Impact Evaluation: Methodological and Operational Issues

This quick reference provides an overview of methods available for evaluating impacts of development programs, and addresses some common operational concerns about their practical application. It is tailored for staff and consultants of the Asian Development Bank and their counterparts in developing member countries, although equally useful for those in similar institutional settings.

About the Asian Development Bank

The work of the Asian Development Bank (ADB) is aimed at improving the welfare of the people in Asia and the Pacific, particularly the nearly 1.9 billion who live on less than $2 a day. Despite many success stories, Asia and the Pacific remains home to two thirds of the world's poor. ADB is a multilateral development finance institution owned by 66 members, 47 from the region and 19 from other parts of the globe. ADB's vision is a region free of poverty. Its mission is to help its developing member countries reduce poverty and improve the quality of life of their citizens.

ADB's main instruments for providing help to its developing member countries are policy dialogue, loans, equity investments, guarantees, grants, and technical assistance. ADB's annual lending volume is typically about $6 billion, with technical assistance usually totaling about $180 million a year.

ADB's headquarters is in Manila. It has 26 offices around the world and has more than 2,000 employees from over 50 countries.

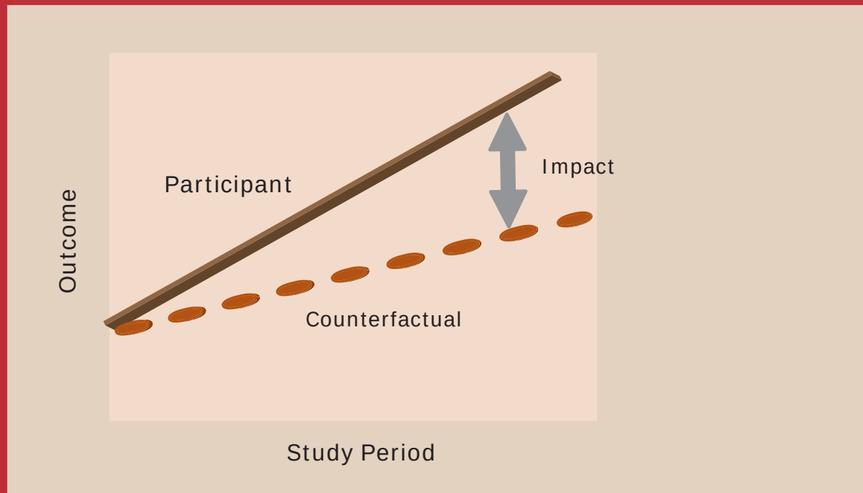
Asian Development Bank
6 ADB Avenue, Mandaluyong City
1550 Metro Manila, Philippines
www.adb.org/economics
Publication Stock No. 080906

Printed in the Philippines

ADB

IMPACT EVALUATION

METHODOLOGICAL AND OPERATIONAL ISSUES



Asian Development Bank

ADB

METHODOLOGICAL AND OPERATIONAL ISSUES

Economic Analysis and Operations Support Division
Economics and Research Department

September 2006

Asian Development Bank

The views expressed in this publication do not necessarily reflect the views and policies of the Asian Development Bank or its Board of Governors or the governments they represent.

The Asian Development Bank does not guarantee the accuracy of the data included in this publication and accepts no responsibility for any consequence of their use.

This quick reference was prepared by Binh T. Nguyen, Economist, Economics and Research Department, Asian Development Bank; and Erik Bloom, Senior Economist, Human Development Department, Latin America and Caribbean Region, World Bank. The authors thank David Green for reviewing an earlier version of this publication; and Juzhong Zhuang, Ajay Tandon, and participants at an ADB seminar on project impact evaluation organized by the Operations Evaluation Department on 27 April 2006 for their useful comments.

Published by the Asian Development Bank, 2006.

Asian Development Bank
6 ADB Avenue, Mandaluyong City
1550 Metro Manila, Philippines
Tel +63 2 632 4444
Fax + 63 2 636 2444
www.adb.org/economics

Publication Stock No. 080906

Contents

I. Introduction

Official development assistance has grown significantly over the past five decades. Currently, every year developed nations and international organizations spend more than $55 billion in grants, technical assistance, and concessionary loans to help poor countries (Dugger 2004). However, the recorded results of development assistance are mixed, and questions of whether and by how much development assistance contributes to economic growth and poverty reduction in recipient countries have often been asked (Rajan and Subramanian 2005, Easterly 2001).

A great deal of effort is being made to assess policy choices *ex ante*, particularly at the designing and monitoring stages, to ensure that project goals and activities are sound and resources are spent to producing outputs. However, considerably less is being done to evaluate *ex post* whether projects actually achieve their ultimate objective of improving welfare for the target population. Some researchers even claim that too few programs are evaluated to the quality standard required and that firm evidence of the likely impact of proposed policy changes is rarely presented (Pritchett 2002). This project management practice of lacking reliable evaluations of impact has resulted in an "evaluation gap" and, consequently, the missing body of knowledge that is required to guide future policy design.[1]

Increasingly, the development community, including donors and governments are looking for more hard evidence on impacts of public programs aiming to reduce poverty. Currently, major institutions like the World Bank, Inter-American Development Bank, and the Organisation for Economic Co-operation and Development are making concerted efforts to have more rigorous impact evaluations that will fill the evaluation gap. The World Bank has started an initiative called Development Impact Evaluation that aims to increase the number of World Bank-financed projects with impact evaluations, and to build knowledge gained from completed evaluations. A database with public access to all completed and ongoing World Bank-financed impact evaluations is available on the World Bank's Poverty Impact Evaluations Database (World Bank 2006b). Several joint actions are being planned, including the creation of an independent entity to sponsor rigorous impact evaluations of social programs in developing countries (see Center for Global Development 2006).

[1] The term "evaluation gap" was coined in a recent report by the Center for Global Development to refer to the lack of quality impact evaluations (Savedoff et al. 2006).

The objective of this quick reference is to provide an overview of methods available for evaluating impacts of development programs and to address some common operational concerns about applying them in practice. The reference is aimed at staff and consultants of the Asian Development Bank (ADB) and their counterparts in developing member countries (DMCs), although it is equally useful for those in similar institutional settings.

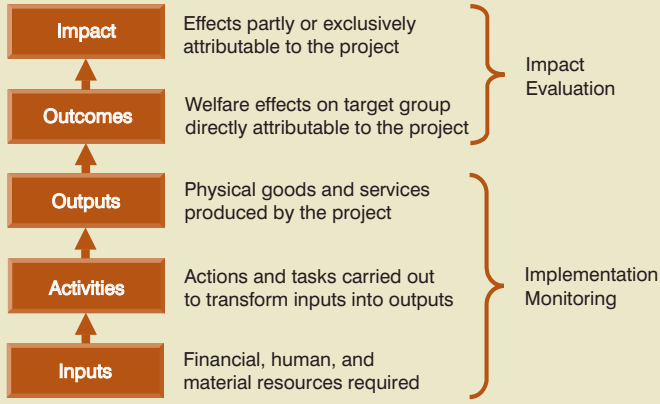
Section II briefly presents general concepts and approaches of impact evaluation. Section III provides an overview of quantitative methods available for evaluating development interventions, and discusses the major technical drawbacks in applying these methods. Section IV addresses some general concerns about impact evaluation, introducing them as "myths" and explaining why impact evaluation is less complicated than often assumed. Section V discusses issues of operational implication including choosing an evaluation method, designing steps, and resource commitments. Section VI introduces two examples of impact evaluation and explains in greater detail how impact evaluation can be realistically implemented. Section VII concludes. The reference also provides a glossary of practical terms frequently used in the field and a summary of recommended readings.

II. What is Impact Evaluation?

The logical framework ("log frame") is used to assess the operational flow of project inputs and outputs and is common in the design of projects, programs, and strategies. However, the higher-order project results ("outcomes" and "impacts") are rarely measured in practice. Oftentimes, evaluation studies focus only on the process or the inputs, activities, and outputs, making it difficult to attribute the observed results to any one particular investment and to convincingly show the outcomes or impacts of the project.

Project impact evaluation studies the effect of an intervention on final welfare outcomes, rather than the project outputs or the project implementation process. More generally, project impact evaluation establishes whether the intervention had a welfare effect on individuals, households, and communities, and whether this effect can be attributed to the concerned intervention. In other words, impact evaluation looks at project results at a higher level. The difference between impact evaluation and process evaluation and project monitoring can be seen using the five distinct components (impact, outcome, output, activities, and inputs) in the project monitoring and evaluation framework in Figure 1.

Figure 1
Project Monitoring and Evaluation Framework



Source: Adapted from a presentation by Savedoff (2006).

In general, impact evaluations can be classified into two approaches: quantitative approach and qualitative approach.

The basic organizing principle of quantitative impact evaluation is the use of an explicit counterfactual analysis. More specifically, quantitative impact evaluation isolates the welfare effect of a specific project by comparing the actual observed outcomes of project participants with counterfactual outcomes, i.e., the hypothetical outcomes that would have prevailed in the absence of the project. Since people are either in or not in the project and cannot be both, these hypothetical counterfactual outcomes cannot be observed. The central objective of quantitative impact evaluation is to estimate these unobserved counterfactual outcomes.

Because of this counterfactual analysis, quantitative impact evaluation makes possible clear specification of the project impact being estimated. It is therefore generally regarded as more authoritative and is usually referred to as rigorous impact evaluation. Nowadays, the World Bank requires a counterfactual analysis to qualify as an impact evaluation.[2]

Why is the counterfactual so important? The answer is to avoid biases in estimating project impacts. One technique frequently used in evaluating development interventions is comparing "before" and "after" outcomes. The problem of this

[2] A review of 78 evaluations done by the World Bank's Operations Evaluation Department since 1979 finds that counterfactual analysis was used in 21 evaluations only. For more than two thirds of evaluations there was no way to assess whether the observed outcomes were in fact attributed to the project being evaluated (Kapoor 2002).

comparison is that it uses the same group of individuals (i.e., project participants) and observes the temporal change in outcome of this group. This gives a potentially biased measure of the project impact because such a comparison fails to account for the changes in outcome that happen with the project participants anyway even without the project. Simply speaking, if one compares one's income between times T_0 and T_1, the difference in income is due partly to one's benefit from the project and partly to one's income change caused by secular changes in the economy in general, even if one did not participate in the project.

Another frequently used technique is comparing the outcomes between a group with the project and a group without the project. People make efforts to make the "with" and "without" groups similar. However, these two groups are only similar in a general sense and there is no guarantee that they are identical or close to identical. An obvious reason is that participating in the project self-selects participants and nonparticipants, making the two groups different. For example, in a micro-enterprise finance program, borrowers and nonborrowers may differ in entrepreneurial capability or willingness to take risk, even if they seem similar in any other observable ways. Because of this failure to control for unobservable differences between the "with" and "without" groups, the estimated impact is biased.[3]

Qualitative impact evaluation does not use a counterfactual analysis but relies on understanding processes (i.e., if A is done, then likely B will occur, and then likely C will occur, etc.); observing behaviors (e.g., consumptions, visits to hospital); and condition changes (e.g., school conditions, irrigation canals). This type of evaluation usually draws inferences from studies like reviewing project implementation processes, interviewing project beneficiaries to get personal opinions, conducting focus group discussions, analyzing supportive secondary data, etc.[4] An example of the qualitative approach is the techniques used in participatory impact assessments that reflect changes using participants' personal knowledge about the conditions in the project area.

While qualitative evaluations build stories and provide contextual insights to what is happening with the project, they

[3] The technical term for this kind of estimation bias is selection bias.

[4] Personal interview is one of the primary ways to collect information in a qualitative evaluation. It is often conducted as an open-ended discussion between the interviewer and respondents to obtain the desired information. This is different from surveys in that surveys are designed with a standard set of structured questions to get objective quantitative information.

often are being criticized for lacking rigor and internal validity. Major critics of this evaluation approach revolve around issues such as subjectivity in data, lack of a reliable comparison group, and lack of statistical robustness often due to small sample sizes.

Quantitative impact evaluations using explicit counterfactual analyses of data from well-designed statistically representative samples are better suited for inferring causal relationships between the program and outcomes. However, there is increasing acceptance that qualitative methods can provide critical insights into the program context and in-depth explanations to the results observed in a quantitative analysis. For this reason, good impact evaluations often combine both quantitative and qualitative methods to the extent possible. This quick reference discusses quantitative evaluation methods only.

III. How to Do an Impact Evaluation: A Methodological Overview

Analysts have at their disposal a number of quantitative techniques to evaluate the impact of interventions. All techniques have limitations and the choice of a particular technique should depend on the availability of data and the nature of the intervention being evaluated.

Broadly speaking, there are two groups of quantitative impact evaluation methods defined by way of constructing the counterfactual. The first is known as **random experiment designs**, which are similar to controlled medical experiments in that they use randomization to obtain the counterfactual. The second consists of **nonexperimental methods**, which use statistical techniques to construct the counterfactual. This section provides an overview of both methods and discusses some of the key issues in using them. A number of useful references are cited during the discussion and in the Recommended Readings to provide further guidance.

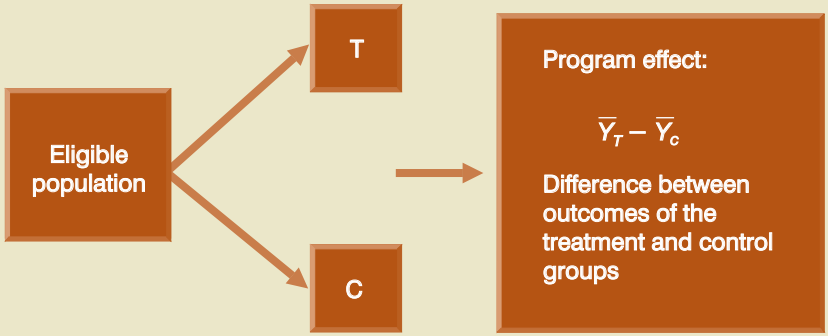
1. Random Experiment Designs[5]

Random experiments to evaluate social programs in general work very much like medical tests of new drugs. Individuals, communities, or other units of analysis from the population of eligible participants are randomly assigned to a "treatment"

[5] Experimental design is a method of research design, while randomized evaluation is an application of it to study development impacts. Kuehl (2000) is a useful reference for understanding sample

group who will receive program benefits, and a "control" group who will not. By construction, with a sufficiently large sample, on average, the two groups are identical at the outset, except for participation in the program.[6] The treatment and control groups ought to pass through the same external events over the same period of time, and thus encounter the same external factors. Therefore, any difference in the outcomes between the treatment and control groups after the program can be attributed solely to the program. For this reason, if designed and implemented properly, randomized evaluations give the best estimate of project impact (see Figure 2).

Figure 2
Random Experiment Design



designs and analyses. Boruch (1996) provides a practical guide on randomized evaluation designs. Moffitt (2004) reviews the use of randomization in evaluating welfare reforms programs. Duflo and Kremer (2003) provide examples of using randomized evaluation for education programs. Many other concrete examples of randomized evaluation can be found on the website of the MIT Poverty Action Lab at http://www.povertyactionlab.com/ (MIT 2006).

[6] This can be seen statistically as follows. Suppose X is a characteristic of the population of interest (X can be observable like education, age, etc. or unobservable like capability, ambition, etc.). Further suppose that X follows a distribution with mean μ and variance $ó^2$. From this population we randomly draw a treatment group X_T and a control group X_C of size N_T and N_C, respectively. Then, with a sufficiently large sample, by the law of large numbers sample averages $\overline{X}_T$ and $\overline{X}_c$ follow a normal distribution with mean μ and variance $ó^2/N_T$ and $ó^2/N_C$, respectively. That is, on average, X_T and X_C are identical. This statistical result can occur only if X_T and X_C are randomly drawn from the population X.

Random assignment to the treatment and control groups can be implemented in different ways, depending on the nature of the intervention. When the program benefits are provided to individuals, *individuals* can be the unit of randomization. When the program benefits are provided to groups, such as schools or communes, *groups* can be the unit of randomization. *Geographical areas* such as cities, counties or villages can also be the unit of randomization. The three most common randomization mechanisms are **lottery design**, **phase-in design**, and **encouragement design.**

In a **lottery design**, applicants are simply randomly assigned to the treatment group and the control group. This is precisely the same as in lottery when everyone has an equal chance of getting in the program. This design is used when the program resources can cover only a fraction of eligible participants and there is no reason to discriminate among applicants. In such cases, lotteries are generally perceived as a fair and transparent means to decide who will receive the program benefits and who will not. For example, a textbook program has only 5,000 textbooks to distribute in a district that has 20,000 school children who are equally qualified to receive a textbook. One simple way to pick the 5,000 beneficiary students is to put the names of the 20,000 eligible students in a basket, mix them up, and then randomly draw them.

Lotteries were applied in Colombia in the mid-1990s to distribute government subsidies. When the country faced shortages of secondary education supply, to encourage children of low-income families go to private schools, the government used lotteries to distribute vouchers partially covering the cost of private secondary school to eligible students (e.g., see Angrist et al. 2002).

The **phase-in design** can be used when the program is designed to cover the entire eligible population but in a phased-in manner. In such a case, everyone is told that they will end up receiving the program benefits but at different times. The timing of actually receiving the program benefits can be randomized. For example, a microfinance program will eventually provide credit to 500 villages in a province in 5 years. However, the program can only support 100 villages each year due to, for example, management capability. In this case, one can randomly pick 100 villages to receive credit in the first year and another 100 villages to receive credit in the second year. The 100 second-year villages can serve as the controls for the 100 first-year villages. In the beginning of the second year, one again randomly picks from the remaining 300 villages another 100 villages to receive credit in the third year. These third-year villages will serve as the controls for the second-year villages. The process is repeated

to cover all 500 villages. In this phase-in design, later-year villages serve as controls for the villages that receive credit in the preceding year.

In 1997 Mexico started a program called Progresa to provide conditional cash transfers for education and health to low-income families. Because of the scale of the program, the government decided to randomly phase in eligible families across the country according to available federal resources. The coverage expanded from about 300,000 families in 1997 to about 2.6 million families in 2000. Currently, the program covers about 4.5 million low-income families or about 20% of all families in Mexico (see Parker and Teruel 2005).

The **encouragement design** is used when everyone is immediately eligible to receive the program benefits and there is enough funding to cover the entire eligible population, but not everyone will necessarily take advantage of the program. In such cases, the program staff can randomly select a group of people and offer them specific incentives to encourage them to use the program. The remaining population without the incentives is used as the control group. For example, consider a vocational training program targeting women between 25 and 40 years of age. Since not all of these women will attend the training—for example because of family constraints—incentives can be randomly offered to some to encourage their participation. The incentives could be bus tickets to cover transportation costs or full or partial compensation of the income lost due to attending the training.

Precautionary Note: Threats to validity of a randomized evaluation

The primary precondition for the "gold standard" (i.e., high internal validity) of randomized evaluation is the integrity of the data from the treatment and control groups. In other words, in order for the evaluator to get a reliable estimate of the program effect, the treatment and control groups must remain clean and unchanged as originally designed throughout the study period. But this can be difficult in practice because evaluators of social experiments usually do not have full control over what is happening in the experiment. Persons in the treatment and control groups are usually free to participate in the study, and may pay attention or not as they like. Some individuals may interact or even leave the program even if by the study design they are not expected to. All these processes can jeopardize the original experiment design and therefore reduce the statistical power of the estimation. Below are three issues that often occur that evaluators should be aware of while designing a randomized evaluation.

(i) **Attrition.** This is the situation when some members of the treatment or control group, or both, drop out from the sample. For example, in a vocational training program, since people in the control group do not benefit from the program, they will tend to drop out from the survey because they do not have incentives from being surveyed. Attrition in the treatment group is generally higher the less desirable the intervention. If the dropouts are systematically different from the stay-ins, this will violate the randomness in the assignment of the treatment and control groups, making the treatment and control groups no longer equivalent at the outset.[a] In such cases, the two identical groups (identical in that observations are the same in every aspect, except for participation in the program) are not being compared, thus the difference between the mean outcomes of the treatment and control groups at the end of the study period is not a correct estimate of the program effect.

More specifically, if the average outcome of the dropouts from the control group is lower than that of the stay-ins, the average outcome of remaining observations in the control group is higher than the average outcome of the original control group comprising both dropouts and stay-ins. The difference between the treatments and the remaining controls is therefore smaller than the difference between the treatment and the original control group. In this case, comparing the treatment and the "altered" control group will result in an underestimate of the true program effect. In the opposite case when the dropouts from the control group happen to have a greater average outcome than the stay-ins, the result is an overestimate of the true program effect. Since dropouts usually are not available for surveying, it is difficult to estimate their outcomes, therefore the direction of the estimation bias cannot be known. Karlan (2001) discusses estimation biases with dropouts and suggests some possible solutions.

[a] As in footnote 6, the condition for the equality between $\bar{X}_T$ and $\bar{X}_C$ is that X_T and X_C are randomly drawn from population X.

Precautionary note on threats to validity. *continued.*

(ii) Spillover. This occurs when the program impact is not confined to program participants. For example, a micro-enterprise finance program that increases business activities in a program town may also cause increased employment opportunities in a neighboring town. In this case, the true program impact is the sum of increased business activities in the program town and increased employment opportunities in the neighboring town. Another example is an agricultural extension program. Program farmers practice the newly learned farming activities in their fields. Nonprogram farmers can observe and apply this in their fields. In this case, the program effect is the yield change in the fields of both program and nonprogram farmers.

When spillovers are positive, not properly accounting them will underestimate the true program impact. On the contrary, when spillovers are negative, not accounting them will overestimate the true program impact. One example is a study of a deworming program in Kenya. Because worms can be transmitted among school children when they play with each other, deworming has a positive spillover effect. Miguel and Kremer (2004) find that when this spillover effect is not accounted for, deworming school children fails to pass the cost–benefit test. However, when it is accounted for, deworming is extraordinarily cost-effective at only $3.5 per additional year of schooling per student.

(iii) Noncompliance. This is another complication in randomized evaluation. It occurs when some members of the treatment group do not get treated or get treated improperly, or some members of the control group get treated. For example, in their study of a deworming program in Kenya, Miguel and Kremer (2004) find that not all students in treatment schools actually got treated. One reason was that individual permission from each child's parent was required because this was a research study. Getting individual permission involved parents coming to school and signing a permission slip in the principal's office. This was not a trivial requirement for many parents because traveling to the school was time-consuming and some parents were reluctant to meet the headmaster when behind on school fees. Another reason was simply because some children missed the day when the deworming medicine was provided. In addition, they also find that for various reasons, about 5% of children in control schools received the medicine.

When noncompliance occurs, the impact being estimated actually corresponds to the impact of the contaminated program, and thus it must be adjusted to get a correct estimate for the impact of the original program. In the above example, at first, it seems natural to exclude from the final evaluation those in the treatment schools who did not get treated and those in the control schools who actually received the medicine. However, if these noncompliers were systematically different from the rest of the children, excluding them will result in a biased estimate of the program impact because the comparison between the treatment and control schools is not made on the two total groups. The evaluator must closely monitor these noncompliers and decide whether to include them depending on how random they are.

The prospective nature of randomized evaluation makes planning and designing the most important stages of the evaluation. Attrition, spillover, noncompliance, and any other sources of estimation bias that may arise from randomization should be anticipated. The evaluation design should be such that these "noises" are minimized or avoided altogether upfront. When the evaluation is already under way, it is almost impossible to redo the evaluation design.

Evaluators of social programs have now accumulated a set of practices to deal with threats to the validity of a randomized evaluation at both the design and analysis stages. Readers interested in more detailed discussions are recommended to read Campbell and Stanley (1966), Cook and Shadish (1994), Dunn et al. (2003), Meyer (1995), and Weiss (1998). For example, to reduce the severity of attrition, the evaluator may undertake several measures simultaneously, including making generous payments to controls for providing outcome data; monitoring attrition rates early to describe them, elicit their possible causes, and take remedial actions; following up dropouts to get relevant information; and designing the experiment to contrast different treatments as opposed to contrasting a single treatment group with a no-treatment control group. In programs where treatment diffusion is likely, the evaluator will have to look for opportunities to study groups that cannot communicate with each other; or will have to design the evaluation such that spillover effects can be measured. At the analysis stage, evaluators also have a number of analytical tools, including, for example, Intend to Treat, to reconcile the program effect when only a fraction of observations in the treatment group actually got treated. However, most of these methods are beyond the scope of this quick reference and thus are not discussed.

A last note is that evaluators of social programs using random experiment designs cannot take for granted that a perfect random assignment plan will be perfectly implemented. Oftentimes, the actual physical assignment process is not done by evaluators themselves but by field staff, thus much depends on the understanding and interpretation of these professionals. Close monitoring to ensure the quality of implementation of the evaluation design and frequent checking on what is happening with the treatment and control groups is a must for randomized evaluations.

2. Nonexperimental Methods[7]

Nonexperimental methods sometimes are also called statistical methods because they use statistical techniques to simulate the counterfactual, i.e., the outcome that would have prevailed had there been no intervention.[8] The most frequently used nonexperimental methods available for evaluating

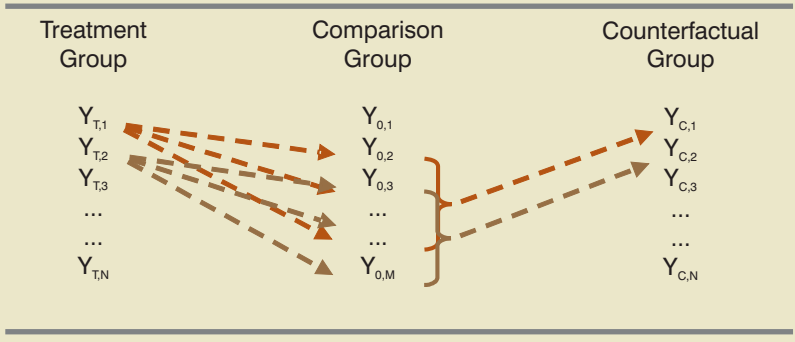
[7] Ravallion (2005) is an excellent survey of impact evaluation methods available for ex-post counterfactual analysis of development programs. The survey has an extensive list of applications of nonexperimental evaluations.

[8] Sometimes also called quasi-experimental methods.

development programs include **propensity score matching (PSM)**, **difference in differences (DD)**, **regression discontinuity design (RDD)**, and **instrumental variables (IV)**.

The basic idea of the **propensity score matching** method is to match program participants with nonparticipants typically using individual observable characteristics. Each program participant is paired with a small group of nonparticipants in the comparison group that are most similar in the probability of participating in the program. This probability (called propensity score) is estimated as a function of individual characteristics typically using a statistical model such as logit or probit model.[9] The mean outcomes of these groups of matched nonparticipants form the constructed counterfactual outcome. This matching procedure is visually illustrated below (see Figure 3).[10] The mean program impact is estimated by the difference between the observed mean outcome of the project participants and the mean outcome of the constructed counterfactual.[11]

Figure 3
Propensity Score Matching: Counterfactual Constructing Procedure



[9] These models have the general form *Probability* (person *i* is in the program) = *G* (characteristics of person *i*), where *G* can take the logistic function (in the logit model) or the standard normal distribution function (in the probit model). See, for example, Wooldridge (2002).

[10] Note that the numbers of observations in the treatment (*N*) and comparison group (*M*) are not necessarily the same.

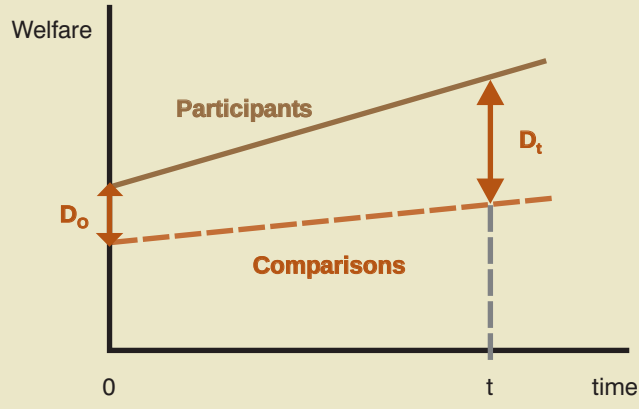
[11] Taking the mean outcome of matched nonparticipants is the simplest PSM estimator. There are other matching algorithms defined by way of assigning the neighborhood for each treatment individual and the weight to each of the matched nonparticipants in this neighborhood. Caliendo and Kopeinig (2005) discuss this in more detail.

The **difference in difference** (or double difference) method entails comparing observed changes in outcome before and after the project for a sample of participants and nonparticipants. Typically, one collects outcome data of both participants and nonparticipants using a baseline survey before the program. One then repeats this survey at some later point(s) after the program is implemented. This repeat survey(s) should be highly comparable with the baseline survey in terms of the questionnaire, the interview, etc. The mean program impact is estimated by comparing the mean difference in outcomes "after" (D_t) and "before" (D_0) the intervention between the participant and nonparticipant groups.[12] This can be seen more clearly in Figure 4.

The underlying assumption of the DD method is that project participants would have the same outcomes as individuals in the comparison group in the absence of the project. Since this is highly unlikely in reality, PSM is a natural choice to select a comparison group before calculating the differences in a DD method. For this reason, the PSM and DD methods are often used together in practice.[13]

The **regression discontinuity design** method can be used when program participation is determined by an explicitly specified exogenous rule. The method stems from the intuition

Figure 4
Double Difference: Graphical Illustration



[12] More specifically, the mean program impact is estimated as $D_t - D_0$, where: $D_t = \frac{1}{N_p}\sum_P Y_t^P - \frac{1}{N_C}\sum_C Y_t^C$, $D_0 = \frac{1}{N_P}\sum_P Y_0^P - \frac{1}{N_C}\sum_C Y_0^C$, Y is a welfare measure, N is the number of observations, P = participants, and C = comparisons.

[13] For more detailed descriptions of PSM and DD methods, see Ravallion (2003).

that individuals around the cut-off point for eligibility are similar and uses individuals just on the other side of the cut-off point as the counterfactual. In other words, RDD compares outcomes of a group of individuals just above the cut-off point for eligibility with a group of individuals just below it.[14]

For example, children of a Filipino family with per capita income equal or below P1,000 per month are eligible to receive a textbook, and ineligible otherwise. It is conceivable to assume that children from families with income per capita, for example, of P1,001 per month are similar to children from families with income per capita of P999 per month. The RDD method uses them as comparisons, and the mean program impact is estimated by the difference between the mean outcomes of the P1,001 group and the P999 group.

The major technical problem of the RDD method is that it assesses the marginal impact of the program only around the cut-off point for eligibility, and nothing can be said of individuals far away from it. In addition, for the RDD estimate to be valid, a threshold has to be applied in practice and individuals should not be able to manipulate the selection score to become eligible. In the above textbook example, children from families with income per capita above, for example, P1,100 per month cannot be manipulated to appear as if their incomes were around or below P1,000 per month. Since these children are unlikely to be similar with children with income around the threshold level, having them in the analysis will bias the estimated impact.

The **instrumental variables** method works exactly as a standard regression analysis. When the program placement is correlated with participants' characteristics, then the estimate of program effect using an ordinary least squares regression model is biased. To correct this, one needs to replace the variable characterizing the program placement with another variable (called instrument) such that it mimics the variable being replaced (i.e., correlated with the program placement) but is not directly correlated with the program outcome of interest.[15]

---

[14] A detailed discussion of the RDD method can be found in Trochim (1984).

[15] The general analytical framework of the IV method is that the project effect can be characterized by the relationship $y = a + bT + cX + e$, with T being the dummy variable for program placement, X the vector of individual characteristics, and e an error term not captured by the model. In this model, an ordinary least squares estimate of b will be biased if T is correlated with the error term. An instrumental variable Z for T is such that Z mimics T but is uncorrelated with e. Because of its standard application, presentation of the IV method is available in most basic textbooks on regression analysis; see for example Wooldridge (2002).

For example, suppose before starting a microfinance program, flyers with information about the program were distributed randomly in some villages. Since reading a flyer may induce program participation but will not affect income *per se*, whether flyers were distributed or not in a given village can be used as an instrument for participation.

Another example is a study by Attanasio and Vera-Hernandez (2004) on the impact of a nutrition program that provides food and child care through local community centers in Colombia. Because local community centers are used by some villagers and not by others, it is conceivable to believe that usage of these facilities is endogenous to health outcomes. An estimate of the program effect using a regression model with usage is therefore biased. The authors use the distance from the household to the community center as the IV for attending the community center. The authors justify their choice of this IV variable on the ground that living near the community center may induce the usage of the center facilities but not directly affect the health outcomes of interest.

It is worth noting that justifications of an IV must ultimately rest on information beyond the confines of the quantitative analysis, including an understanding of program contexts, theoretical arguments, empirical observations, and common sense. To derive an appropriate IV, it is therefore necessary to have information from sources other than typical household survey data, including qualitative open-ended type of interviews.

Precautionary Note: Estimation biases when using nonexperimental methods

Since nonexperimental methods use statistical techniques to model the behavior of participants and nonparticipants, using them requires a high level of precaution to avoid or minimize estimation biases.[a] The first kind of estimation bias arises from failing to account for observable variables, called *omitted variables* estimation bias. For example, if education is one of the determinants of the program participation, not including it in estimating the probability of program participation in a PSM model will result in a biased estimate of that probability. Apparently, this will give wrong matched nonparticipants and consequently a wrong counterfactual constructed from these wrong matches. The estimated project impact is therefore incorrect.

The second kind of estimation bias is called *selection bias* and comes from endogenous program placement. Assignment of poverty reduction programs often is determined by selection criteria, for example, income below a certain level. This endogenous program placement effectively makes program participants and nonparticipants different in some set of characteristics (e.g., in income level). Even when participation is voluntary, that participants self-select into the program makes them different from nonparticipants. For instance, borrowers in a microenterprise finance program may be intrinsically more entrepreneurial or more willing to take risk than nonborrowers. Because of these endogenous program assignment and self-selection participation, those who are in the program are often not a good comparison for those in the program.[b] The observed difference in the outcome of interest is therefore attributable to both the program and the pre-existing differences between participants and nonparticipants.

Another source of estimation bias comes from a *misspecification* in modeling the behavior of participants and nonparticipants. For example, one may specify labor income as a linear function of individual attributes such as education, age, and work experience. By construction, this assumes that causation goes in one direction: from individual attributes such as education, age, and work experience, to labor income. In reality, it could well be a reverse interaction from labor income to individual attributes, e.g., level of education. Specifying a model of a one-way direction therefore is erroneous and the estimated program impact is biased.

Antidotes to these estimation biases exist. However, they often are technically complex and data-intensive. In this regard, nonexperimental methods usually require extensive data of high quality to control for all factors, both observable and unobservable, that determine project participation. Also, because sometimes some of these methods may involve intense computations, they may not always be readily applicable for routine monitoring and evaluation.

---

[a] Glazerman et al. (2002) conduct a replication study of 16 randomized evaluations performed during 1982–2002 using various nonexperimental methods including PSM and ordinary least squares regression. They find that nonexperimental methods only occasionally replicate the findings from the experimental impact evaluations.

[b] In extreme cases when no subgroups of similar individuals among participants and nonparticipants can be found, there will be no way to get a reasonable comparison group of any size from nonparticipants. For instance, it happens that all project participants have a graduate degree of education and all nonparticipants only have a primary school level. In this case, of course, one cannot get a subgroup with the same education among participants and nonparticipants.

IV. Myths and Realities of Impact Evaluation

Practitioners often hesitate including meaningful impact evaluation in their project design. The justifications given vary, often citing the impracticality of carrying out a rigorous evaluation in project primarily designed to improve the living standards of the target population. While it is unnecessary and impossible to carry out impact evaluation studies for every project, they are often much easier and less costly to carry out than perceived. This part looks at some common misperceptions about impact evaluation and provides some simple responses.[16]

Myth 1: It is difficult

"The methodology for impact evaluation is quite demanding in terms of time and statistical understanding."

While it is true that the underlying theory of evaluation has significant statistical underpinnings (to correct for biases that are inherent in all evaluations), in practice many types of project impact evaluation are quite easy to carry out.

In evaluations using experimental methods, the evaluator simply compares the results between the treatment and control groups. Little additional work is required. Evaluations using statistical methods do tend to require a relatively sophisticated understanding of statistics. But nowadays, statistical software, e.g., Stata, have many of these routines built in, making the work much easier. The evaluator needs only to understand the underlying techniques. Given the importance of results-based management within ADB, it is realistic for the project officer to seek technical assistance from within ADB or to hire a consultant using technical assistance or staff consultant resources if the project does not have sufficient resources for the technical work involved.

[16] For those interested in understanding underlying causes of why impact evaluations are not popular, Pritchett (2002) is highly recommended. This paper builds on a political economy model to explain the *status quo* underinvestment in rigorous impact evaluations.

Myth 2: It is expensive

"I understand the value of impact evaluation, but my project has a tight budget and does not have the resources to support an in-depth impact evaluation."

Of course, evaluation requires some extra resources. However, many projects already include a budget for monitoring and evaluation (M&E) activities (typically 1–5% of the total project cost). This allocation is used for routine project monitoring activities and for the preparation of baseline and final evaluations of the project. If planned in advance, many types of impact evaluation discussed in this guide can usually be included into ongoing M&E activities at no additional cost. In some cases, the survey requirements for impact evaluation are no more than for M&E at all. Thus impact evaluation does not necessarily have a major cost impact if there already is a plan for M&E in the project.

Typically, it is the government's responsibility to shoulder the cost of M&E as part of the project cost. However, with the importance given to results-based management and the possibility to generate important research findings for other activities, it may be possible for the project officer to seek technical assistance support to finance some aspects of the impact evaluation.

Myth 3: It is unethical

"Many of the proposals for experimental design involve experimenting with people and denying people basic services. This raises ethical issues."

People are rightly concerned about "experiments" that involve humans and are uncomfortable with the idea that necessary services might be randomly denied to some people and offered to others.

These concerns are real but are not difficult to manage. The experiments proposed for evaluation simply introduce policy changes that may have or may not have positive welfare impact. In many cases, projects have severe resource limitations that require some sort of rationing and randomization is generally perceived as a fair means to do it. In other cases, the timing of the project will require that some groups receive the benefits of the project before other groups. For example, the construction

of health clinics is divided into different phases or packages. The timing of the activities can be done randomly to allow for testing of the impact of the intervention. Similarly, in a rural roads project, one district may get roads before other districts. The selection of the order of the districts can be done randomly.

Myth 4: Governments will not agree

"Impact evaluation is important but the government will not agree to support it, both because of the cost and because it is concerned about the results."

Governments often have mixed feelings about M&E. While it is universally accepted that it is important to monitor activities and to correct poorly implemented activities, M&E also can raise politically uncomfortable issues. As previously argued, if M&E is part of project costs, the additional cost of including a detailed impact evaluation is often trivial.

Governments can also gain immensely from impact evaluation in several ways. First, impact evaluation can provide valuable insights into what works, which is important for governments to make decisions for their development plans. Second, it can show the public and legislature evidence of government success. Third, it can serve as an important tool to raise resources from development partners—showing that clear success is a strong incentive to increase support. In practice, many governments have supported serious impact evaluation when carefully explained.

Myth 5: It will not work in many sectors

"Impact evaluation has been used primarily in social sectors and the techniques have been designed for projects in those sectors."

Of course there are some sectors and types of projects for which it is more difficult to perform meaningful impact evaluation. For example, it is hard to measure the impact of large projects such as expressways and power grids. These projects tend to have an impact on everybody in the economy and it is hard to develop a reasonable counterfactual to isolate the project effect. It is also difficult to measure impact of program loans aiming at improvements of governance and changes to institutions. While these programs could have a major effect, it is hard to separate ADB's efforts from other changes in the

economy. Likewise, attribution is also an issue if a project is being done as part of a sectorwide approach or making extensive use of budget support.

However, many of the techniques presented in this guide can be applied with some simple innovative design. In many cases, there is a timing aspect to the construction of infrastructure. For example, in a rural roads project, not all rural roads will be built at the same time. This timing allows the evaluator to add some randomness at little additional cost. The order of road construction can be done randomly to see if the communities that get the road first have better outcomes.

Myth 6: It may show no results now

"My project's real impact will not be felt for many years. I am worried that by doing a careful evaluation now, it will give a false result about what the project's real impact is."

One concern that project implementers have with evaluation is that the results are collected before the project can have a real impact. Thus an evaluation may prematurely show a project as a failure on the basis of early results. This can lead to tension between the evaluator and the implementer.

This does not need to be an issue with impact evaluation. Since an impact evaluation requires the active participation of the project design team, they can include realistic benchmarks that can be evaluated within a given timeframe. This will allow necessary time for real changes to ensure an understanding of impacts as they happen. After all, the ultimate purpose of the evaluation is to provide knowledge of the project to the government and ADB to inform future project design.

Myth 7: No institutional mandate

"ADB does not mandate impact evaluation as a compulsory project component and there is little incentive for doing this type of project evaluation."

ADB has always been interested in evaluating the impact of its interventions and several ADB-financed projects have been subject to rather rigorous impact evaluations. However, as part of the reform agenda, ADB has very clearly signaled that it will increase the role of impact evaluation and will take steps to ensure that lessons learned from past projects are incorporated into future interventions. One indication of this is the

requirement that all interventions and country strategies have a framework specifying expected outcomes and impacts (the project framework).

One key element of the reform agenda is the Management for Development Results initiative, which focuses on measuring how ADB can contribute to a country's efforts to reach its Millennium Development Goals. Within this institutional context, while ADB does not expect every project to use a rigorous framework to evaluate its impact, it certainly appreciates and rewards the efforts of those responsible for designing projects that develop a serious plan for impact evaluation.

Myth 8: ADB already evaluates projects

"ADB already evaluates projects, both at project completion and through an independent evaluation a few years after the project ends."

At project completion, ADB operations staff prepare a project completion report that reviews the activities and outputs of the project. Several years after the completion of the project, the Operations Evaluation Department may review the project in depth to generate a set of lessons learned.

While these reports are important and may contribute to ADB operations, by default, they do not necessarily include a valid assessment of project impact. As argued, impact evaluation focuses on the project impact on final welfare outcomes and usually requires planning during project design to set up a system that will allow a valid assessment of the treatment effects on the basis of a counterfactual analysis. The standard ADB evaluation focuses on inputs, activities, and outputs and lacks a counterfactual analysis to allow an unbiased estimate of project impacts and reliable inferences of attribution.

V. Operational Implications

Undertaking a rigorous impact evaluation can be quite challenging requiring not only solid technical knowledge but also, and more importantly, firm implementation commitments from all parties involved including the government, operational staff and management. More specifically, it involves choosing the right evaluation methodology, planning and designing an evaluation component carefully from the start of the project, and sufficient continuous financial and human resources inputs. After all, because of their public good nature, impact evaluation

studies need strong institutional support. These matters are discussed in this section.

1. Choosing an Evaluation Method

What evaluation method to use depends on the nature of the intervention being evaluated, and choosing a particular method involves **trade-offs**. In general, nonexperimental methods are more popular. However, these methods may suffer **estimation biases** due to sample selection (not having a perfect control) and model specification (using incorrect statistical model). In addition, because these methods usually involve complex statistical modeling, they often require intensive data, making the evaluation more expensive and the computation often quite involved.

The greatest advantage of randomized evaluations is their high internal validity (considered as "**gold standard**") because of the high quality of the counterfactual. Also, they are relatively easy to understand and to present results. Compared to nonexperimental methods, they are less costly because they usually do not require as large samples. However, they may be **more selective in applicability**. For example, it is very difficult to do randomized evaluations of large infrastructure projects or projects designed to benefit a large part of or the entire population. In addition, the internal validity is highly subjective to the project design and implementation and thus the result can be biased if problems such as attrition, spillover, contamination, randomization biases, etc. are not properly taken care of.

2. Designing an Evaluation

Traditionally, evaluation is carried out at the end of a project cycle (often several years after the project closes) and is essentially a backward-looking exercise. It focuses on questions like: "What did the project do?" "Did it work as planned?" "What did it accomplish?"

In fact, planning and designing from the start of the project is the key for impact evaluation. Regardless of the size, program type, and methodology used, a typical evaluation study begins with determining whether or not to carry out an evaluation, setting clear objectives, identifying the evaluation method, investigating data needs, and subsequently designing samples and collecting data. Then, the collected data are analyzed and

the findings are presented to stakeholders to inform future project design. These main steps are shown in Box 1.[17]

Box 1
Designing Steps

During project identification and preparation

1. Determine whether to carry out an evaluation
2. Clarify the objectives of the evaluation
3. Investigate data availability
4. Select the evaluation method
5. Form the evaluation team
6. If data collection is needed, then
   a. design and select samples
   b. develop questionnaires
   c. staff and train for fieldwork
   d. pretest survey

During and after project implementation

7. Conduct baseline and repeat surveys
8. Analyze data
9. Write up the findings and discuss with stakeholders
10. Incorporate the findings in future project design

Source: Baker (2000).

3. Resource Requirements

Evaluation is not free. It requires the outlay of resources from the project to pay for a variety of expenditures including staff, consultants, surveys, analysis, and reports, etc. Generally, an evaluation could cost from a few thousand to a few million dollars, depending on the scale and complexity of the intervention and questions studied. However, typically, a thorough impact evaluation involves a few hundred thousand dollars.[18] In terms of time, a typical evaluation takes several years to complete, depending on how long it takes for the project to show impacts.

Regarding data collection, ideally, an evaluation needs a baseline survey and at least one follow-up survey during or after project implementation of both with- and without-project households of, typically, a few hundred to a few thousand observations. The time lag between the baseline and follow-up

[17] A thorough discussion of these steps can be found in Baker (2000).

[18] The literature on evaluation rarely provides specific information on the overall or unit costs of evaluations. One such rare example is Montgomery et al. (1996) reporting the cost of $250,000 for an impact evaluation of a credit program provided by microfinance organization BRAC in Bangladesh in 1994.

surveys should be at least as long as to allow the project impact to occur. Also, the surveys should be comparable in the questionnaires, design, interview, timing, etc. to minimize unobservable differences between the treatment and comparison groups. The sample should be designed such that the two groups are as similar as possible.

Typically, the questionnaires should contain information on **household and household member characteristics** including measures of welfare outcomes of interest and determinants (income, expenditure, assets, health status, age, education, occupation, work experience, ethnicity); **program characteristics** (locations, time, selection criteria, placement, benefits); as well as **characteristics of surrounding conditions** (local markets, schools, roads, administrative centers).[19, 20]

The data requirements for randomized evaluations are usually less intensive. Many projects have ongoing M&E systems that are already collecting data for the monitoring purposes. Most of these data and possibly a few additional indicators could be used to assess project impact. In such cases, the additional cost for data collection is minimal.

4. Evaluating Large-scale Interventions

In general, impact evaluation works best in cases where the development intervention is discrete and well-targeted. A clear example is a pilot project, where the explicit justification of the project is to learn what works with a specific population before the intervention is expanded to the general population. In contrast, interventions that are difficult to evaluate tend to have a large regional or national coverage or tend to generate public goods. For example, investing in agriculture research and development will create a public good, effectively ruling out the identification of a true comparison group. Impacts of program loans are also difficult to measure using this methodology because of their economywide impact that benefits everybody. (See Box 2.)

Many ADB-financed projects could be evaluated using the impact evaluation techniques presented. This is because projects are typically designed as stand-alone activities (albeit within a larger government strategic framework), with a well-defined set of interventions and a well-defined beneficiary group.

[19] The World Bank's household living standards surveys are an excellent illustration of the most extended type of survey and data required. See World Bank (2006a).

[20] Deaton (1997) provides a thorough discussion on treatments of household survey data.

Box 2
Evaluating Untargeted Interventions

By nature, it is difficult to evaluate the impact of large-scale, untargeted interventions. In such cases, evaluators could develop a chain of causality in which they identify, *a priori*, the likely series of events that is caused by the intervention. This should be done prior to the intervention. Then during implementation and afterward, the indicators in the chain of events need to be monitored to map out likely impact.

Qualitative assessments can be especially valuable in these cases as they help the evaluators understand the change in the thinking that is a result of the intervention.

VI. Case Studies

Impact evaluation studies are becoming increasingly common for project and program evaluation. Although much of the work has been research, ADB and other development financiers have supported a number of detailed impact evaluation studies. This section reviews two examples of impact evaluation, one financed by ADB and one by the World Bank, and shows their relevance for ADB, focusing primarily on the methodology used.[21]

1. Cambodia Contracting Experiment[22]

In 1996, ADB approved a $25 million loan for the basic health services in Cambodia. Because of shortages of public health care providers, the government was considering administering contracts to nongovernment organizations (NGOs) rather than directly providing these services. Two models of contracting were considered. In contracting-out, the NGOs would have the full responsibility for delivery of all district health services, including staff employment, procurement of drugs and operational supplies, management, etc. and are fully accountable for achievements of health targets. In contracting-in, the NGOs would provide only management of district health services, with actual staff delivering the services remaining with the ministry and operational inputs still being provided by the government.

[21] Readers interested in more case studies are referred to read Baker (2000).

[22] For more details, see Keller and Schwartz (2001).

Contracting NGOs to provide public services was rather new in Cambodia, so it was decided to introduce contracting as a pilot experiment subject to a rigorous evaluation before full-scale project implementation. Since quite a number of districts qualified for contracting (using the pre-established criteria), it was decided to ***randomly allocate*** which districts would get what type of contracting model. Several districts were also randomly chosen to be control districts with a similar budget supplement but not using a contracting model. A careful tracking survey was done in all districts to monitor the health impacts of the different type of contracting arrangements.

Evaluation Design

The evaluation involved four control districts, three districts with health care services contracted-out to NGOs, and two districts with health care services contracted-in to NGOs. A sample of 270 households in the nine districts was surveyed.

Cost of Evaluation

Monitoring and evaluation were important components in the experiment. The cost of evaluation, however, was not that great for the contracted districts, since these were anyway surveyed under the usual government delivery. Additional project resources were needed to do surveys in the control districts. Also, due to the nature of experiment, some additional technical assistance resources were required to cover the cost of researchers, additional survey work, and publications.[23]

Evaluation Findings

The evaluation finds clear and convincing evidence that the contracting-out model outperformed the contracting-in model in delivering health services to the intended beneficiaries, using impact, outcome, and process indicators.

[23] The evaluation was a component of a two-component Technical Assistance to Cambodia for the Second Basic Health Services Project (15 May 2001). The other component of the technical assistance was the preparation of a follow-up investment project. The total technical assistance cost was $850,000, including a $150,000 counterpart fund from the government.

Consequences of Evaluation

The evaluation gave strong evidence that led to the extension and expansion of contracting-out and significant co-financing. Other donors are considering contracting as an alternative for Cambodia. The evaluation generated substantial international interest in the Cambodia experience.

2. Viet Nam Rural Roads[24]

In 1997, the World Bank financed the Rural Transport Project I (RTPI) in Viet Nam. The project was designed to construct and rehabilitate about 5,000 km of district and commune level roads in 18 poor provinces over 5 years. The total project cost was about $61 million. The overall objective of the project was to raise living standards in poor rural areas by enhancing access to markets and public services such as schools, health centers, and political and administrative facilities.

The purpose of this impact evaluation was to determine how household welfare was changing in communes in the project area compared with those not in the project area. The evaluation was designed and conducted by researchers in the World Bank's Development Research Group, and began concurrently with project preparation.

Evaluation Design

This evaluation was the first comprehensive attempt to rigorously assess whether rural roads really reduced poverty by isolating the welfare change due to the project from other factors ongoing in the economy. The design of the evaluation centers on using household survey data of a sample of project and non-project communes, collected before and after the intervention. Six out of the 18 project provinces were selected for the surveys. In these selected provinces, a sample of 3,000 households in 100 project and 100 nonproject communes was chosen. Four rounds of panel survey data of these households and communes were collected in 1997 (baseline survey), 1999, 2001, and 2003. The questionnaires of the surveys were designed similarly to those used in the World Bank's Household Living Standards Surveys for Viet Nam. The surveys were carried out by a local Vietnamese research institution.

[24] For more details, see van de Walle and Cratty (2005).

The analysis combined two impact evaluation methods, the PSM and DD methods. First, PSM was used to select ideal comparison communes from among the sampled nonproject communes, using a logit model. Second, the impact of the roads is then estimated by the difference between outcomes in the project areas after and before the project, minus the corresponding outcome difference in the matched nonproject comparison areas.

Cost of Evaluation

The total cost of the evaluation was about $200,000, or about 3% of the total project cost, covering everything except World Bank staff time and travel expenses (Baker 2000).

Evaluation Findings

More uses of the survey data continue. One finding so far is that the project contributes insignificantly to rehabilitated road increments in the project districts. The evaluation finds that about one third of the investment intended to expand serviceable road length was displaced, suggesting a significant level of aid fungibility.

VII. Conclusion

Impact evaluation is an important tool for measuring empirically the impact of development projects. Impact evaluation differs from traditional project implementation and monitoring evaluation in several ways. First, it focuses on impact and outcome variables while traditional evaluations look mostly at the inputs and outputs without a meaningful evaluation of the consequences. Second, it provides a quantitative estimated measure of project impacts and causal inferences about project and outcomes on the basis of an explicit counterfactual analysis. Third, impact evaluation is built into the project design from early stages, requiring the active participation of the project team including the government, operational staff, and management. This allows the evaluation to focus on variables that are of use to implementers, and allows for midcourse adjustments. Fourth, depending on the techniques used, it can be quite economical and much less complicated than usually perceived. In some cases, it can use existing data sources with minimal additions rather than creating new data instruments.

No single impact evaluation method is ideal for all projects. A thorough impact evaluation begins with choosing the right

evaluation methodology and this usually involves some trade-offs. While random experiment designs can be a powerful tool to produce a reliable estimate of project impacts, using them requires a well thought out design at the project planning stage and continuous care throughout project implementation. Most particularly, a randomized evaluation needs to be designed and implemented such that spillover effects, attritions, noncompliances, and any other potential randomization biases are well taken care of to maintain the validity of the estimated project impact. Nonexperimental methods, on the other hand, while more popular, often can be subject to estimation biases. Critics of these methods revolve around the imperfection of the constructed counterfactual due to endogenous program placement. Generally, this class of evaluation methods requires especially careful analytical treatments to minimize the likelihood of misspecifying the statistical model and to control for observable and unobservable pre-existing differences between individuals in the project and comparison groups.

Regardless of what evaluation method is used, planning and designing an evaluation component carefully from the start of the project is always a *sine qua non.* In addition, the success of an evaluation requires firm commitment from the government, operational staff, and management to provide sufficient, continued financial and human resources inputs throughout the evaluation.

Glossary

Comparison group—a group of units in nonexperimental evaluations that are similar to the treatment group but do not receive the program benefits.

Control group—a group of units in randomized evaluations that are randomly drawn from the eligible population and not to receive the program benefits.

Counterfactual—the situation that would have prevailed had the intervention not occurred.

Endogeneity—a statistical term referring to a simultaneous causal relationship between the dependent and independent variables. In the context of project impact evaluation, it refers to the situation when the project placement is determined by some criteria that correlate with the outcome of interest. For example, subsidies are provided to people with income below a certain level.

Estimation bias—a statistical term referring to the difference between an estimate and the true value of the parameter being estimated. Obviously, a good estimate is the estimate that has no estimation bias.

Ex-ante analysis—an analysis undertaken prior to the program implementation, usually during program design, to estimate program's potential outcomes.

Exogeneity—a statistical term referring to the genuine independence of independent variables. In the context of project impact evaluation, it refers to the project placement that is independent of the beneficiaries' characteristics.

Ex-post evaluation—an evaluation of outcomes undertaken after the program is completed and the outcomes are known.

External validity—the credibility or reliability of an estimate of project impact when applied to a context different from the one in which the evaluation was carried out.

Impact evaluation (*syn.* impact assessment)—an evaluative study of a program's impact on outcome indicators of interest. For example, how a microfinance program contributed to changes in client's income.

Instrumental variable—a variable used in statistical models to replace a variable that correlates with the error term. An instrumental variable should have two properties: (i) it is correlated with the variable it replaces, and (ii) it is uncorrelated with the model's error term.

Internal validity—the credibility or reliability of an estimate of project impact conditional on the context in which it was carried out.

Outcome indicator—an indicator of an outcome of interest. For example, income, consumption, number of children, etc.

Quasi-experiment (*syn.* non-experiment)—an impact evaluation design in which treatment and control groups are formed not by a random assignment (e.g., by statistical matching methods.)

Random assignment—an assignment of the treatment and the control group based on a random draw.

Selection bias—an estimation error due to observable or unobservable pre-existing differences between the treatment and control groups.

Treatment group—a group of units in both randomized and nonrandomized evaluations that receive the program benefits.

Recommended Readings

Impact evaluation is a major field of research. There is a significant literature focusing on both the theoretical and empirical aspects of impact evaluation. There is also a large number of "how-to" guides, explaining the steps that an evaluator should follow.

Baker, Judy. 2000. *Evaluating the Impact of Development Projects on Poverty—A Handbook for Practitioners.* **World Bank, Washington, DC.**

This is a reference book at the introductory level for project managers and policy analysts. Targeted toward readers with a general knowledge of statistics, the book focuses on nonexperimental evaluation methods. It is organized into two parts: basic conceptual methodological matters and 15 real-life case studies. The methodological part presents an overview of basic concepts and techniques for project impact evaluation; key steps and related issues in designing and implementing an impact evaluation; and analytical techniques to correct for selection bias in the propensity score matching, difference in difference, and instrumental variables methods through a hypothetical case study. The 15 case studies include a mix of countries; types of projects (infrastructure, microfinance, education, health, job generation, agricultural extension); and evaluation methodologies.

Bourguignon, Francois and Luiz A. Pereira da Silva, eds. 2003. *The Impact of Economic Policies on Poverty and Income Distribution—Evaluation Techniques and Tools.* **World Bank, Washington, DC.**

This book reviews the most robust techniques and tools available for evaluating the poverty and distributional impact of economic policies. It aims at readers with intermediate to advanced knowledge of statistics, economics, and quantitative impact evaluation methods . The book covers both microeconomic and macroeconomic methods and is more suitable for technical users of the methods than for general project managers. Each chapter describes the overall framework and, in some cases, steps of a specific evaluation technique and its applications; including how household survey data are used for descriptions of economic welfare distribution. Readers interested in actual applications of these techniques will need to refer to more in-depth discussions in reference papers provided at the end of each chapter.

Rossi, Peter H., Mark W. Lipsey, and Howard E. Freeman. 2004. *Evaluation: A Systematic Approach.* 7th ed. Thousand Oaks, CA: Sage Publications.

This is one of the most comprehensive books on program impact evaluation for social programs. The book is written as a textbook with intuitive terms, and being an overall slow read, is not a suitable reference for beginners. The book provides an introduction to evaluation methods covering basic concepts, evaluating indicators, planning and designing aspects and concerns, and general issues in formulating survey questions. It also contains a wide discussion of operational matters covering indicators, target population, program outcomes, including validity of evaluations and randomized and nonrandomized evaluation methods and methods of data analysis.

Weiss, Carol. 1998. *Evaluation.* 2nd ed. New Jersey: Prentice Hall.

This is one of the basic reference books on evaluation. Using nontechnical language, the book provides insight into the steps necessary to conduct a meaningful program evaluation, essentially providing a step-by-step guide. It includes a broad discussion of potential biases and how they can influence the interpretation of the results. The book includes an extensive list of references.

References

Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer. 2002. "Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment." *The American Economic Review* 92(5):1535–58.

Attanasio, O., and A. M. Vera-Hernandez. 2004. Medium and Long-run Effects of Nutrition and Child Care: Evaluation of a Community Nursery Programme in Rural Colombia. Working Paper No. EWP04/06, Centre for the Evaluation of Development Policies, Institute of Fiscal Studies, London.

Baker, J. 2000. *Evaluating the Impact of Development Projects on Poverty—A Handbook for Practitioners.* World Bank, Washington, DC.

Boruch, R. F. 1996. *Randomized Experiments for Planning and Evaluation: A Practical Guide.* Thousand Oaks, CA: Sage Publications.

Bourguignon, F., and L. A. Pereira da Silva, eds. 2003. *The Impact of Economic Policies on Poverty and Income Distribution—Evaluation Techniques and Tools.* World Bank, Washington, DC.

Caliendo, M., and S. Kopeinig. 2005. Some Practical Guidance for the Implementation of Propensity Score Matching. IZA Discussion Paper No. 1588, Institute for the Study of Labor, Bonn.

Campbell, D. T., and J. C. Stanley. 1966. *Experiment and Quasi-Experimental Designs for Research.* Chicago: Rand McNally.

Center for Global Development 2006. *A Major Step Forward on Impact Evaluation.* Available: http://www.cgdev.org/content/article/detail/8207/.

Cook, T. D., and W. R. Shadish. 1994. "Social Experiments: Some Developments over the Past Fifteen Years." *Annual Review of Psychology* 45:545–80.

Deaton, A. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore, MD: Johns Hopkins University Press.

Duflo, E., and M. Kremer. 2003. "Use of Randomization in the Evaluation of Development Effectiveness." Paper prepared for the Conference on Evaluation and Development Effectiveness, 15–16 July, World Bank, Washington, DC.

Dugger, C. 2004. "World Bank Challenged: Are Poor Really Helped?" *The New York Times*. 28 July. Page 4.

Dunn, G., M. Maracy, C. Dowrick, J. L. Ayuso-Mateos, O. S. Dalgard, H. Page, V. Lehtinen, P. Casey, C. Wilkinson, and J. L. Vazquez-Barquero. 2003. "Estimating Psychological Treatment Effects from a Randomized Controlled Trial with Both Non-compliance and Loss to Follow-up." *British Journal of Psychiatry* 183:323–31.

Easterly, W. 2001. *The Elusive Quest for Growth: Economists' Adventures and Misadventures in the Tropics.* Cambridge: The MIT Press.

Glazerman, S., D. M. Levy, and D. Myers. 2002. Nonexperimental Replications of Social Experiments: A Systematic Review. Interim Report/Discussion Paper Mathematica MPR Reference No. 8813-300, Mathematica Policy Research, Inc., Princeton, NJ.

Kapoor, A. G., 2002. *Review of Impact Evaluation Methodologies Used by the Operations Evaluation Department over 25 Years.* Operations Evaluation Department, World Bank, Washington, DC.

Karlan, D. 2001. "Microfinance Impact Assessments: The Perils of Using New Members as a Control Group." *Journal of Microfinance* 3(2):76–85.

Keller, S., and J. B. Schwartz. 2001. Final Evaluation Report: Contracting for Health Services Pilot Project (CHSPP). A Component of the Basic Health Services Project. ADB Loan No. 1447-Cam. Phnom Penh.

Kuehl, R. O. 2000. *Design of Experiments: Statistical Principles of Research Design and Analysis.* 2nd ed. California: Brooks/Cole.

Massachusetts Institute of Technology. 2006. *Poverty Action Lab*. Available: http://www.povertyactionlab.com/papers/.

Meyer, B. D. 1995. "Natural and Quasi-experiments in Economics." *Journal of Business and Economic Statistics* 13(2):151–61.

Miguel, E., and M. Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica* 72(1):159–217.

Moffitt, R. A. 2004. "The Role of Randomized Field Trials in Social Science Research. A Perspective from Evaluations of Reforms of Social Welfare Programs." *American Behavioral Scientist* 47(5):506–40.

Montgomery, R., D. Bhattacharya, and D. Hulme, 1996. "Credit for the Poor in Bangladesh." In D. Hulme and M.P. Mosley, *Finance Against Poverty*, Vol. 2. London and New York: Routledge.

Parker, S. W., and G. M. Teruel. 2005. "Randomization and Social Program Evaluation: The Case of Progresa." *The ANNALS of the American Academy of Political and Social Science* 599:199–219.

Pritchett, L. 2002. "It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." *The Journal of Policy Reform* 5:251–69.

Rajan, R., and A. Subramanian. 2005. Aid and Growth: What Does the Cross-country Evidence Really Show? IMF Working Papers 05/127, International Monetary Fund, Washington, DC.

Ravallion, M. 2003. "Assessing the Poverty Impact of an Assigned Program." In F. Bourguignon and L. A. Pereira da Silva, eds., *The Impact of Economic Policies on Poverty and Income Distribution—Evaluation Techniques and Tools.* World Bank, Washington, DC.

_____ . 2005. Evaluating Anti-poverty Programs. World Bank Policy Research Working Paper No. 3625, World Bank, Washington, DC.

Savedoff, W. D., 2006. "The Evaluation Gap: An International Initiative to Build Knowledge." Center for Global Development, Washington, DC.

Savedoff, W. D., R. Levine, and N. Birdsall. 2006. *When will We Ever Learn? Improving Lives through Impact Evaluation*. Center for Global Development, Washington, DC. Available: http://www.cgdev.org/content/publications/detail/7973.

Trochim, W. 1984. *Research Design for Program Evaluation: The Regression–Discontinuity Approach*. Beverly Hills: Sage Publications.

van de Walle, D., and D. Cratty. 2005. Do Donors Get What They Paid For? Micro Evidence on the Fungibility of Development Project Aid. Policy Research Working Paper No. 3542, World Bank, Washington, DC.

Weiss, C. 1998. *Evaluation.* 2nd ed. New Jersey: Prentice Hall.

Wooldridge, J. M. 2002. *Introductory Econometrics: A Modern Approach*. 2nd ed. Cincinnati, OH: South-Western College Publication.

World Bank. 2006a. *Living Standards Measurement Study of the World Bank*. Available: http://www.worldbank.org/lsms.

_____ . 2006b. *Poverty Impact Evaluations Database.* Available: http://www1.worldbank.org/prem/poverty/ie/evaluationdb.htm.