4 PROGRAM EVALUATION: CONCEPTS AND METHODS

There is a growing awareness regarding evaluation of development programs to assess impact on desired outcomes. Evaluation is a powerful tool for learning about what works and what does not. There is an extensive literature now available on appropriate methodologies. This chapter presents a short summary of the concepts and methods of program evaluation.

4.1 INTRODUCTION

Interventions and social programs by public sector call for huge amounts of resources that can be spent for alternative programs. The costs of these interventions are especially higher among the resource scarce developing economies. It is, therefore, imperative to evaluate the costs and benefits of these programs from time to time and assess their impact on the presumed outcomes. As Heckman et. al. (1999) put it, “An emphasis on objective publicly accessible evaluations is a distinctive feature of the modern welfare state, especially, in an era of limited funds and public demands for accountability.”

The literature on evaluation methods in economics is vast and continues to grow (Blundell and Dias, 2002). The initial evaluations in economics draw heavily from the labour market and safety net literature, with seminal works by Ashenfilter (1978), Ashenfelter and Card (1985) and Heckman and Robb (1985).

In this chapter, we put forth the evaluation problem and different approaches that have evolved in literature and empirical studies to address this issue. Section 4.2 lays out the various conceptual aspects of program evaluation. In section 4.3, we present the formal evaluation problem. Section 4.4 discusses the experimental and non-experimental studies and econometric tools to estimate effects. We present a summary in section 4.5.

4.2 ASPECTS AND ELEMENTS OF PROGRAM EVALUATION

4.2.1 What is Program Evaluation

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, rapid appraisals and beneficiary assessments and indicator monitoring. Program evaluation has also been used a tool for doing a cost-benefit analysis of development interventions. Though the objectives of an evaluation program could differ, however, each is an exercise to measure some aspect of the program. We find that some of the programs are mainly to access impact and are called impact assessment.

The fundamental aspect of an impact evaluation is an assessment of the impacts on participants that can be attributed to direct participation in a program or intervention. It tries to determine whether the program as implemented achieves its goals. Any impact evaluation would aim to answer the following type of questions about a program:

- Does the program or intervention achieve the stated goals?
- Does it have unintended effects on participants?
- Do program impacts differ among individuals?
- Is the program cost effective as compared to other options?
- Reasons for success or failure, as the case may be
- What can be remedial measures or recommendations to improve the performance of the program?

Conducting an impact assessment requires a plan that (i) establishes evaluation objectives (ii) determines appropriate evaluation methods (iii) provides a data collection strategy and identify available sources and (v) establish a timeline for producing and disseminating findings.

4.2.2 Objectives and Outcomes

The principal objective of any evaluation is always to determine whether the program is helping beneficiaries or not. Also, it is important to know whether the program has targeted the right set of people or not. This is especially true for developing economies where resources are relatively scarce and leakages are greater. Evaluation objective should consist of the desired policy questions to be answered balanced against the likely resource constraints, including time, money and data availability. The statement of the objective should 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, in order to assess impact of a health and education program, clear outcomes in terms of grade school annual attendance, annual sick leave, immunization rates etc should be specified outcomes.

Programs also need to be assessed in terms of directly targeted beneficiaries as well as the ones that are affected indirectly. This would involve estimation of effect by taking into account the externalities generated by programs.

4.2.3 Program Costs

Programs usually entail very high costs to the economy. In developing economies, where resources are relatively scarce, an assessment of efficiency of program becomes imperative. For the economic evaluation of policy interventions, valid estimates of the costs are important. There has not been as much empirical work in estimating costs or efficiency of programs as has been in estimating effects or effectiveness or efficacy of interventions. Also, most of the times only direct costs are taken into consideration, while the indirect costs are neglected, partly due to the problems involved therein. Schmidt (1999) suggests that the full costs of program participation should include three components: the time cost of participation, the time cost of delivery of program; and the financial costs of the program.

4.3 THE EVALUATION PROBLEM

In recent years, the evaluation literature in statistics and econometrics has developed a unified formal framework that facilitates the exploration of the potential and the limits of both experimental and non-experimental evaluation strategies, following Rubin (1974).

Any impact evaluation attempts to answer essentially counterfactual questions: How would individuals who did benefit from the program have fared in the absence of the program? How would those who did not benefit have fared if they had been exposed to the program? (Duflo 2001). The problem in answering this question is that the same individual can not be observed simultaneously on both the situations. An individual is observed either exposed to the program or not exposed. Comparing the same individual overtime, in most cases, will not give reliable estimates of program impact as many things would have changed. We,
therefore, cannot seek to obtain an estimate of the impact of the program on each individual. The only possibility is to be able to obtain the average impact of the program on a group of individuals by comparing them with a similar group that was not exposed to the program.

Thus, the critical objective of impact evaluation is to establish a credible comparison group, who, in the absence of the program, would have had outcomes similar to those who were exposed to the program. The average effect on individuals of a certain type would be population average treatment effect (ATE), which would be the outcome if individuals were assigned to treatment, and the effect of treatment on untreated (TU).

In order to lay out the formal evaluation problem, we use the explanation given by Blundell and Dias (2002).

If a policy intervention at time k for which we want to measure the impact on some outcome variable, Y. This outcome is assumed to depend on a set of exogenous variables, X, the particular relationship being dependent on the participation status in each period t. Let D be a dummy variable representing the treatment status, assuming the value 1 if the agent has been treated and 0 otherwise. The outcome’s equations can be generically represented as follows,

\[ Y_{it}^s = s_i^s(X_i) + U_{it}^s \]

\[ Y_{it}^0 = s_i^0(X_i) + U_{it}^0 \]

where the superscript stands for the treatment status and the subscripts i and t identify the agent and the time period, respectively. The functions \( s_i^0 \) and \( s_i^1 \) represent the relationship between the potential outcomes (\( Y_0, Y_1 \)) and the set of observables X and (\( U_0, U_1 \)) stand for the error terms of mean zero and assumed to be uncorrelated with the regressors X. The X variables are not affected by treatment (or pre-determined) and are assumed known at the moment of deciding about participation. For comparison purposes, this means that agents are grouped by X before the treatment period and remain in the same group throughout the evaluation period.

We assume that the participation decision can be parameterised in the following way: For each individual there is an index, IN, depending on a set of variables W, for which enrolment occurs when this index raises above zero. That is:
\[ \text{IN}_i = f(W_i) + V_i \]

where \( V_i \) is the error term, and,

\[ D_{it} = 1 \text{ if } \text{IN}_i > 0 \text{ and } t > k \]
\[ D_{it} = 0 \text{ otherwise} \]

Except in the case of experimental data, assignment to treatment is most probably not random. As a consequence, the assignment process is likely to lead to a non-zero correlation between enrolment in the program - represented by \( D_{it} \) - and the outcome’s error term – \((U_0, U_1)\). This happens because an individual’s participation decision is probably based on personal unobservable characteristics that may well affect the outcome \( Y \) as well. If this is so, and if we are unable to control for all the characteristics affecting \( Y \) and \( D \) simultaneously, then some correlation between the error term and the participation variable is expected. Any method that fails to take such problem into account is not able to identify the true parameter of interest. Under the above specification, one can define the individual-specific treatment effect, for any \( X_i \), to be

\[ \alpha_{it}(X_i) = Y_{1it} - Y_{0it} = (g_{it}(X_i) - g_{0t}(X_i)) + (U_{1it} - U_{0it}) \text{ with } t > k. \]

The different parameters of interest measured in period \( t > k \), can then be expressed as:

**Average Treatment Effect:**

\[ \alpha_{ATE} = E(\alpha_{it}|X = X_i), \]

**Average Treatment on the Treated Effect:**

\[ \alpha_{TTE} = E(\alpha_{it}|X = X_i, D_t = 1), \]

**Average Treatment on the Untreated Effect:**

\[ \alpha_{TU} = E(\alpha_{it}|X = X_i, D_t = 0). \]

Having laid out the formal problem, we now turn to the methods that are used to estimate impact.
4.4 EXPERIMENTAL AND NON-EXPERIMENTAL STUDIES

4.4.1 Experimental Studies

The key concept of any experiment is the randomized assignment of individuals into treatment and control groups. It eliminates bias in the estimates of impacts. These biases result from pre-existing or ex-ante differences between the participants and comparison groups that can be confounded with the effects of program participation. A completely random assignment of individuals or households into two groups, a treatment group that participates in the program and a control group that does not, ensures that any observed differences over time between the groups will be due, on average, to participation in the program.

There are two conditions are needed to ensure that random assignment eliminates pre-existing differences between treatments and controls; first, the assignment procedures must be truly random and the process should not affect the program itself and second, control group should not have access to the program or a close substitute. Though there could be cases where control groups refuse to participate or deviate from normal behaviour (Heckman et. al. 1999), still randomization is the preferred evaluation methodology (Barnow and King 2000; Orr 1999; LaLonde 1986). Despite the fact that randomized experiments give best estimates for impact of interventions, they are not always feasible due to the high costs and detailed prior planning involved. Also, there are ethical issues involved by purposely denying the control group benefits of interventions. Yet, these have been used extensively.

4.4.2 Non-Experimental/Observational Studies

If random assignment is not possible, then the data at hand is an account of how individuals fared after an intervention. The objective of any observation study is then to use this information in an appropriate way so as to replace the comparability of treatment and control groups by design. In the case of non-experimental studies, the researcher has very little control over the data. He observes the individuals after they have participated in programs. The non-experimental methods can be categorized as (i) those that primarily address bias due to observable characteristics and (ii) those concerned with bias from unobservables, including multivariate regression model, reflexive comparison, double-difference and instrumental variable methods (Smith 2000).
The Problem of Selection Bias

Selection bias relates to unobservables that may bias outcomes (for example, individual ability, preexisting conditions). Randomized experiments solve the problem of selection bias by generating an experimental control group of people who would have participated in a program but who were randomly denied access to the program or treatment. The random assignment does not remove selection bias but instead balances the bias between the participant and nonparticipant samples. In quasi-experimental designs, statistical models (for example, matching, double differences, instrumental variables) approach this by modelling the selection processes to arrive at an unbiased estimate using nonexperimental data. The general idea is to compare program participants and nonparticipants holding selection processes constant. The validity of this model depends on how well the model is specified. A good example is the wages of women. The data represent women who choose to work. If this decision were made, we could ignore the fact that not all wages are observed and use ordinary regression to estimate a wage model. Yet the decision by women to work is not made randomly—women who would have low wages may be unlikely to choose to work because their personal reservation wage is greater than the wage offered by employers. Thus the sample of observed wages for women would be biased upward. This can be corrected for if there are some variables that strongly affect the chances for observation (the reservation wage) but not the outcome under study (the offer wage). Such a variable might be the number of children at home.

Source: Greene (1997)

Multivariate regression is used to account for possible differences between participants and the comparison group on measurable characteristics. This is done by measuring impact of program participation on outcome holding all the factors constant. 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. However, if there is bias due to endogenous program placement or due to selection, then the regressions estimates could not give correct estimate and more robust techniques are required.
Matching: A matched comparison is formed by selecting individuals or households that are very similar to the participant group using measured characteristics. The principal idea of exact matching is to assign to one or more of the individuals in the intervention sample as matching partners for one or more individuals from the non-experimental control sample who are similar in terms of their observed individual characteristics. The method used for matching is the “propensity score matching”. All observed characteristics are used to predict the likelihood of program participation. The non-participants are then matched with participants for closest propensity score. The impact estimate is then the difference between the mean outcome of the participants and the comparison group mean. The precision and usefulness of matching methods is questioned by some (Dehejia & Wahba 1999; Heckman 1997) as the ability of a matched comparison design to reduce bias depends on the selection of the matching characteristics and the sub-sample comprising the comparison group. Yet, due to its low cost and ease of use, the propensity score method has been used extensively in safety net and job training evaluations (Heckman, LaLonde and Smith 1999) and is gaining popularity in other evaluations as well.

Difference-in-Difference Method: The double difference method attempts to eliminate bias from unobservable characteristics when the comparison group is externally selected. The idea is that the treatment and control group differ from one another mainly in ways that are not observable, but these do not change overtime, and can be subtracted out from an estimate of program impact. The impact estimate is the difference between the before- and-after program change in outcome for the participants (the first difference) and the before- and-after program change outcome for the comparison (the second difference). Under the assumption of a fixed unobservable difference between the two groups, the difference-in-difference method provides a consistent estimate of the overall program impact.

Heckman, LaLonde and Smith (1999) show that participants in many social programs, especially labour market and safety net transfer program exhibit a sharp decline in earning just before participation which the non-participants do not exhibit, contradicting the view that there is simply a fixed unobservable difference in income levels.
**Instrumental Variable (IV) Method:** Using the regression model, the outcome variable is regressed on the indicator of program participation and other variables that affect the outcome. If participants and comparisons differ due to unobservable characteristic 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. An ‘instrument’ would be a variable that’s highly correlated with program participation but not related to the outcome. Instrumental variable method has been used for Argentina TRABAJOAR evaluation, the PROBECAT evaluations and Bangladesh Food-for-Education evaluations.

**Before-After Comparison:** Probably the most common evaluation strategy for attempting to construct a counter factual is a comparison of treated individuals with themselves before and after an intervention. The comparable non-participants, then are the participants themselves before the program was implemented. This requires longitudinal data.

**Cross-Section Estimates:** In before-after comparisons, program participants serve as their own controls. In the absence of any longitudinal data, one could use cross-section data. Here, one needs to take the means over the corresponding observations in random samples of participants and non-participants and estimate impact of intervention.

**4.5 SUMMARY**

Program evaluation has become an important tool to assess effectiveness and efficiency of programs. Its initial applications in the labor market have provided useful insights into how programs should be designed and implemented. The estimation of effects in experimental studies does not involve the problem of bias. However, most of the interventions are quasi- or non-experimental and the bias need to be corrected by use of robust econometric techniques. We presented these in this chapter. The use of anyone of these depends on the nature of data available. We use the method of propensity score matching in our analysis to arrive at more robust estimates.
Summary of Quantitative Methods for Evaluating Program Impact

The main methods for impact evaluation are discussed below.

Experimental or Randomized Control Designs

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

Non-experimental or Quasi-Experimental Designs

- Matching methods or constructed controls, in which one tries to pick an ideal comparison that matches the treatment group from a larger survey. The most widely used type of matching is propensity score matching, in which the comparison group is matched to the treatment group on the basis of a set of observed characteristics or by using the “propensity score” (predicted probability of participation given observed characteristics); the closer the propensity score, the better the match. A good comparison group comes from the same economic environment and was administered the same questionnaire by similarly trained interviewers as the treatment group.

- Double difference or difference-in-differences methods, in which one compares a treatment and comparison group (first difference) before and after a program (second difference). Comparators should be dropped when propensity scores are used and if they have scores outside the range observed for the treatment group.

- Instrumental variables or statistical control methods, in which one uses one or more variables that matter to participation but not to outcomes given participation. This identifies the exogenous variation in outcomes attributable to the program, recognizing that its placement is not random but purposive. The “instrumental variables” are first used to predict program participation; then one sees how the outcome indicator varies with the predicted values.

- Reflexive comparisons, in which a baseline survey of participants is done before the intervention and a follow-up survey is done after. The baseline provides the comparison group, and impact is measured by the change in outcome indicators before and after the intervention.

Source: Baker (2000)