The higher the variance, the larger the sample needed to achieve a given level of precision. The formulas can be adjusted to account for additional regression controls if the data permit.

The size of the desired sample also depends on the outcomes to be assessed, as well as the relative size of the population groups of interest. If the evaluation will examine the effect of a program on an event or condition that is not common among the population (implying the outcome has a low mean and/or high variance), then a larger sample will be required to detect impacts. The large sample sizes in the Bolivia Social Fund evaluation – more than 7,000 households -- were in part dictated by the emphasis on under-five mortality, a rare occurrence within the population. Similarly, if a special vulnerable or

minority group is the focus, larger samples will typically be required although this can sometimes be handled by oversampling and statistical weighting techniques.

Costs

There is considerable variation in the cost of impact evaluations. Information from OED suggests that impact evaluations are generally expensive, but can range anywhere between US$ 200,000 and US$ 900,000 depending on the use of methodology and extent of data collection (World Bank 2002, pg. 21). Another source suggests that the cost of an evaluation is between about 5 and 7 percent of the cost of the program being evaluated as a rule of thumb (W. K. Kellogg Foundation, 1998). From a sample of evaluations conducted in LAC, Judy Baker found that the average cost of the impact evaluations was US$ 433,000, but as a percentage of the loan or credit or total project cost, amounted to only about 0.56 percent. An informal survey of private firms that conduct impact evaluations found that a rigorous evaluation can be conducted in the range of US$ 300,000 – 350,000.[7]

It is difficult to obtain information on the true costs of impact evaluations. Often, larger evaluations are conducted by private firms who do not wish to reveal their costs. But even where information is nominally available, it is difficult to know the total costs or to disentangle different elements. For example, the World Bank and the government of Argentina worked together to evaluate the TRABAJAR program in Argentina, so it would require adding together staff time of Bank personnel and expenditures from the Argentine government. Notwithstanding these difficulties, Table 3.1 estimates the costs of available impact evaluations, broken down into analysis and data collection activities where possible.

The table suggests that there is very little correlation between the cost of an impact evaluation and the size of the program being evaluated. There are three main features which affect the cost of an evaluation.

- ?? Objectives -- the policy questions to be addressed determine how complex the study design must be and consequently how costly. Simpler is usually better.

- ?? Availability of representative socioeconomic data -- if national data exists, alternative methods of selecting a comparison group can be considered which may be cheaper than the ideal random assignment design.

- ?? Timeframe – Evaluators frequently do not have the luxury of months or years for data collection. This lack of time necessitates more creative use of

[7] Based on queries by the author of five international firms that have conducted impact evaluations in developing countries. Two scenarios were considered: a) experimental design of 1,000 treatment and 1,000 control households involving specially administered baseline and followup surveys and b) a nonexperimental matched comparison of the same sample size using national data.

data and more complicated statistical techniques, potentially raising the overall price of evaluation.

The largest single cost is usually analysis. This is frequently because expatriate analysts are employed at high rates while data collection is paid for at local rates. Among the social fund evaluations, data collection accounted for more than a third of the evaluation costs on average. Using existing data can be much cheaper for the evaluation, since it has already been collected. Among the social fund evaluations shown in Table 3.1, Bolivia's is by far the most expensive. The Bolivia evaluation was the only one of the social funds that did not have access to national household survey data, but did use a large health survey to draw comparisons. The evaluation used random assignment and special baseline and followup surveys of the sample, increasing the total cost. But note that large scale socioeconomic surveys are quite expensive to conduct in the first place. An average cost of the LSMS is about $170 per interviewed household, including interviewing and data processing. Using the smaller Core Welfare Indicators Questionnaire (CWIQ) reduces the cost to between $30 and $60 dollars per interviewed household, still expensive. (World Bank 2002). Fortunately, many countries have LSMS data. However, often, the data either doesn't contain the necessary outcome variables or doesn't cover a large enough number of program participants and eligible participants, or both. In such cases, a special survey of the program population will be required.

Table 3.1

Estimated Costs from Selected Impact Evaluations

		Evaluation Costs		
Program	**Methods and Sample Size**	**Data Collection (% of total)**	**Analysis (% of total)**	**Total (US $)**
Nicaragua School-Based Management	Matched comparison of schools, two followup surveys plus focus group surveys (242 schools, 400 teachers, 3,000 students)	35	65	495,000
El Salvador School-Based Management	Random assignment, 200 schools, 2,000 students	59	41	443,000
Columbia Voucher Program	Matched comparison of schools (150 schools, 2,000 students)	69	31	226,000
Armenia Social Fund	Matched comparison group using LSMS and facility survey (2,260 households, 53 facilities)	18	82	111,000
Bolivia Social Fund	Random assignment and matched comparison with baseline and followup surveys (7,000 individuals, 380 facilities)	69	31	878,000
Honduras Social Fund	Matched comparison with pipeline communities, special household and facilities surveys (2,320	32	68	263,000

		Evaluation Costs		
Program	**Methods and Sample Size**	**Data Collection (% of total)**	**Analysis (% of total)**	**Total (US $)**
	households, 81 facilities)			
Nicaragua Social Fund	Matched comparison using LSMS and facility survey (2,000 households, 400 facilities)	56	41	449,000
Peru Social Fund	Matched comparison group with pipeline communities, special surveys (5,120 households, 520 facilities)	Na	Na	350,000
Trinidad and Tobago Youth Training	Tracer surveys, sample from existing national data (2,500)	63	37	238,000
Argentina TRABAJAR Workfare	Matched comparison using national survey, followup survey (2,800 individuals, 120 projects)	90	10	390,000
Bangladesh Microfinance	Matched comparison using national data (1,798 households)	Na	Na	Na
Bangladesh Food For Education	Unmatched comparison using statistical controls from national data (3,625 households)	36	54	140,000
Czech Republic Active Labor Market Program	Matched comparison group using national data and followup survey (9,477)	20	80	250,000
Mexico PROGRESA	Random assignment of localities , also nonexperimental methods (24,407 households)	17	83	2,415,000
Honduras PRAF II	Random assignment of municipalities (4,000 treatment households, 1600 controls – 5,600 households)	Na	Na	Na
Nicaragua Red de Protección Social	Random assignment of census areas with baseline and followup surveys, also qualitative institutional study (773 treatment households, 812 controls)	Na	Na	Na
Zambia Social Recovery Project	Matched comparison group with pipeline communities using national survey with special module followup plus facility survey (2,950 households, 100 facilities)	76	24	174,000
Average from Firms		85	15	350,000

Costs are estimates. For evaluations of World Bank projects, costs do not include those of counterpart teams and resources not financed as part of the loan or credit. *Sources*: Baker, 2000; Rawlings and Rubio 2002, Fretwell et. al. 1999; Jalan and Ravallion 1999; other project files.

Estimates of special surveys of a sample of the program population and comparison or control group also vary depending on survey costs in the country, the complexity of

the survey effort, and the desired sample size. The large evaluation of the Progresa program in Mexico had survey costs of about US $17 per completed questionnaire.

IV. Will an Impact Evaluation be Conducted?

Impact evaluations can answer some important questions about the effectiveness of program interventions, but they can also be relatively expensive to conduct. Are they worth the effort and money? Under what conditions should a program or intervention be evaluated? And as importantly, will an evaluation be undertaken? Whether an evaluation is undertaken will depend on a combination of factors, including an assessment of the raw policy value of the exercise and the political economy associated with the policy environment.

In Theory

If any of three basic questions can be answered affirmatively, a rigorous impact evaluation should be considered[8]:

- ?? Is the program considered to be of strategic relevance for national public policy?
- ?? Can the evaluation results influence the design of the program?
- ?? Will the evaluation contribute to improving the state of knowledge about a type of program or policy and does the information generated have potential future research value?

These three questions are at the heart of the technocratic decision of whether to evaluate and they correspond to the principle benefits attributed to impact evaluations.

Policy impact. The greatest motivator for evaluation is the desire to reform or validate a program. If the evaluation can affect policy, it can largely be justified if it is also a strategic intervention. For example, the evaluations of the Honduras and Nicaragua social funds are translating directly into policy and program reform. The Nicaragua social fund has suspended the financing of new sewerage projects temporarily and will finance more integrated infrastructure projects and is strengthening its own project appraisal and monitoring and evaluation capacities. In Honduras, impact evaluation results have led to redesigning the criteria for supporting water systems, developing baseline data on incoming projects to facilitate future impact assessments, strengthening subproject supervision, and ensuring more systematic consultation with beneficiary communities. (SIF 2000 report). While not specifically safety nets programs, the evaluations serve to illustrate the policy value of impact evaluations. In Bangladesh, the evaluation of the Rural Food Rationing Program found significant leakages of benefits to the nonpoor and other inefficiencies which led to the abolition of the program.

[8] There are many other formulations of the questions that guide an evaluation decision. See for example Wholey et. al. (1994) and Prennushi et. al. (2001). Most can be represented as special cases of the three questions presented above.

It was estimated that the action resulted in a budget savings of about US $60 million per year. (Babu 2000).

Box 2: Improving Program Performance and Beyond -- The U.S. National JTPA Study

The National Job Training Partnership Act (JTPA) Study was commissioned in the mid 1980's by the US Department of Labor to study the effectiveness of programs funded by the Job Training Partnership Act. The evaluation is one of the largest, most comprehensive of its kind ever undertaken. Some 20,000 program applicants from across the country were included in the experimental design to estimate impacts on earnings, employment and welfare receipt of individuals served by the programs. The evaluation was commissioned in 1986 and results were publicly released in 1994.

Among the findings of the study were that JTPA had very different effects for adults and for youth. For adults, the program was successful in raising earnings of participants by 7 to 11 percent and provided benefits of about $1.50 for every $1.00 invested. However, the program had no significant impact on earnings for youth and costs exceeded benefits to society.

The US Congress reduced the budget for the JTPA youth component by more than 80 percent and the budget for the adult component was increased by 11 percent following the release of the evaluation findings. The evaluation directly influenced the design of the program and saved taxpayers over US$ 500 million.

Beyond the direct effects on the JTPA program, the study has yielded longer-term benefits in improved knowledge and basic research. The rich data set produced by the JTPA study has been used by academic researchers and others to study a range of topics from different aspects of job training interventions to evaluation methodology itself.

Source: Orr 1999, author.

Expanded knowledge. In addition to the local program benefits of impact evaluation, there can be a significant contribution to general knowledge on the effectiveness of social programs. An informed understanding of the likely effects and appropriateness of various public interventions can be gained only through the accumulation and synthesis of rigorous evaluation findings. For example, there is a large body of evaluation results stemming from the various welfare-to-work experiments conducted by the states in the 1970's-1990's in the United States. Collectively, these evaluations can inform policymakers elsewhere about the potential benefits and limitations of these types of programs in their country settings and contexts. Work done by the Manpower Demonstration Research Corporation and others in the US has helped to synthesize some of the early findings and make them available to a wider audience.[9] Another example is the effort underway by researchers at the World Bank to assess lessons from recent and ongoing evaluations of Conditional Cash Transfer Programs in Latin America.[10]

[9] See for example Gueron and Pauly (1991) and Barnow and King (2000).
[10] Rawlings and Rubio (2002).

In Practice 1: Constraints on the Use of Evaluation

The process of the evaluation itself has repercussions for programs and policy formation that can also limit their use. These repercussions reflect the perceived inadequacy of results and the possible interference to the operation of the program through the evaluation in terms of resources and program operations.

Limitations of Resources. As noted earlier, evaluations can be expensive. Governments frequently cannot justify the high costs of evaluating programs which rely on scarce resources. The fact that in developing countries evaluations will be frequently funded from loans makes evaluation costs harder to justify. Further, donors typically are not encouraged to spend significant resources on non-lending activities.

Program Effects. Frequently, there are concerns that the evaluation itself interferes with the operation of the program. This is most often heard in connection with random assignment designs, which involve significant staff time and interaction with eligible participants. It is argued that ethically, agencies cannot randomly deny benefits or services to eligible needy individuals. In the US National JTPA Study, there was a major concern among training centers about negative publicity from using random assignment (Smith 2000). Nearly 200 training centers had to be contacted to find 16 that were willing to participate in the study.

Limitations of Results. Despite good planning and adequate resources, impact evaluations can still lead to ambiguous policy implications. Three main issues are at stake: (i) non-replicability of results; (ii) untimely research findings; and (iii) unanswered policy questions.

The fact that evaluations can produce ambiguous or even conflicting results depending on data sources and methodology casts a shadow on their use for policy. This problem is also known as lack of replicability, and is most commonly experienced with non-experimental evaluation designs. Several non-experimental evaluations of the Comprehensive Employment and Training Act in the United States produced widely varying estimates of impacts, for example, and fueled the debate on the use of non-experimental methodology. Two additional examples from Latin America highlight the policy dilemma. In Peru, there were two separate evaluations of the Social Fund, completed a year apart. Using different methodologies and data, they arrived at opposite conclusions on key impacts.[11] The earlier evaluation used national data and an instrumental variables approach to determine targeting and district level impacts of social fund projects. The later evaluation implemented a special household survey and used a nonexperimental matched comparison group design from communities and individuals who were approved for future funding from the program (pipeline communities). The two evaluations reached conflicting conclusions, particularly on the impact on education enrollment.

[11] Rawlings, Sherburne-Benz and Van Domelen (2000).

A second example comes from Mexico. The national Programa de Becas de Capacitacion para Trabajadores (PROBECAT) provides short-term training and job placement for unemployed workers. During the 1990's, two evaluations were conducted.[12] Both used the same data sources and same initial sample. The first found positive impacts of the program on employment and wages and concluded that PROBECAT benefits exceed program costs. The second evaluation, however, found no positive effects and an unfavorable cost-benefit assessment.[13] How is this possible? The second evaluation used a slightly different methodology to develop the comparison group than the earlier study and used additional controls for selection bias in both program participation and the decision to work. The estimated program impacts turn out to be quite sensitive to such seemingly small methodological differences. In the face of such sensitivity, it is difficult for anyone, let alone policymakers, to know what to conclude regarding the effectiveness of the program.

Lack of timely results is another criticism of rigorous impact evaluation. It is argued that the longer it takes to conclude an evaluation, the less likely the findings will directly influence policy. The evaluation of the Bolivian social fund lasted nearly a decade – the study was designed in 1991, baseline data was collected in 1993 with a followup household survey administered in 1997-98. Initial impact findings were not available until late 1999. However, there are exceptions, such as the JTPA evaluation in the US. The evaluation took nearly 8 years, but resulted in a major overhaul of the program (see Box 2).

Finally, a frequently cited criticism of impact evaluations is that they leave important policy questions unanswered. Questions such as how the impacts might change if the program design were changed – the selection criteria for participation, program compliance rules, level of benefits, or upscaling the program – cannot be addressed, nor can assessments of complementarities or substitutions between program components and policies. This criticism often is raised in connection with the cost of evaluations. They are expensive, so why can't they answer all questions? Despite its demonstrated policy value and continued use for research, this charge has been levied against the JTPA Study in the US on several occasions, for example.

A response. Two points should be borne in mind when considering both the criticisms of un-timeliness and un-answered policy questions. First, rigorous impact evaluations are not designed to be quick policy assessments. They are time-consuming primarily because they are attempting to determine real impacts on program participants, and it takes time for these to be observed in the data. Second, all policy studies are open to the criticism of not addressing every question that might be of interest. It is important to determine up front which are the most important prior to undertaking the evaluation, as discussed in section II. One of the legacies of good evaluations is the data they leave behind. The use of the US JTPA data has already been mentioned. The rich data from the Progresa program and evaluation have similarly been used to address a range of

[12] There were actually three evaluations, but the first two produced the same general results using the same but updated data. We refer in the text to the second evaluation done in 1995 and that done in 1999.
[13] See Wodon and Minowa (1999), STPS (1995) and Baker (2000).

policy questions. The conduct of a formal evaluation does not preclude other studies that address particular issues related to the program. The recent ex-ante simulation of potential effects from program changes to Brazil's Bolsa Escola is a case in point.

The most troubling criticism is the non-replicability of results a la the PROBECAT and Peru Social Fund examples. However, these examples only reinforce the point that evaluation design is critical. The best evaluations are those that use an experimental design and/or use a variety of methods to estimate impacts, thereby providing an assessment of the robustness of the findings. The Bolivia Social Fund evaluation used nearly every quantitative method available to estimate impacts, including randomization, matching, reflexive comparisons, double differences and instrumental variables. Similarly, the Progresa evaluation and the US JTPA Study both used randomization and various nonexperimental methods, as well as qualitative information. Sensitivity of impacts to the data and choice of estimator is not a sufficient reason not to undertake impact evaluation.

Box 3: What are the Key Design Features of a Good Impact Evaluation?

To provide the highest value, an impact evaluation should include:

?? Clear objectives. Evaluation questions should be determined early, should be simple and measurable.

?? Credible evaluator. The evaluator should be independent of the agency or institution whose program is being evaluated.

?? Rigorous methodology. Ideally should include an experimental design or a well-chosen matched comparison group.

?? Adequate sample size. The sample should be large enough to detect program effects of plausible size. In addition, the size should permit assessment of program impacts on key subgroups of the target population, as appropriate to the program. Minimum detectable effects should be determined prior to the implementation of the evaluation.

?? Baseline data. Need to establish the appropriate comparison group and to control for observable program selection criteria.

?? Sufficient followup. Followup data should be collected after enough time has passed to plausibly detect an impact, and should measure the relevant outcome variables.

?? Multiple evaluation components. The impact evaluation should do more than detect program effects. It should also examine program process, reasons for observed outcomes, and cost effectiveness.

Source: author, Ezemanari et al . 1999

In Practice 2: Political Economy of the Policy Environment

Politics and political economy play an important role in the decision of whether to conduct an evaluation of a program. Some of these aspects act to increase the likelihood of undertaking evaluation, but most serve to deter evaluation. The essential issues can be characterized as stemming from principal-agent problems, where stakeholders do not have an incentive to support an evaluation. Stakeholders refer variously to the funding

agency (government or donor), the implementing agency, the program beneficiaries, and the public.

Political costs and incentives. Unfortunately, there are many potential costs perceived by stakeholders. Program funders and administrative agencies alike see the risk of negative evaluation findings to terminate the program and hurt careers. In Indonesia, attention is being paid to developing evaluation and monitoring capacity in government. Public officials express concern about the risk of being seen as conveyers of bad news. The value of evaluation for improved general knowledge can be large, however. Yet as with any public good, the individual policymaker or agency may not have adequate incentive to undertake a specific evaluation. This can be a good reason to subsidize evaluation efforts. World Bank and others help, but more could be done in developing countries. Incentives within the donor organizations that are more favorable to evaluation and assessment would be beneficial in this regard.

Political benefits. One of the main reasons formal impact evaluations are undertaken is to gain political support for a program. This is particularly true for programs that are seen as strategically important for national policy, or for programs that are introducing innovative approaches. The Mexican government paid for the evaluation of Progresa in part because the conditional cash transfer model was relatively new and was viewed as a potential replacement for certain subsidy programs. Support from within government and among the public was needed to expand the approach. In the early days of the Women Infant and Children (WIC) nutrition program in the United States, a random assignment evaluation with cost-benefit assessment was carried out to assess the potential for national expansion. The Head Start was similarly evaluated. NGOs and charity foundations may also evaluate their programs in order to demonstrate their own legitimacy for funders in governments or international organizations. For example, the W.K. Kellogg foundation has produced its own handbook on evaluation for its funded projects.

Establishing an Evaluation Culture

For an evaluation to have an effect on policy it is necessary to have both a good supply of evaluation ingredients and an enthusiastic demand for the evaluation product (OECD 1999). Most of this paper has discussed the supply ingredients. But securing demand for evaluation should not be overlooked. This section briefly looks at some of the factors influencing the demand for evaluation generally, with reference to country experiences with attempts to institutionalize evaluation practices.

Convincing Stakeholders. In any setting, whether a developed country with a tradition of evaluation or a small developing country with little previous exposure to program evaluation, it is necessary to convince the program's stakeholders of the value of the evaluation. The key stakeholder in this regard is the program's funding agency or institution. But in many cases, it is the Ministry of Finance or the donor institution that is in the position of advocating for evaluation. Ryan speaks of the importance of identifying

"policy entrepreneurs" who can assist in the advocacy and outreach needed to support an evaluation.[14]

A principal method of convincing stakeholders is to present evidence regarding what is involved in impact evaluation and what can be expected. This is especially true among developing countries with limited evaluation experience. Often, an international donor is in the position of convincing a country to evaluate a program. An advantage of an experimental design is that there is less controversy regarding impact estimates than with non-experimental designs. It is also relatively straightforward to understand the concept of random assignment, making the method easier to explain to policymakers.

Frequently, money is an issue, especially in developing country contexts. Countries typically do not want to borrow in order to finance an expensive evaluation. In Russia, for example, a planned evaluation of a labor market program was recently cancelled when the government decided not to borrow the estimated half million dollars needed to fund it. This points to a role for international donors and financial institutions in financially supporting evaluation efforts. This approach has been followed with some success by institutions like the World Bank and the Inter-American Development Bank, where evaluations have been explicitly included in loan packages and the costs deferred by contributing staff expertise.

Failure to get sufficient support from key stakeholders can have less obvious, but serious consequences for the quality of an evaluation. For example, the inability to convince the employment agency of the value of the evaluation severely compromised the survey instrument in the impact evaluation of active labor market programs in the Czech Republic.[15] For the evaluation to be implemented well and actually used, it is important that all stakeholders are involved. These include other policymakers, program managers, and civil servants implementing the program. Some proponents of evaluation advocate the use of participatory evaluation methods as a way to involve a wider range of stakeholders in the process.

Stakeholders can be convinced by the demonstration effect of actually doing an evaluation. This may operate as more of a forced approach to reluctant stakeholders, but funding agencies can make evaluation mandatory. Some governments have followed this strategy, imposing the evaluation requirement from a level of authority that cannot be challenged by the program proponents. For example, prior to the overhaul of the welfare system in the United States, individual states were allowed to develop pilot programs that differed from the standard national program rules. Encouraging innovative welfare programs at the state level was viewed as an important step in creating an improved system, and stood to benefit the states that undertook them in terms of benefit savings and increased efficiency. However, in order to receive these waivers, the federal government required that states had to agree to a formal impact evaluation of the pilot program. Further, the evaluation had to be conducted by an independent, openly bid contractor. The use of evaluation in the welfare arena has increased the acceptance and expectation

[14] Ryan, 2002.
[15] Baker, 2000.

of evaluations in other areas. The US government has been one of the largest consumers of formal evaluations in the world, and a rather large evaluation industry developed as a result.

Transparency and Communication. Communicating findings is very important to help gain support for the evaluation and ensure that the results are actually used for policy. There are many strategies to disseminate findings, and the appropriate strategy will depend in part on how large and how politically important the evaluation is perceived to be. An impact evaluation of a national program might involve presentations of findings to parliament or congress and a public media campaign. Smaller efforts may involve dissemination throughout a ministry, to donors and civil society groups, and presentations at conferences and administrative forums.

Box 4: Institutionalizing Evaluation

Examples of countries' efforts to institutionalize evaluation, including impact evaluation, illustrate the link between evaluation and the larger mandate of improving public sector management:

Sweden. Evaluations were first undertaken in the 1950s by public commissions preparing policy decisions. The commission system combined the views of stakeholders and introduced research findings into decision-making. From the 1960s evaluation began to be viewed as an activity that could continuously provide decision-makers with information. Specialised research bodies and agencies were founded with evaluation as their main task.

Australia. Evaluation became an integral part of Australia's public management reform process in the mid-1980s when the government's evaluation strategy was launched. Evaluation has been systematically integrated into corporate and programme management and planning. All public programmes, or significant parts thereof, are reviewed once every three to five years; all major new policy proposals include an evaluation strategy, ministries are currently required to provide an annual evaluation plan; and results of major evaluations are expected to be made public.

The European Commission has for some time had institutionalised evaluation practices in certain central policy areas of Community policy, such as development aid, research and technology policy and programmes financed through the Structural Funds. A policy of systematic program evaluation was adopted in 1996. An important link between budgeting and evaluation has been created through the requirement that all new program proposals must be based on an exante evaluation and are accompanied by an evaluation plan. There is "no appropriation without evaluation."

Indonesia. A Steering Committee was established in 1994 to oversee the development of a national strategy and framework for performance evaluation. Evaluation efforts focus on two separate tracks. The first is the formulation of performance indicators for development projects. The second is carrying out evaluation studies of development projects, programs and sectors. Success so far in establishing evaluation capacity has been attributed largely to the sustained efforts of a few key senior public servants and the stance they took to promote both the supply and demand for performance evaluation.

Chile. The first initiative involved developing a system of performance indicators for government programs. By 1994 the Budget Directorate requested indicators on performance from all government agencies. The second component is program evaluation, focusing on cost, timeliness and feedback to decisionmaking. There has been political agreement to evaluate all programs in Chile. This clear target has helped to deal with initial resistance by agencies.

Source: OECD 1999, World Bank 1998.

Most countries with well-established evaluation cultures have policies of open communication and availability of results. In Sweden, the public has traditionally had access to all government material including evaluation reports. In the US as well, evaluation findings are generally available through the relevant government agencies and on websites, and are frequently covered in the media. In Mexico, the results of the Progresa evaluation have been widely circulated domestically and internationally. Even in less developed countries, communication of findings has been a policy. Columbia, for example, has commissioned independent evaluations of key government projects and has actively promoted dissemination of findings to the public (Guerrero, 1999).

One of the problems frequently encountered with evaluation is that program implementers and proponents do not support the findings, particular if negative, and hinder their use to reform a program. This response can be mitigated if these stakeholders have been involved in the evaluation at any early stage and if results have been shared with them before public dissemination. An example of effective communication can be taken from the Bangladesh Rural Rationing program evaluation. The preliminary negative findings of the operational performance study were first shared with administrators and study collaborators in government through several in-house seminars late 1991. The evaluation researchers used the close working relationship with government to share their research-based information on the program to higher level policymakers, including the Minister of Finance and the Cabinet. By the spring of 1992, the decision was made to abolish the Rural Rationing program. Subsequent review of the experience revealed three key features of the evaluation process that enhanced its ability to influence policy: (i) the research offered quantitative parameters and results, outlining specific courses of action; (ii) the evaluation team facilitated the use of information by collaborating closely with policymakers and operating within the decision framework; and (iii) the results were timely in responding to the need for information.[16]

Institutional Capacity. In the long run, the key to establishing an evaluation culture in a country is to develop the capacity of institutions, particular government funding agencies and line ministries. Many of the countries where evaluation has taken hold are those that have put a priority on public sector performance management. Where governments see a need to improve public sector performance, monitoring and evaluation of programs is usually accepted. In Chile, for example, the new government in the early 1990's recognized the need for improving the public sector and embarked on a concerted effort to inculcate a culture of monitoring and evaluation. The first step was to develop a set of performance indicators for each program, and then to focus on systematic monitoring and evaluation. There has been general political agreement to evaluate all programs over time.[17] Additional detailed examples of countries efforts to institutionalize evaluation as part of a larger public sector management mandate can be found in Box 4 below and in various Operations Evaluation Department (OED) publications from the World Bank.

[16] See Babu, 2000.

[17] It should be noted that most of these evaluations will not be full impact evaluations as described in this paper, but will include important elements such as process studies.

Again, donors can play an important role in developing institutional capacity. Donor-supported evaluations should incorporate capacity building within the exercise, making the evaluation part of the overall project rather than simply an ex-post procedural requirement. While independent evaluation is to be encouraged, too often evaluations are conducted largely by international consultants who leave the report with the donor and government with little transfer of knowledge or followup. There have been few examples of successful donor capacity building in this regard. In Indonesia, the World Bank helped establish the Social Monitoring and Early Response Unit to monitor the country's response to the Asia Crisis. It has subsequently become an independent NGO specializing in analyzing and evaluating policies and programs. Establishing evaluation capacity and advancing the role of evaluation in decision-making and social science knowledge can be further be promoted through conferences, training courses and technical assistance activities, setting up newsletters and journals on evaluation and policy.

VI. Conclusions

There are a number of conclusions that emerge from this review. First and foremost impact evaluations should be undertaken strategically. Not all programs should be evaluated rigorously. Full impact evaluations are relatively expensive, and only those programs that are of strategic relevance for policy, that will improve the state of knowledge about the intervention, or that can directly influence program design should be considered. The political economy of the associated institutions and stakeholders must be taken into account. Among the main conclusions:

Planning is crucial. Impact evaluation cannot answer all questions. A clear statement of measurable objectives is required. Estimate costs and establish a timeframe for the evaluation early in the planning process. Examine existing data sources to determine the best methodology. Experimental designs are the best method of obtaining accurate estimates of impact, but may not be suitable to the situation at hand.

Gain the support of policymakers and program administrators under evaluation to help ensure success. Cooperation and buy-in are needed to obtain data, to set up an experiment, to pay for the evaluation, and to trust the findings in order to incorporate them into future policy and program reform. An open and transparent communication process is essential.

Donors and international financial institutions have a role. For cost and political economy reasons, impact evaluations may not be undertaken in developing country contexts without the support of donors. These donors can help defer the costs of financing as well as enhance the institutional capacity of the client government and help to establish a culture of evaluation.

Impact evaluations can provide unique information on the efficacy and value of social programs. Judicious use can help in the formulation of sound social policy and expand the state of knowledge about what helps the poor and vulnerable.

References

Babu, Suresh. 2000. "Impact of IFPRI's Policy Research on Research Allocation and Food Security in Bangladesh." Impact Assessment Discussion Paper No. 13. International Food Policy Research Institute.

Baker, Judy L. 2000. *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*. Directions in Development Series. Washington DC: The World Bank.

Bamberger, Michael. 2000. "The Evaluation of International Development Programs: A View from the Front." *American Journal of Evaluation*. Vol. 21, No. 1: 95-102.

Barnow, Burt S. and C. T. King, eds., 2000. *Improving the Odds: Increasing the Effectiveness of Publicly Funded Training*. Washington, DC: Urban Institute Press.

Barnow, Burt S. 1987. "The Impact of CETA Programs on Earnings: A Review of the Literature." *Journal of Human Resources*. 22(2): 157-193.

Bassi, Laurie J. 1995. "Stimulating Employment and Increasing Opportunity for the Current Work Force." In *The Work Alternative: Welfare Reform and the Realities of the Job Market*. Eds. D.S. Nightingale and R. H. Haveman. Washington DC: The Urban Institute.

Bell, Steve, L. Orr, J. Blomquist, and G. Cain. 1995. *Program Applicants as a Comparison Group in Evaluating Training Programs.* Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

Bourguignon, Francois, F.H.G. Ferreira, and P.G. Leite. 2002. "Ex-ante Evaluation of Conditional Cash Transfer Programs: The Case of Bolsa Escola." The World Bank: Washington DC. Processed.

Coady, David P. 2000. "Final Report: The Application of Social Cost-Benefit Analysis to the Evaluation of PROGRESA," Report submitted to PROGRESA. International Food Policy Research Institute: Washington, DC.

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

Dehejia, Rajeev and S. Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*. 94(448): 1053-1062.

Ezemanari, Kene, G. Rubio, A. Rudqvist, and K. Subbarao. 2001 "Impact Evaluation: A Position Paper." Poverty Reduction and Economic Management Network, World Bank. (see "Good Practice Examples" link on PREM Impact Evaluation site)

Ezemanari, Kene, A. Rudqvist, K. Subbarao. 1999. "Impact Evaluation: A Note on Concepts and Methods." Mimeo, Poverty Reduction and Economic Management Network. World Bank.

Fretwell, David H., J. Benus, C. J. O'Leary. 1999. "Evaluating the Impact of Active Labor Programs: Results of Cross Country Studies in Europe and Central Asia." World Bank Discussion Paper. World Bank, Washington D.C.

Friedlander, Daniel, D. H. Greenberg, P. K. Robins. 1995. "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods." *American Economic Review* 85 (September): 923-37.

Greenberg, David and M. Shroder. 1991. "Digest of the Social Experiments." Institute for Research on Poverty. Special Report No. 52. Madison, WI: University of Wisconsin.

Grosh, Margaret and P. Glewwe. 2000. "Chapter 1: Introduction." In *Designing Household Survey Questionnaires for Developing Countries: Lessons from 15 Years of the Living Standards Measurement Survey*. Margaret Grosh and P. Glewwe, eds. The World Bank.

Gueron, Judith and E. Pauly. 1991. *From Welfare to Work*. New York: Russell Sage Foundation.

Guerrero, R. Pablo. 1999. "Comparative Insights from Columbia, China and Indonesia." In *Building Evaluation Capacity.* R. Boyle and D. Lemaire, eds. New York: Transaction Publishers.

Heckman, James J. 1991. "Randomization in Social Policy Evaluation." In *Evaluating Welfare and Training Programs*, Charles Manski and Irwin Garfinkel, eds. Cambridge, MA: Harvard University Press.

Heckman, James J., R. LaLonde and J. Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics, Volume 3A*, Orley Ashenfelter and David Card, eds. Amsterdam: North- Holland: 1865-2097.

Heckman, James J. and J. Smith. 1997. "The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study." NBER Working Paper 6105 (July). National Bureau of Economic Research, Cambridge MA.

Heckman, James J. and J. Smith. 1995. "Assessing the Case for Randomized Social Experiments." *Journal of Economic Perspectives*, No. 9: 85-110.

Jalan, J. and M. Ravallion . 1999. "Income Gains to the Poor from Workfare: Estimates for Argentina's Trabajar Program," mimeo, Development Research Group, World Bank, Washington D.C.

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

Manski, Charles F., and I. Garfinkel. 1991. "Issues in the Evaluation of Welfare and Training Programs." ." In *Evaluating Welfare and Training Programs*, Charles Manski and Irwin Garfinkel, eds. Cambridge, MA: Harvard University Press.

Manski, Charles F. 1990. "Where We Are in the Evaluation of Federal Social Welfare Programs." Institute for Research and Poverty, University of Wisconsin-Madison. *Focus* Vol. 12 No. 4 (Fall): 1-5.

Morduch, Jonathan. 1998. "Does Microfinance Really Help the Poor? New Evidence From Flagship Programs in Bangladesh." Mimeo, World Bank, Washington D.C.

Newman, John, L. Rawlings and P. Gertler. 1994. "Using Randomized Control Designs in Evaluating Social Sector Programs in Developing Countries." *The World Bank Research Observer* 9(2): 181-201.

Organization for Economic Cooperation and Development. 1999. "Improving Evaluation Practices: Best Practice Guidelines for Evaluation and Background Paper." OECD Public Management Service PUMA/PAC(99) 1.

Orr, Larry L. 1999. *Social Experiments: Evaluating Public Programs with Experimental Methods*. Thousand Oaks, CA: Sage Publications.

Prennushi, G., G. Rubio and K. Subbarao. 2001. "Monitoring and Evaluation," in the PRSP Sourcebook.

Ravallion, Martin. 1999. "The Mystery of the Vanishing Benefits: Ms. Speedy Analyst's Introduction to Evaluation." Mimeo, Development Economics Research Group. The World Bank.

Ravallion, Martin, E. Galasso, T. Lazo and E. Philipp. 2001. "Do Workfare Participants Recover Quickly from Retrenchment?" Mimeo, Development Economics Research Group. The World Bank.

Rawlings, Laura B. and G. M. Rubio. 2002. "Evaluating the Impact of Conditional Cash Transfer Programs: Lessons from Latin America." Draft, Human Development Network, Latin America and Caribbean region. The World Bank.

Rawlings, Laura, L. Sherburne-Benz, and J. Van Domelen. 2002. *Evaluating Social Fund Performance: A Cross-Country Analysis of Community Investments.* Draft. The World Bank.

Rossi, Peter H. and H. E. Freeman. 1993. *Evaluation: A Systematic Approach*. Fifth Edition. Newbury Park California: Sage Publications, Inc.

Rubio, Gloria and K. Subbarao. 2001. "Impact Evaluation in Bank Projects: A Comparison of Fiscal 1998 and 1999." Mimeo, Poverty Reduction and Economic Management Network, World Bank.

Ryan, James G. 2002. "Synthesis Report of Workshop on Assessing the Impact of Policy-Oriented Social Science Research in Scheveningen, The Netherlands, November 12-13, 2001." Impact Assessment Discussion Paper No. 15. International Food Policy Research Institute.

Smith, Jeffrey. 2000. "A Critical Survey of Empirical Methods for Evaluating Active Labor Market Policies." Mimeo, Department of Economics, University of Western Ontario.

STPS (Secretaria del Trabajo y Pervision Social). 1995. *Evaluación del Programa de Becas de Capacitacion para Desempleados*. México DF.

Wholey, J.S., H.P. Hatrey, and K.E. Newcomer. 1994. *Handbook of Program Evaluation*. San Francisco: Jossey-Bass Publishers.

W. K. Kellogg Foundation. 1998. *Evaluation Handbook*. Battle Creek: Michigan.

Wodon, Quentin and M. Minowa. 1999. "Training for the Urban Unemployed: A Reevaluation of Mexico's Probecat." Background paper for Mexico Poverty Assessment. Latin America and Caribbean Region, World Bank.

World Bank. 2002. *Monitoring and Evaluation: Some Tools, Methods and Approaches*. Operations Evaluation Department, Washington DC: World Bank.