A Review of Recent Developments in IMPACT EVALUATION

February 2011

Asian Development Bank
Contents

Foreword v
Abbreviations vii
Figures viii

I. Introduction 1

II. The Evaluation Problem 2

III. Review of Evaluation Methods 4
   A. Before–After Comparison 4
   B. With–Without Comparison 5
   C. Difference-in-Differences Method 7
   D. Regression Discontinuity Designs 9
   E. Instrumental Variables 11
   F. Randomized Evaluation 12
   G. Summing Up 16

IV. Review of Impact Evaluations in the Main Development Sectors 17
   A. Education 17
      1. Conditional cash transfer programs 18
      2. In-kind transfer programs 19
      3. School construction and physical inputs to education 20
      4. Private participation 20
      5. Improving teacher quantity and quality 21
      6. School management reform 22
   B. Health 23
      1. CCT health programs 24
      2. Subsidized insurance programs 24
      3. School or community-based health interventions 26
      4. HIV/AIDS-targeted interventions 27
      5. Pricing and take-up of health services 27
   C. Infrastructure 28
      1. Transport 28
      2. Energy 30
      3. Water supply and sanitation 32
Several billion dollars of official development assistance each year flow to developing countries. Whether and how these funds are able to generate the intended project or program impacts are important issues that the development community must address in the face of competing demands for scarce resources. As a signatory to the 2002 "Monterrey Statement," the 2005 Paris Declaration on Aid Effectiveness, and the 2008 Accra Agenda for Action, the Asian Development Bank (ADB) is committed to better measuring, monitoring, and managing for development results.

Impact evaluation answers the need for empirical evidence showing that development interventions result in measurable impact. A systematic and rigorous assessment of project and/or program outcomes through impact evaluation supports evidence-based decision making, helps guide the design of more effective interventions, and contributes to the efficient allocation of scarce resources. Although the number of impact evaluation studies has been growing, there is still a large need for more studies to fill in the knowledge gaps in the effectiveness of various development interventions and contexts where these interventions work. ADB is responding to this need by better integrating impact evaluation in its project cycle through additional funding support for capacity building for and conduct of impact evaluation studies on selected interventions financed by ADB.

The purpose of this report is to serve as a reference in support of these initiatives. It first introduces the fundamental challenge of impact evaluation, which is to credibly attribute the impact, if any, to the intervention concerned. It then discusses the merits and limitations of various impact evaluation methods. This is followed by a survey of impact evaluation studies carried out in key development sectors including education, health, infrastructure, finance, and agriculture, among others. The survey focuses on what types of development interventions have been evaluated, what methods have often been applied, and what findings have been generated in these sectors. The report also presents six case studies that are deemed representative of, or innovative in, the selected sectors, with a view to providing development practitioners and researchers with practical examples of impact evaluation. Finally, the report discusses practical steps in implementing an impact evaluation exercise.
The report was prepared by a team led by Yi Jiang, former Economist at the Economics and Research Department (ERD) and currently Environmental Economist at East Asia Department, under the overall guidance of Juzhong Zhuang, Deputy Chief Economist and concurrently Assistant Chief Economist. Other contributors include Michiko Katagami, former Economist at ERD and currently Natural Resources and Agriculture Specialist at the Regional and Sustainable Development Department, and Hyun Hwa Son, Senior Economist at ERD. Ganesh Rauniyar, Natalie Chun, and many other colleagues at ADB provided valuable suggestions and comments at various stages. Lilibeth Poot and Lea Ortega provided excellent research assistance. Anneli Lagman-Martin, Eric Van Zant, and Cherry Lynn Zafaralla provided editorial assistance.

It is hoped that the review of recent applications in impact evaluation as well as discussions on the practical considerations in implementing impact evaluation will be of benefit to practitioners in and outside ADB in conducting impact evaluations with greater analytical rigor.

Changyong Rhee
Chief Economist
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADB</td>
<td>Asian Development Bank</td>
</tr>
<tr>
<td>ATE</td>
<td>average treatment effect</td>
</tr>
<tr>
<td>ATT</td>
<td>average treatment-on-the-treated</td>
</tr>
<tr>
<td>CCT</td>
<td>conditional cash transfer</td>
</tr>
<tr>
<td>CGD</td>
<td>Center for Global Development</td>
</tr>
<tr>
<td>DID</td>
<td>difference-in-differences</td>
</tr>
<tr>
<td>DIP</td>
<td>Dropout Intervention Program</td>
</tr>
<tr>
<td>DPEP</td>
<td>District Primary Education Program</td>
</tr>
<tr>
<td>EDUCO</td>
<td>Educacion con Participacion de la Comunidad</td>
</tr>
<tr>
<td>ICFES</td>
<td>Instituto Colombiano para la Evaluación de la Educación</td>
</tr>
<tr>
<td>ICS</td>
<td>International Child Support</td>
</tr>
<tr>
<td>ID</td>
<td>identification</td>
</tr>
<tr>
<td>IEG</td>
<td>Independent Evaluation Group</td>
</tr>
<tr>
<td>ITT</td>
<td>intent to treat</td>
</tr>
<tr>
<td>IV</td>
<td>instrumental variable</td>
</tr>
<tr>
<td>MDG</td>
<td>Millennium Development Goal</td>
</tr>
<tr>
<td>MFI</td>
<td>microfinance institution</td>
</tr>
<tr>
<td>MIT</td>
<td>Massachusetts Institute of Technology</td>
</tr>
<tr>
<td>NGO</td>
<td>nongovernment organization</td>
</tr>
<tr>
<td>PACES</td>
<td>Programa de Ampliacion de Cobertura de la Educacion Secundaria</td>
</tr>
<tr>
<td>PROGRESA</td>
<td>Programa de Educación, Salud y Alimentación</td>
</tr>
<tr>
<td>PS</td>
<td>propensity score</td>
</tr>
<tr>
<td>RCT</td>
<td>randomized controlled trial</td>
</tr>
<tr>
<td>RDD</td>
<td>regression discontinuity design</td>
</tr>
<tr>
<td>RPS</td>
<td>Red de Proteccion Social</td>
</tr>
<tr>
<td>SHG</td>
<td>self-help group</td>
</tr>
<tr>
<td>SME</td>
<td>small and medium enterprise</td>
</tr>
</tbody>
</table>
Figures

Figure 1  Before–After Comparison  
Figure 2a  Differences-in-Differences  
Figure 2b  Failure of Differences-in-Differences  
Figure 3  Regression Discontinuity Design  

5  
7  
9  
10
I. Introduction

Impact evaluation aims to answer whether and to what extent a development intervention has delivered its intended effects. For instance, evaluation of a rural water infrastructure project could show whether the investment has improved the health status of local people. This knowledge would enable evidence-based policy making and help improve the efficiency of resource allocation. The desire for more hard evidence on the effectiveness of development interventions has fueled a growing interest in impact evaluation in the international development community in recent years. For instance, the Center for Global Development (CGD) has been continuously promoting and sponsoring rigorous impact evaluation of social programs in developing countries (CGD 2006). A great number of interesting, policy-relevant impact evaluations focus on development economics even though the related key methodological innovations started in the field of labor economics.

The biggest challenge in impact evaluation is to isolate the impact solely attributable to the intervention concerned. Doing so requires a credible “counterfactual,” often in the form of a plausible comparison group, which in many cases is difficult to obtain. A rigorously conducted impact evaluation produces reliable impact estimates of an intervention through careful construction of the counterfactual using experimental or nonexperimental approaches. An impact evaluation study often involves evaluation design, data collection, economic modeling, and econometric analysis.

An earlier Asian Development Bank publication (ADB 2006) provides an overview of impact evaluation, including a brief presentation of the methodology, two case studies, and discussions on practical issues related to carrying out impact evaluation. Building on that publication, this report has three objectives. First, it provides more in-depth discussions on the evaluation problem as well as major evaluation methods. Second, it carries out a comprehensive survey of impact evaluation studies in key development sectors with a view to providing development practitioners and researchers with a clear picture of what types of development interventions have been evaluated, what methods have often been applied, and what findings have been generated in these sectors. Third, it presents six well-known impact evaluations as case studies.

The rest of this report is organized as follows. Chapter II describes the fundamental problem underlying impact evaluation, which poses a substantial challenge to development researchers and practitioners. Chapter III reviews commonly used evaluation methods, focusing on their merits and limitations as well as the circumstances under which they are able to address the evaluation problem. Chapter IV reviews the literature on impact evaluation in major development sectors, and Chapter V describes six evaluation studies in detail. Chapter VI discusses experiences and lessons learned from existing practice in four key areas. Chapter VII concludes.
Impact evaluations attempt to assess the “treatment effects” of a development program, which is the difference between the outcome of an economic agent (individual, household, firm, village, etc.) participating in a program (treated) and that of the same agent not participating in the program (untreated). However, a particular agent can only be in one state, treated or untreated, at any given point: there is no way to observe both outcomes for the same agent simultaneously. In other words, a true counterfactual that shows how an agent that has (not) been exposed to a program would have fared in its absence (presence) is always unavailable. In this sense, the evaluation problem is one of “missing counterfactual” and falls into what Holland (1986) refers to as the “fundamental problem of causal inference.” See Appendix 1 for a technical formulation of the evaluation problem.

Comparing the same agent before and after a program is unlikely to be a good strategy because many factors that affect the outcome could occur during the program. For instance, an unemployed individual completes his job training when the economy recovers from a recession. Despite getting a job after the training, it would be difficult to tell, based on the before–after comparison, whether his re-employment is due to the job training or not. To convincingly evaluate the treatment effects of a program, one must rely on comparisons between agents participating in the program and those not participating in the program. A good comparison group should provide reliable information on what would have happened to the “treatment group” had it not been exposed to the program.

Selecting nonparticipants as the comparison group would be an intuitive choice. But selection bias could arise when decisions on program participation are not made randomly, but based on some unobserved factors that are correlated with the outcomes. For instance, individuals may decide whether or not to participate in a program based on factors such as preferences, opportunity costs, and expected gains from the program; and program administrators may have a set of criteria by which they screen interested individuals to determine program participation. If these factors are private information to the individuals or program administrators and not accessible to the evaluator, but are correlated with the outcomes of the program being evaluated—a situation called “selection on unobservables”—the evaluation results could be biased. Selection on unobservables makes it difficult to establish a credible counterfactual: those who do not participate in the program are generally a poor comparison group for those participating, since it is difficult to tell whether the differences in outcomes of the two groups are due to the differences in the unobservables or the treatment received from the program.
A simple example helps illustrate why selection on unobservables invalidates nonparticipants as a credible comparison group. Suppose two people living in the same community are both eligible for a government job training program. The training center is 5 kilometers away from their community. One person has a bicycle to get there, but the other has no transport and has to forego the training. Two job opportunities emerge when the training ends. They are almost the same except that one is within the community and the other near the training center. The job near the training center pays $10 more per month than the job within the community. The person with the bicycle takes the higher-paying job, while the other person takes the local one. Someone observing only the program participation and wages of the two persons may attribute the wage difference of $10 to the training program. A more careful evaluator may argue that commuting costs, say $4, should be factored into the wage difference and conclude that the training leads to a $6 increase in wage. However, observing that the differences in program participation and wages are both caused by bicycle ownership, one would agree that the actual effect of the training program is being exaggerated still and probably much smaller.

The problem of selection on unobservables introduces selection biases and plagues impact evaluation and other empirical social science studies that aim to draw causal inferences using observational (nonexperimental) data. The goal of impact evaluation is to obtain reliable estimates of program impacts by removing or mitigating the selection bias. Econometric methods have been developed for evaluation using observational data. The validity of each method depends on the specific context to which it is applied and the soundness of the underlying assumptions. For instance, regression and matching models using cross-sectional data essentially assume that factors determining program participation are all observable to the evaluator. This assumption is implausible in many cases where unobservables play an important role in determining program participation. In recent years, randomized field experiments have emerged as a promising way to address the problem of selection bias in evaluating impacts of development programs. The randomized controlled trials (RCTs), which have a long tradition in biological and medical research, assign program participation randomly. In theory, this approach generates a comparison group of the “gold standard,” producing unbiased estimation of program impacts. In practice, however, as the experiments are often carried out in the field instead of a laboratory, a number of issues need to be addressed even in cases of RCTs, such as in program design and implementation, and results interpretation.
Methods used in impact evaluation can be classified into two broad categories: nonexperimental and experimental. Nonexperimental methods—including before–after and with–without comparisons, cross-sectional regression, matching, difference-in-differences (DID), regression discontinuity design (RDD), and instrumental variables (IV)—use observational data and rely on different assumptions about underlying selection processes. For instance, before–after comparison, cross-sectional regression, and matching usually assume the absence of selection on unobservables, while DID, RDD, and IV take selection on unobservables seriously. See Appendix 2 for the technical formulation of nonexperimental evaluation methods. The experimental approach typically evaluates program impact through RCTs. In this section, we introduce both nonexperimental and experimental evaluation methods focusing on their respective merits and limitations, rather than the technical details, and the circumstances in which each method best fits.

A. Before–After Comparison

Before–after comparison is conceptually straightforward and probably involves the least additional cost for evaluation design and data collection. Basically, the method compares the preprogram and postprogram outcomes of a group of economic agents participating in the program. For the average effect on the participants, one needs only to subtract the preprogram mean from the postprogram mean. When individual data are available and some variables vary over time, one may do a regression analysis, with the outcome as the dependent variable and a time indicator (taking the value of 1 for postprogram observations and 0 otherwise) as well as time-varying variables as the explanatory variables. The coefficient in front of the time indicator represents the treatment effect of the program.

However, the chance for a before–after comparison to yield reliable estimates of program impact is rather small, since the key assumption for the method to work—that the status of participants (such as health or education outcomes) would otherwise be the same had the program not been administered—rarely holds in most circumstances. Trends over time, economywide shocks, and other developments or interventions during program implementation can cause the outcome to change over time even in the absence of the program. It is therefore invalid to attribute

---

1 Here, the discussion is focused on quantitative methods for impact evaluation. In many cases, qualitative methods such as in-depth interviews can provide useful supplements, both ex ante and ex post, to the quantitative analysis.
the changes in the outcome entirely to the program if these confounding factors are significant and cannot be ignored. This point is illustrated in Figure 1, where individuals are assumed to take part in a job training program at time $t$. If one only compares their average wages before and after the program, at $t-1$ and $t+1$ respectively, the evaluator may mistakenly conclude that the program has no effect on the recipients’ average wage, at $900 before and after the program, whereas the fact is that the downward trend in wages would continue even without the program, declining to $700. Thus, the actual impact of the job training program is $200. The before–after comparison in this case underestimates the actual positive effect of the job training program.

**Figure 1. Before–After Comparison**

![Figure 1. Before–After Comparison](image)

**B. With–Without Comparison**

Instead of comparing the same group over time, this approach compares the outcomes of two groups—participants and nonparticipants—at the same postprogram time point. The idea is that if everything else is equal or can be fully controlled for in an empirical model, the difference in the outcomes of the two groups can be viewed as the treatment effect of the program. The validity of this method relies on two key assumptions. First, conditional on the observed influencing variables—sometimes called covariates or control variables—the treatment status of an agent is independent of the potential outcomes of participating in the program. This assumption requires the absence of selection on unobservables, also referred to as “unconfoundedness” in the literature (e.g., Imbens and Wooldridge 2008). This means that all covariates that are associated both with the treatment status (that is, whether or not being treated) and the potential outcomes can be observed by the evaluator and taken into account in the model. The second assumption necessary for the with–without comparison to work is called the “overlap assumption,” which requires that there are both treated and untreated agents for all possible values of the covariates.

With–without comparison is most often done by estimating a multivariate linear regression model using cross-sectional observations of participants and nonparticipants. The coefficient in
front of the treatment indicator is taken as the program impact. However, linear regressions often yield results highly sensitive to minor changes in model specification. Consequently, the estimated treatment effects could be biased in cases where the linear relationship does not hold. To address the specification bias associated with linear regression, recent literature has moved to semiparametric or nonparametric methods. One particular example of such methods is matching.

The basic idea of matching is to construct the counterfactuals for the treated agents with the untreated ones without imposing strong assumptions on model specification. Intuitively, the matching untreated agents come from the neighborhood, defined based on the observable characteristics in various ways, of each treated agent. The mean impact of the program is the average of the differences in outcomes between the matched treated and untreated agents. However, when the matching needs to be carried out on a large number of characteristics, the high dimension of matching variables becomes a challenge. Propensity score matching, however, offers a solution to this problem. The propensity score equals the probability of receiving the treatment conditional on observable characteristics. It summarizes the information of all observables into one variable and thus matching based on the propensity score deals with a much smaller dimension of variables. While matching estimation generally involves a flexible function of matching variables, model specification bias is less a concern for matching than for regression models. Nevertheless, the overlap requirement is more likely to be an issue for matching, especially when the sample size is small.

The primary concern about with–without comparison is selection on unobservables. A number of ways have been developed to test the plausibility of the absence of selection on unobservables necessarily assumed for the cross-sectional regression and matching (e.g., Rosenbaum 1987, Heckman and Hotz 1989). One way is to identify two comparison groups, make one the “pseudo” treatment group, and compare the two. A nonzero “treatment” effect for the pseudo treatment group would suggest that at least one of the two groups is invalid. However, this is not a sufficient test in that the result that the two comparison groups are indifferent in the outcomes does not necessarily uphold the assumption of the absence of selection on unobservables since both could be biased comparison groups in the same direction. A second test is to measure the treatment effect on a variable, which is unlikely to be affected by the program. Lagged outcomes, whose values are determined before the program, are good candidates for such tests. If the impacts are estimated to be zero on these variables, the absence of selection on unobservables assumption holds more plausibly.

The with–without comparison approach may seem to be less demanding on data requirements as it does not require baseline information. However, to mitigate concerns with selection on unobservables, it is necessary to collect and control for as many covariates as possible. For instance, in evaluating the impact of a water supply project on child health, one needs to know the hygiene behaviors of children, which are often associated with their mothers’ education. Moreover, the overlap assumption requires a large number of observations, particularly for the comparison group.

In sum, with–without comparison provides an option when impact evaluation is to be carried out after the intervention has started or has been completed and the baseline data are missing. Choosing a right comparison group is critical to measuring treatment effects given wide concerns over the selection bias. There are ways to test the severity of the selection bias, but the test results may not provide sufficient evidence for the absence of the selection bias. It is also important for impact evaluation studies to collect adequate samples as well as sample characteristics that could potentially be correlated with program outcomes. Finally, matching is being substituted for traditional linear regression as a more sophisticated econometric technique to estimate the treatment effects.
C. Difference-in-Differences Method

The DID method has become popular in empirical economics since its genesis in Ashenfelter (1978) and is probably the most widely used method for impact evaluation. It is applicable when data are available for the time periods before and after the program for the treatment group as well as the untreated comparison group. The average treatment effects are obtained following a two-step differencing procedure. The first step is to calculate the difference of the mean of an outcome variable before the implementation of the program from the mean after the implementation of the program for the treatment group and untreated comparison group, respectively. The second step is to calculate the difference between the two differences.\(^2\) Put in econometric language, DID is largely a two-period panel model with fixed effects assumed for the treatment status and time period. The coefficient for the interaction term of the treatment status and time period (equal to 1 if an observation is from the treatment group and postprogram period and 0 otherwise) is interpreted as the mean impact of the program. One advantage of such a panel model over calculating differences in means twice is that other observables that vary across individuals and over time are controlled in the model.

Unlike before–after and with–without comparisons, which basically ignore the selection on unobservables, DID addresses this issue in a somewhat clever way. It allows the comparison and treatment groups to be different in observable or unobservable ways from the outset, but assumes that the outcome variables of the treatment group would have experienced the same changes over time as those of the comparison group in the absence of the program. The over-time difference in the comparison group would capture the impacts of external shocks or interventions that affect both the treatment and comparison groups other than the program. DID can produce consistent estimates of program impact in the presence of fixed differences between treatment and comparison groups that are unobservable to the evaluator. In other words, DID relies on a less restrictive assumption than those for before–after and with–without comparisons to identify the program impact. Let us consider the job training case as presented in Figure 2a.

---

Figure 2a. Difference-in-Differences

---

\(^2\) The same result could be obtained by first differencing across groups for the before and after periods, respectively, and then differencing between the two periods in the cross-group differences.
Now we observe a group of program participants and a group of nonparticipants. Before the program, both groups’ wages were trending up while the treatment group earns $200 less on average. Assume that the preprogram trend as well as the difference between groups will remain the same over time without the program, which means that the treated group would earn $1,400 at \( t + 1 \) without the program. Thus, the true program effect is an increase of $200 ($1,600 – $1,400) in the average wage. DID first calculates the average wage increases for participants ($1,600 – $1,000 = $600) and nonparticipants ($1,600 – $1,200 = $400), respectively, and then subtracts the wage increase of the nonparticipants from that of the participants ($600 – $400 = $200) to reach the unbiased estimate of the treatment effect. In contrast, before–after comparison produces an overestimate ($1,600 – $1,000 = $600) of the treatment effect arising from failure to take into account the upward trend of wages. On the other hand, with–without comparison yields a zero program effect as it does not take into account the preexisting difference between participants and nonparticipants.

DID will generate biased estimates if the key assumption is violated, i.e., the treatment groups would have followed a distinct growth path in the outcome from that of the comparison group without the program. One well-known example is the Ashenfelter’s Dip (e.g., Heckman and Smith 1999), which refers to the phenomenon where the earnings of participants in government training programs drop before program entry. Since the dip in earnings is found to be primarily transitory, participants would have recovered somewhat even without undertaking the training. On the contrary, individuals not experiencing an income decline are less likely to self-select into the program and less likely to have a similar income growth pattern to that of the participants. DID using these nonparticipants as the comparison group would produce upward biased estimates of the program impact.

Figure 2b provides a simple illustration of the case of Ashenfelter’s Dip. The participant group experienced a deep drop in average wage from $1,000 to $700 at \( t – 1 \). After participating in the program, the group’s average wage increased to $1,400 at \( t + 1 \). Meanwhile, the nonparticipants’ average wage rose from $1,200 to $1,600 between \( t – 1 \) and \( t + 1 \). However, as the wage drop is temporary, the participants’ wages would recover to $1,400 as of \( t + 1 \), even if they did not participate in the program, suggesting a true program effect equal to zero. By comparison, the two groups would have different paces of wage increase without the program, which violates the necessary assumption for DID to apply. The DID estimate of treatment effect equals $300 [(1,400 – 700) – (1,600 – 1,200)], which overestimates the actual program impact.\(^3\) Therefore, it is critical to select the comparison group for DID especially when the outcomes of interest involve dynamic variables and the prior outcomes affect program participation.

In addition to the approaches discussed in the previous section in assessing the quality of the comparison groups, i.e., comparing multiple comparison groups and inspecting the exogenous variables that should not be affected by the program, it is useful to examine whether the treatment and comparison groups tend to move in parallel before the program. To do this, a long time series of the outcome measures is desirable, although collecting them may add extra burden to the evaluation project. In theory, data collection does not need to include time-invariant characteristics since they will be automatically removed from a DID estimation. However, they are highly informative with regard to who is participating in the program and why—questions that could attract policy interest and are useful for assessing the plausibility of the comparison group. Finally, special attention needs to be paid to the attrition issue in evaluation design and data collection. Nonrandom

---

\(^3\) Neither before–after comparison nor with–without comparison yields unbiased estimates in this case.
withdrawal from the program or missing postprogram observations could undermine the unbiased characteristic of DID estimation.

**Figure 2b. Failure of Difference-in-Differences**

![Graph showing wages over time for untreated and treated groups with a discontinuity at the cutoff point.]

**D. Regression Discontinuity Designs**

In certain types of programs the assignment of treatment is determined by an administrative agent on the basis of a continuous (forcing) variable and a threshold value. For instance, the Government of Chile introduced an education enhancement program (P-900) in 1990 targeted at about 900 low-performing public schools. Those with mean fourth-grade test scores in 1988 below a cutoff value received interventions such as infrastructure improvement and teacher training workshops, while those with mean scores above the cutoff value did not (Chay, McEwan, and Urquiola 2005). RDD is appropriate in such cases to overcome potential biases in estimating the impact of the program (as discussed below).

The rationale behind the RDD method is that under certain conditions individual units around the cutoff point should be highly similar except for the treatment status. Therefore, untreated units right below the cutoff constitute a sound counterfactual for the treated units right above the cutoff. Any difference in the outcomes between these two groups can be viewed as a causal effect of the program treatment. Figure 3 illustrates this idea. The outcome would have a smooth, nonlinear relationship with the forcing variable (shown as the black solid curve and the dashed red curve) if the program is not available. When the program is offered with the eligibility determined by the individual's forcing variable relative to the cutoff, c0, the outcomes of the treated units would move to the solid red curve given positive program impact. The jump at the cutoff, Δ, should correctly capture the program impact. Figure 3 also demonstrates that linear model specification (dashed black and red lines) may generate a biased estimate, such as Δ', of the program impact. Therefore, impact estimation by RDD should involve a flexible function of the forcing variable.
Two situations may arise from implementing the assignment rule. One is that the assignment rule with respect to the forcing variable is strictly carried out. All units above the cutoff receive the treatment while all units below the cutoff do not. The other is that the assignment rule is not enforced seriously. In the Chile P-900 program, the schools preselected on the mean test scores were subject to government officials’ review and some were removed from eligibility. While in the first situation, called “sharp RDD,” one observes a zero-to-one change in treatment status at the cutoff, in the second situation, “fuzzy RDD,” one sees a sizable jump in treatment probability at the cutoff. Although sharp RDD and fuzzy RDD call for different estimation techniques and the results may have somewhat different interpretations, the key idea of exploiting the discontinuity in treatment status at the cutoff point to identify program impacts remains the same.

Figure 3. Regression Discontinuity Design

Two concerns need to be looked into before applying RDD in program evaluation. First, if there is another unobserved factor that also displays discontinuity at the cutoff value of the forcing variable, we would be unable to determine whether it is the program or the unobserved characteristic that is responsible for any difference in the outcome. For instance, a second intervention assigned by the same rule would prevent us from disentangling the two treatment effects. The other concern is whether individuals manipulate their forcing variable to change the treatment status. Hypothetically, had the P-900 program selected schools based on the 1990 test scores instead of the 1988 test scores, and schools had information on the cutoff score, some schools could have reported lower scores to be eligible for additional resources from the program. Checks on these possibilities can be done formally through statistical tests. Graphical checking, however, provides an alternative way to assess the credibility of the RDD (Imbens and Lemieux 2008). To see whether the treated and untreated units are indeed similar at the cutoff except for the treatment status, one can plot the covariates against the forcing variable. That all the covariates are smoothly crossing the cutoff point suggests similarity between the treated and untreated. When the distribution of observations along the forcing variable is plotted, clustering of observations on one side of the cutoff point indicates that further exploration is needed to look into the possibility of manipulation.
Although application of RDD requires that programs have a special assignment mechanism, such programs are actually not rare in the developing world and a significant number of RDD studies have emerged since the method became widely known to evaluators. A thorough understanding of the assignment mechanism of the concerned program is very important for choosing between sharp and fuzzy RDD and assessing the extent of manipulation. Baseline data are not necessarily needed for RDD, but a large sample with plenty of observations close to the cutoff point from both sides is essential for the estimates to possess enough statistical power. In theory, RDD only identifies program impacts on units close to the cutoff. Caution should be given to extrapolating RDD results to the entire sample.

E. Instrumental Variables

Put simply, selection bias results from the presence of unobserved factors that influence an individual’s decision or eligibility for program participation, while being correlated with the outcome of interest. The IV approach to solving this difficult problem is to identify one or more variables (instruments) that affect program participation and do not affect the outcome in any way except through program participation. However, it is often very difficult to find qualified IVs that satisfy both conditions. The IV estimation will be biased in finite samples when the IVs only weakly affect program participation (Bound, Jaeger, and Baker 1995). The second condition requires that IVs are exogenous or statistically orthogonal to the unobservables correlated with the outcome. Since this is not a testable condition, a strong case needs to be made to defend the exogeneity of the instruments chosen. Empirical studies often defend the selection of instruments on the basis that they are external to the outcome variables of interest, which does not necessarily imply the instruments are exogenous (Deaton 2009). Using invalid IVs could result in estimates with exacerbated bias or bias in the opposite direction.

A popular way to identify IVs is to use so-called natural or quasi-experiments (Meyer 1995). Natural or quasi-experiments refer to settings where the nature, public policy, legislative changes, or other events induce some exogenous variations in the key explanatory variables. They are not experimentally controlled by researchers for the purpose of a causal study. However, they effectively serve the purpose in the same way as a controlled experiment does in that they provide an exogenous source of variation in the explanatory variable. An example of using this type of IV is Duflo and Pande (2007), which uses land gradient as an instrument for dam construction in a study of the impacts of dams on poverty, in which it is argued that the land gradient where a dam is constructed does not affect poverty otherwise. Again, the question whether land gradient is indeed a good instrumental variable or not is still left open (Deaton 2009).

Sometimes, program eligibility or placement is determined on a random basis. However, those who are randomly assigned program eligibility or treatment would make a decision on whether to participate in the program or not. In the meantime, those not eligible or for whom the program is not available may find their way to get into the program or receive comparable treatment. Selection bias arises if we compare the treated and untreated individuals directly. In these cases, however, the random program eligibility or availability could be employed as an instrumental variable for the final treatment status (Heckman 1996) on the grounds that eligible individuals are generally more likely to participate in the program than the ineligible ones, and the eligibility is uncorrelated to the outcome of interest since it is random.
**F. Randomized Evaluation**

Concerns with the bias in measuring impacts of social programs using observational data have led to an increasing enthusiasm for conducting randomized controlled trials in a real-world context for program evaluation. Randomization is believed to generate gold standard evidence, which would be subject to less criticism. A randomized experiment randomly assigns program treatment to the units in the field and takes measures to prevent the treated and untreated from interacting with each other. In this circumstance, the untreated are thought to be a sound counterfactual for the treated units. In the development field, a lack of solid evidence that past projects have reduced poverty and promoted economic growth has led to the wide acceptance of randomized evaluation as a convincing way to demonstrate the impacts of development interventions (Banerjee and Duflo 2008).

In addition to producing reliable estimates of program impacts, RCTs also allow researchers to study the effects of a single specific intervention, which tend to be veiled in data observed on a development package encompassing multiple interventions. This allows testing important, creative theories that cannot be done without a controlled experiment (e.g., Ashraf, Karlan, and Yin 2006), and pursuing innovative measurement or data collection (e.g., Olken 2007).

The most common situation in which randomized evaluation can be carried out is a pilot project preceding a large-scale program. Carefully evaluated pilot projects allow stakeholders to determine whether the program has had the intended impacts and to make informed decisions on whether to extend the scale. For example, the well-known Programa de Educación, Salud y Alimentación (PROGRESA)—which reached 2.6 million families in Mexico (10% of families) with a budget of $800 million (0.2% of gross domestic product) by 2000—started in 1998 with a pilot project covering 1% of the total 50,000 targeted communities. The positive results obtained by randomized evaluation gained extensive political support for scaling up the program. Duflo (2004) provides another example, in which the randomized evaluation failed to justify a full-scale two-teacher program in India on a cost–benefit basis.

Often, demand for a program with a fixed budget is so high that the subscriptions to participate exceed capacity. Randomly selecting participants, for example by lottery, is probably the best way to address fairness or transparency concerns with the selection procedure in this case. It also facilitates assessment of the program impact. For instance, Colombia allocated vouchers for private schools by lottery—the program evaluated in Angrist et al. (2002). In some cases a program, though targeted on a large scale, will be introduced in a sequence of phases. Randomized evaluation may be possible in such a way that a proportion of beneficiaries is randomly selected to receive the program in each phase, with those not yet selected being the control group. The evaluation of a remedial education program in Banerjee et al. (2003) is such a case.

Although RCTs are gaining enormous momentum in program evaluation, it is not without criticism (e.g., Heckman 1992, Heckman and Smith 1995, Deaton 2009, Ravallion 2009). The fact that there is an ongoing experiment might lead individuals to change their behavior and thereby bias the evaluation results, introducing randomization bias. For example, a control group may act differently after being informed that a program will be expanded to cover them. One way to address randomization bias is not to inform individuals, whenever proper, that they are excluded from the program.4 See the box for a discussion of design and implementation issues in RCTs.

---

4 In medical trials, people who are assigned into the control group are often given a vitamin pill without being told.
Design and Implementation Issues in Randomized Controlled Trials

When the project team decides that randomization is the appropriate way to evaluate the impacts of an intervention, a series of design and implementation issues needs to be sorted out. Duflo, Glennerster, and Kremer (2008) provide practical guidelines for conducting a randomized evaluation, including a detailed discussion of these issues, as summarized below.

**Level of Randomization**
When the possibility is open, one needs to choose whether to randomize the intervention at the level of the individual, the household, or the community, etc. Randomization at an aggregate or group level, say, the village, may often be easier and cheaper to implement. It allows researchers to capture the spillover effects when the untreated gets exposed to the treatment in some way. Further, it mitigates potential resistance to cooperation from the control group, which sometimes arises when the intervention, perceived as desirable, is randomized by individuals within a closed community. On the other hand, the larger the group subject to randomization, the larger the sample size needed to achieve a given statistical power, which puts higher demands on budgets and other inputs.

**Testing Different Treatments Simultaneously**
A development program is often a package of interventions. It is of great interest to know which interventions or combinations of interventions work as well as how the interventions interact with each other. Crosscutting designs can be employed in experiments to answer these questions. For instance, two interventions, A and B, are being planned. The researcher may generate four random groups: one treated with A and B, one A only, one B only, and one no interventions. By this design, we can test not only the effects of intervention A alone, B alone, and a combination of A and B, but also examine whether A (B) has a different effect when combined with B (A).

In cases when the costs of crosscutting designs are too high, the evaluator faces a choice between evaluating individual interventions separately and evaluating the combined package of interventions. Policy makers may prefer the latter in that it is more likely to be shown as effective. However, economists’ concern with the combined intervention package is that the evaluation results convey little knowledge of which intervention(s) caused the response, which is important for generalizing lessons learned. In practice, it is plausible to first evaluate the combined package and then examine individual components if the package is proven to work.

**Sample Size**
Sample size is one of the key issues that need to be elaborated in experimental designs. A sufficiently large sample enables researchers to reject the null hypothesis of zero treatment effect at a given statistical significance level with a given effect size. In other words, the design owns sufficient statistical power for testing the treatment effect. However, oversampling means high cost of data collection and waste of statistical power. Sophisticated statistical calculation is generally needed to determine the optimal sample size, taking into account the available budget. A related issue is the division of the sample between the treatment and comparison groups. Duflo, Glennerster, and Kremer (2008) suggest that an equal division

*continued on next page*
between treatment and comparison groups is optimal if the main cost of the evaluation is data collection. A larger comparison group is desirable if the treatment is much more expensive than data collection. In the case of multiple treatments, the sample size of the comparison group should be larger than that of each treatment group.

**Baseline Data Collection**

In principle, randomized evaluation does not rely on baseline data to identify and measure program impacts in that it balances the characteristics of the treatment and comparison groups. However, baseline data offer a few benefits for experimental trials. First, because the variables collected through the baseline survey partly explain the variability in the outcomes, controlling these variables increases statistical power and thus reduces the required sample size. Thus, the trade-off is to conduct a large experiment without baseline data or a small experiment with baseline data. Second, the availability of baseline data makes it possible to investigate heterogeneous impacts varying by pre-intervention conditions. Third, a check on the randomness of the experiment can be done with baseline data. Nevertheless, collecting baseline data in retrospect in the post-intervention survey is not recommended due to concerns with the contamination of recalled information. Whenever appropriate, available administrative data should be exploited as a substitute for the baseline survey.

**Partial Compliance**

Interventions are often randomized across groups or communities, even if they are targeted at the individual level. In these cases, a fraction of individuals in the treatment group may not take up the treatment and some individuals in the control group may find a substitute for the program in some way, which sometimes is referred to as partial compliance. When partial compliance occurs, the randomization only affects the probability of an individual being exposed to the program. Comparing treated individuals to the untreated is likely to reintroduce selection bias because those deviating from their group assignment are unlikely to be random. One valid approach is to compare all individuals in the treatment group to all those in the comparison group, which yields the so-called intention-to-treat effect. The intention-to-treat effect is a parameter of interest in some contexts, but not so in many others. Certain conditions satisfied, the randomization can be employed as an instrumental variable for the actual treatment to estimate a local average treatment effect—the effect on those complying with initial group assignment (Imbens and Angrist 1994). If the effect on everybody in the treatment group is really what is of interest, a smaller sample size may be preferable and great care needs to be taken to ensure full compliance.

**Attrition**

Attrition occurs when some individuals are absent from final data collection although they were in the original sample design. While random attrition is possible, nonrandom attrition is more likely and may cause serious bias in the estimation results. For instance, highly skilled workers may be more likely to migrate out of a program region than low-skilled ones. Attrition imposes a significant challenge on program administrators and is difficult to solve ex post. Special attention and measures may be essential to prevent or reduce attrition. Tracking those dropping out of the program helps in assessing the potential bias and mitigating the bias, but it usually requires lots of resources.

Another major concern with RCTs is the external validity of the results. Recognizing that randomized evaluations have higher internal validity (consistent impact estimates) than nonrandomization, the question often asked is to what extent the results can be generalized, or whether one would get the same results if the same experiment were carried out in a different environment and/or by a different implementer. To answer this question confidently, replication studies have to be conducted in different locations with different teams. They may confirm or refute the findings in the initial studies (e.g., Bobonis, Miguel, and Puri-Sharma 2006; Banerjee et al. 2008). Either case contributes to knowledge about the effectiveness of the intervention and the context in which it works. In this sense, systematic replication efforts should be encouraged and highly valued.\(^5\) Certainly, the issues associated with external validity do not only arise from experimental evaluation. Many studies applying nonexperimental methods estimate program impacts for a subgroup of the population, which is not necessarily representative of the entire population.

A related concern is that the estimated effects for pilot projects may not be generalized when the program is scaled up. Consider an education-enhancing project first piloted on a small scale, which is demonstrated to result in a better job market performance among beneficiaries. When it is to be administered for the whole region or country based on positive evidence from the pilot project, similar effects should not be assumed. When such a project covers the entire job market, it will increase the overall supply of skilled labor in the market and push the equilibrium returns to education downward. In other words, experimental assessment of a pilot project often reveals partial equilibrium effects, which are likely to be different from the general equilibrium effects resulting from scaling up the project. Duflo, Glennerster, and Kremer (2008) suggest that randomization at a level where general equilibrium effects work—in this case, in an integrated regional job market—would help capture the general equilibrium effects. However, it would be less helpful if general equilibrium effects work at the national level when one can only randomize at the regional level.

Most impact evaluations focus exclusively on the mean impact of a program on either the population as a whole or on participants, and the mean treatment effects are also the most convenient statistics to estimate by randomized controlled trials. However, there are other parameters of interest to the policy maker. For instance, one may want to know the median gain through the program or the fraction of participants who have benefited. What the experiment offers are the distributions of outcomes in treatment and control groups. It does not yield the distribution of the treatment effects. There is no way to know the quantile (e.g., median) treatment effects without additional assumptions because the quantile treatment effects are not equal to the differences in quantiles of the treatment and control outcomes. Banerjee and Duflo (2008) argue, however, that experiments retain advantages over nonexperiments in that experiments invoke fewer assumptions to know other important aspects of the treatment effects. In addition, they suggest exploring heterogeneous effects across subgroups defined by covariates with experimental data while the subgroups are specified ex ante with some guidance from theory.

---

\(^5\) One situation in which replication is less informative is when the randomized experiments can only be carried out and evaluated in very specific locations or with specific program implementers, precisely because they involve randomization, or other program implementers do not agree to randomize in their programs (Heckman 1992).
G. Summing Up

To conclude, before–after and with–without comparisons are less likely to produce credible estimates of program impact, although they may demand less evaluation planning and data collection. RDD and IV are powerful econometric techniques to evaluate program impacts. Nonetheless, they are only applicable in special settings where program assignment is determined or affected in a particular way. Although program implementers may be able to intentionally create such a program assignment mechanism (e.g., based on a continuous score) to facilitate evaluation, such occasions are relatively few. From a practical point of view, the DID method is recommended in that it addresses a significant portion of selection issues that before–after and with–without comparisons fail to take into account, while being more broadly applicable than RDD and IV. To conduct a quality impact evaluation with the DID method, one needs to synchronize evaluation with the program and carefully select the comparison group.

Evaluation through randomized controlled trials is promising, especially for pilot projects intended to indicate whether a scale-up is desirable. A good experiment often combines policy interest with economic theory and creative design. Programs targeted at individuals or local communities are likely to be strong candidates for randomized evaluations (Duflo and Kremer 2008). In practice, many such randomized evaluations have been conducted in health and education sectors, with some in local governance, finance, and agriculture, and on issues related to gender and ethnicity. In contrast, nonexperimental approaches still seem to be essential in evaluating projects in sectors such as transport and energy.

\[\text{6 See project list by theme at J-PAL (2011) for a sample.}\]
IV. Review of Impact Evaluations in the Main Development Sectors

This section reviews selected impact evaluations that have been implemented in major development sectors including education; health; infrastructure (particularly transport, energy, and water supply and sanitation); microfinance; agriculture; and others. The review focuses on three key aspects: (i) typical interventions being rigorously evaluated in each sector, (ii) evaluation methods commonly used, and (iii) key evaluation findings.

A. Education

Education is one of the sectors in which impact evaluations have been carried out most actively. Education interventions take various forms, and most are designed to either remove access barriers to education for the poor or improve education supply and quality. In this review, we group the education interventions into six broad categories: (i) conditional cash transfer (CCT) programs that involve offering cash grants to poor families conditional on their children’s school enrollment and/or attendance; (ii) in-kind transfer programs that attempt to improve education outcomes through food or goods, instead of cash; (iii) school construction and provision of physical inputs to education; (iv) initiatives to encourage private sector participation in the education sector, which is considered a quick and efficient way to increase public education supply; (v) programs targeted at improving teacher quantity and quality; and (vi) school management reforms.

Education interventions in general fit well with the evaluation framework in that they usually deal with a large number of units (students, teachers, classes, or schools) with some receiving the treatment and others not. This feature facilitates the construction of counterfactuals using the untreated units, unlike in a systemwide intervention that affects everybody.

When using observational (nonexperimental) data to evaluate an education intervention, one encounters the typical evaluation challenge discussed earlier: people do not take the treatment in a random way. Parents are well motivated to make the decision on how much schooling their children should get. Teachers with better skills may be more passionate about the job than those without. A program may be provided to schools or regions where it is affordable or is expected to have the largest effect. When evaluators fail to observe the key information underlying the behavioral choice or program assignment, a naïve comparison of the treated and untreated will produce incorrect estimates of the impacts.
The DID method is probably the most practical and widely used method to address some of these issues in nonexperimental evaluations of education sector interventions (e.g., Attanasio, Meghir, and Santiago 2005; Chaudhury and Parajuli 2006; Chin 2005; Duflo 2001; Ravallion and Wodon 2000). Meanwhile, in recent years, there has been growing interest in conducting randomized field experiments and evaluating education interventions with experimental data. Mexico’s PROGRESA and Colombia’s Programa de Ampliación de Cobertura de la Educación Secundaria (PACES) are the best-known nationwide education (and health) programs that reach the beneficiaries in a random way. Not only did the programs meet the challenge of allocating limited public resources in a fair and transparent way, randomized implementation also facilitated objective and accurate evaluations of program effects. There are also small-scale randomized education programs, often introduced as pilot projects before expanding to a wider coverage. Examples include Pratham’s remedial education program and Seva Mandir’s teacher camera monitoring program in India, and the girls’ scholarship programs in Kenya.\(^7\) Evaluation of these randomized programs yields valuable insights into how public education can be promoted effectively and efficiently in developing countries.

Selected studies under each of the six broad categories of education interventions in developing countries are reviewed below. A critical review of the literature can be found in Holla and Kremer (2009) and Banerjee and Duflo (2006).

1. **Conditional cash transfer programs**

PROGRESA, launched by the Government of Mexico in 1997, is among the first CCT programs in developing countries. Its purpose is to improve the educational, health, and nutritional status of poor families, particularly children and their mothers. The education component consisted of cash grants to mothers conditional on their children’s school enrollment and a monthly and annual minimum attendance rate of 85\%.\(^8\) The program covered eligible communities in a randomly phased-in fashion. Encouraged by the success of PROGRESA, several other Latin American and Asian countries have also designed and implemented CCT programs in recent years. Examples include Colombia’s Familias en Acción, Nicaragua’s Red de Protección Social (RPS), and Indonesia’s Jaring Pengamanan Sosial. Given the disadvantage of girls against boys in many developing countries, a number of CCT or similar programs have particularly targeted girls’ access to education, such as the female stipend program in Pakistan, a scholarship program in Cambodia, and the Girls Scholarship Program in western Kenya.

PROGRESA, one of the most heavily studied CCT programs, has been found promising for increasing school enrollment and attendance. A large number of studies utilize experimental data generated from the program to evaluate its impact. Schultz (2004) shows that the program raises enrollment rates of poor children in the treated localities compared to the control localities, with the positive impact often larger for girls than boys, and larger at secondary than at primary schools. Behrman, Sengupta, and Todd (2005) find that the program reduced dropout and repetition rates, and raised the probability of advancing to the next higher grade as well as the reentry rate from the dropout state. Attanasio, Meghir, and Santiago (2005) and De Janvry, Dubois, and Sadoulet (2007) reveal similar positive findings on school enrollment and grade progression, although the magnitudes of the effects differ somewhat. Lalive and Cattaneo (2009) and Bobonis and Finan (2009), meanwhile, present evidence showing that children in ineligible households also acquired more schooling when the subsidy was introduced in their local village or neighborhood. They

---

\(^7\) Pratham and Seva Mandir are nongovernment organizations in India.

\(^8\) The health and nutrition component of PROGRESA is discussed in the health section.
argue that this is because the program indirectly influenced the decisions of ineligible households by causing changes in peer group schooling; the social interaction effect is almost as important as the direct effect. De Janvry et al. (2006) identify an additional benefit from PROGRESA in that the conditional transfers act as safety nets for schooling of poor children in the presence of short-run economic shocks.

Evaluations of other CCT programs produce similar favorable results: CCT programs generally achieved the intended outcomes, including higher school enrollment, attendance, and grade progression, and lower dropout rates and child labor force participation (Attanasio et al. 2005, Attanasio et al. 2006, Maluccio 2009). Indonesia’s Jaring Pengamanan Sosial program was found effective in keeping large numbers of children from dropping out of school in the aftermath of the 1997/98 Asian financial crisis (Cameron 2002, Sparrow 2006). Schady and Araujo (2006) find that the randomized unconditional cash transfer program, Bono de Desarrollo Humano, in Ecuador effectively increased enrollment, reduced dropout rates, deterred children from working, or shortened their work hours.

Evaluation studies of CCT programs targeting girls in particular consistently showed significant positive effects on girls’ school enrollment (Chaudhury and Parajuli 2006; Kim, Alderman, and Orazem 1999; Filmer and Schady 2006). Based on a randomized evaluation, Kremer, Miguel, and Thornton (2009) find that not only did the merit scholarship program in Kenya lead to substantial exam score gains among girls in program schools, but girls who were unlikely to win a scholarship and boys in one of the two program districts also gained.

2. In-kind transfer programs

Bangladesh’s Food for Education program in the mid-1990s provided a free monthly ration of rice or wheat to poor families if their children attended primary school. Meanwhile, many unconditional in-kind transfer programs aim at lowering the costs associated with school attendance to keep children in school. Also in Bangladesh, the government and the World Food Programme jointly launched the School Feeding Program in food-insecure areas of Bangladesh in 2002. All children in the treatment schools received nutrient-fortified biscuits. In western Kenya, the International Child Support (ICS), a Dutch nongovernment organization (NGO), implemented a randomized subsidized school meals program during 2000–2002. The Philippines’ Dropout Intervention Program (DIP) also consisted of a school feeding intervention. Two randomized interventions in western Kenya distributed uniforms, which are often a significant cost for poor families, to pupils in primary schools for free.

The effects of in-kind transfer programs are found to be largely positive. Evaluating the Food for Education program of Bangladesh using econometric methods, Ravallion and Wodon (2000) and Ahmed and Del Ninno (2002) show that the program promoted school enrollment and attendance and reduced grade repetition and dropout. Ahmed (2004) finds similar effects of the School Feeding Program in Bangladesh. The free school meals experiment in Kenya had a significant positive effect on school participation and resulted in higher curriculum test scores in schools with experienced teachers. But because of capacity constraints and pupil transfers, the treatment schools raised their fees while nearby comparison schools lowered theirs (Vermeersch and Kremer 2004). The DIP

9 The other three interventions in DIP were multilevel learning materials, school feeding combined with parent–teacher partnerships, and multilevel learning materials combined with parent–teacher partnerships.
school feeding intervention in the Philippines, whether combined with parent–teacher partnerships or not, had a positive and significant impact on math and English test scores (Tan, Lane, and Lassibile 1999). Evans, Kremer, and Ngatia (2008) and Duflo et al. (2006) provide evidence from randomized experiments that uniforms were a financial barrier to school attendance and that the positive effects of free uniforms on school attendance were larger for poorer students who did not own a uniform before the program. An unexpected impact identified in Duflo et al. (2006) is that teenage childbearing declined in response to the uniform program.

3. School construction and physical inputs to education

During a major program in 1973–1974 and 1978–1979, the Government of Indonesia built over 60,000 new schools, with the number of schools in each district proportional to the number of children of primary school age not enrolled in 1972. Argentina implemented a large program during 1993–1999 that financed the construction of 3,531 classrooms for preschool education throughout the country. Priority of program placement was given to poor areas with low levels of preschool enrollment. Two programs in Pakistan established private girls’ schools to increase girls’ enrollment. To address the concern that shortage of school inputs results in poor performance of students in less developed areas, ICS administered two randomized interventions to supply physical inputs, flip charts, and textbooks in schools in rural Kenya. Pratham implemented a computer-assisted learning program in some urban schools in India. Students participating in the program got a chance to play math education games 2 hours per week.

The literature shows that large-scale school construction programs contributed to local human capital accumulation. Duflo (2001) adopted a DID strategy showing that Indonesia’s school construction program significantly increased the years of education for children in sparsely populated regions. As a consequence, hourly wages increased 10.6% for beneficiary cohorts and regions, while the effects were larger for regions where population density was low and the average education of cohorts not exposed to the program was low. Berlinski, Galiani, and Gertler (2009) find that the preschool facility construction program in Argentina increased preschool attendance by an average of 7.5 percentage points. Furthermore, attending preschool had positive effects on subsequent third-grade Spanish and math test scores as well as behavioral skills such as attention, effort, class participation, and discipline. Kim, Alderman, and Orazem (1998) show that construction of private girls’ schools in Pakistan increased girls’ enrollment.

The impacts of school inputs programs may depend on the actual inputs supplied. The evidence comes from two randomized evaluations. Glewwe et al. (2004) and Glewwe, Kremer, and Moulin (2009) show that using flip charts and supplying textbooks did not result in significant changes in test scores. While the best students with high pretest scores benefited from the textbooks, other students did not. This may be because the textbooks were written in English, most students’ third language. Banerjee et al. (2007) find that Pratham’s computer-assisted learning program increased math scores significantly and the benefits remained substantial though smaller than the initial year after the program.

4. Private participation

Private participation in the education sector is often considered a quick and efficient way to increase public education supply. In 1981, Chile started a nationwide initiative that provided vouchers to every
student attending private school to cover part of their costs. With strong financial incentives created by the program, more than 1,000 private schools entered the market. The Government of Colombia’s PACES, established in late 1991, is another well-known voucher program. PACES randomly offered vouchers, which cover about half of the costs of private secondary education, to a subset of applicants from households in the lowest socioeconomic strata. In 1995, PACES provided vouchers to about 100,000 students to attend private school.

However, findings about the voucher programs to encourage private participation in public education supply seem to be mixed. Hsieh and Urquiola (2006) find no evidence that Chile’s voucher program, which removed restrictions on choice of school, had improved average education outcomes as measured by test scores, repetition rates, and years of schooling. The evidence suggests that the voucher program gave rise to sorting, as the “best” public school students left for private schools. The randomized evaluations, however, produce some positive results. Angrist et al. (2002) find that the voucher winners in the PACES program were more likely to receive a private school scholarship, attend a private school, complete more years of schooling, and had lower repetition rates than losers. Angrist, Bettinger, and Kremer (2006) also find that the vouchers raised registration rates for college entrance examination, and that the voucher winners scored higher on the language section of the exam.

5. Improving teacher quantity and quality

Several programs have been designed to address the teacher shortage issue. The Government of India launched the Operation Blackboard program in 1987 that provided a second teacher to all one-teacher primary schools, and a teaching-learning equipment packet to all primary schools. Somewhat similar to Operation Blackboard, ICS Africa implemented the Extra Teachers Program in Kenya in 2005, but on a much smaller scale. Participating schools received funding to hire an additional local first-grade teacher with the same academic qualifications as a civil service teacher, and split the class into two. The program experimented on combinations of class size reduction with alternative reforms, including teacher incentive improvement, and grouping students by initial achievement. Pratham conducted a remedial education program (Balsakhi) in some Indian urban schools, which hired young women to teach lagging students.

To motivate teachers for better teaching, ICS administered an incentive program in primary schools in western Kenya. Half of the participating schools won prizes based on the performance of the school on the district exams in each year. Each teacher in the winning school received a prize equivalent in value to 21%–43% of an average teacher’s monthly salary. In a larger experiment run by an Indian state, teachers of rural primary schools were given bonuses based on the average improvement of their students’ test scores in independent learning assessments. To address the problem of rising teacher absenteeism, Seva Mandir implemented an innovative program in 2003–2005 to enforce the monitoring of teacher attendance in India. The main element of the program was to photograph, with date and time function, the teacher and students at the start and end of each day. Teachers were paid according to the number of days in which photos satisfied the specified rules.

Supplying additional teachers may not be sufficient for better school performance. Chin (2005) reveals that the Operation Blackboard program did not send enough teachers to the intended places, teachers per school did not increase, and class size did not decrease. The effect of the program on school inputs seemed to have been the redistribution of teachers from larger schools to smaller
ones. Nevertheless, the program raised primary school completion rates, especially for girls and the poor. Evaluation of Kenya’s randomized Extra Teachers Program by Duflo, Dupas, and Kremer (2008) suggests that class size reduction alone did not improve student test scores. Rather, combining class size reduction with measures that improved teacher incentives, such as hiring local teachers on short-term contracts, or empowering parent committees to oversee the program, led to a significant increase in test scores. In addition, class size reduction combined with student grouping based on initial achievement also led to test score increases—an evidence of the benefits of homogeneous classes. Banerjee et al. (2007) find that Pratham’s remedial education program led to large gains in the test scores of children at the bottom of the class.

While ensuring an adequate number of teachers is important, providing incentives and monitoring teacher efforts may be more relevant. Glewwe, Nauman, and Kremer (2003) find that western Kenya’s teacher incentive program, which was implemented as a randomized experiment, led to significantly higher test scores among students in program schools. But the program was unsuccessful in stimulating more teacher effort aimed at increasing long-run learning. Teacher attendance did not improve, homework assignment did not increase, and pedagogy did not change. In addition, more test preparation sessions were conducted to raise short-run test scores, but students were not able to retain the gains in test scores after the end of the program, confirming concerns about teacher incentive programs. Nevertheless, Muralidharan and Sundararaman (2009) produce somewhat different evidence in evaluating the 2-year teacher incentive program in India. They show that the program improved students’ performance on both “conceptual” as well as “mechanical” components of the tests, suggesting that the gains in test scores were an actual increase in learning outcomes. Positive spillover was also identified on subjects without incentives. Teacher incentive intervention was more effective than school inputs intervention that had similar costs. Finally, Duflo, Hanna, and Ryan (2008) show that Seva Mandir’s class camera experiment in India had significant positive effects on teacher attendance, raising the mean teacher attendance rate from 58% before the program to 79%—mainly due to its financial incentives—and consequently improved children’s test scores.

6. School management reform

During 1978–1994, Argentina decentralized all federally managed public schools to provincially controlled public schools. In 1991, El Salvador’s Ministry of Education started Educacion con Participacion de la Comunidad (EDUCO), a program decentralizing preschool and primary education by strengthening direct involvement and participation of parents and community groups. Similarly, in 1999 Honduras implemented a community-based education project, El Proyecto Hondureno de Educacion Comunitaria (PROHECO). The project decentralized aspects of school management to the school level, emphasizing parent and community participation in the administration process. In 1991, Nicaragua established councils in all public schools to ensure the participation of the educational community, parents, and students in making school decisions.

Evidence favoring school decentralization reforms is somewhat weak. As most interventions in this area are nonrandom, evaluation is mainly based on econometric methods. Evaluation of Argentina’s school decentralization indicates that overall, it had a positive impact on student test scores. But the gains were not distributed equally and the poor may not benefit from such reform (Galiani, Gertler, and Schargrodsky 2008). Jimenez and Sawada (1999) show that students in EDUCO schools missed less school days due to teacher absences than students in traditional schools. In a later study, Jimenez and Sawada (2003) find that the impacts of the program on the likelihood of school continuation were positive but sensitive to model specifications. Sawada and Ragatz (2005)
report that the EDUCO program increased teachers’ time spent teaching and meeting with parents. Yet, they could not find significant results to firmly assert the impacts of the decentralization program on student outcomes. Di Gropello and Marshall (2005) find evidence consistent with the idea that decentralization under PROHECO streamlined administrative processes in their baseline model. When more controls were added to the model, however, most results for teacher behaviors were not statistically significant. The authors attribute this result to the small sample size. King, Ozler, and Rawlings (1999) demonstrate an increase in the autonomy of schools and the perceived influence of council members after Nicaragua’s school autonomy reform, but failed to assess how the reform affected teacher and student performance.

B. Health

Three out of the eight Millennium Development Goals (MDGs) are related to human health improvement: reducing child mortality; improving maternal health; and combating HIV/AIDS, malaria, and other diseases. Developing countries and development agencies have funded a large variety of public health programs to address problems that hinder the achievement of these goals, as well as other prevalent issues such as limited access of the poor to basic health care. However, the effectiveness of these programs are not always self-evident because many factors—such as genes, motivation, habits, standard of living, and family circumstances—work interactively to affect individuals’ health outcomes even though the programs may change only one or a few of these factors. Given public budget constraints, it is important to evaluate the actual impacts of the health interventions and assess their cost effectiveness.

A few categories of interventions, including (i) CCT health program, (ii) subsidized insurance program, (iii) school or community-based health intervention, and (iv) HIV/AIDS-targeted interventions, have been carefully examined in the literature. In addition to various health interventions, an important issue that has increasingly been debated and assessed is whether or how to charge for public health services and products. Advocates argue that charging fees could screen out low-valuation consumers, induce more frequent use of products or services, and make the program more financially sustainable. On the contrary, opponents raise the concern that the take-up of health services is very low in some cases even if services are available for free. How charging fees would affect the utilization of basic health services and how to promote utilization in case it is low are two interrelated questions attracting more and more attention.

To the extent that health status is closely correlated with individuals’ genetic factors, motivation, attitudes, habits, and daily behaviors, it should not be assumed that the interventions will naturally work to improve the health outcomes of the beneficiaries, particularly if they did little to change health-related behaviors. Moreover, because factors such as motivation, attitudes, and habits, which are not observable to the evaluator, have great influence on one’s health as well as program participation, it is quite a challenge to obtain reliable estimates of the impacts of health interventions when they are not randomly assigned. The single difference using propensity score matching and DID are common nonexperimental methods used in health program evaluation. The results of studies using these methods should always be interpreted with caution since they normally do not address the selection bias properly.

Similar to education program evaluation, a recent trend is to conduct randomized field experiments to better understand whether and how a particular health intervention works. Kenya’s school-based deworming experiment is one of the most famous examples adopting the experimental approach. Free of selection biases, the evidence derived from randomized field experiments may
be considered more convincing. The experience of the deworming project shows that once an intervention is proved effective and efficient through a controlled experiment, it is likely to be scaled up and applied in other regions.

1. **CCT health programs**

The CCT program is as popular in public health as it is in education. Mexico’s innovative PROGRESA program provided cash transfers to eligible families—conditional on availment of preventive health services and attendance of children aged 0–5 and lactating mothers in nutrition monitoring clinics, to receive nutritional supplements and education on nutrition and hygiene and to measure the growth of the children. Similarly, pregnant women were to obtain prenatal care, nutritional supplements, and health education. Modeled after PROGRESA, both the Government of Colombia’s Familia en Accion program and Nicaragua’s Red de Proteccion Social (RPS) program had nutrition and health components that gave eligible households cash contingent on attendance at educational workshops and on bringing children under 5 for growth and development checks, vaccination, and provision of nutritional supplements.

PROGRESA and other CCT programs are found generally effective in promoting health care and improving health outcomes. Gertler and Boyce (2003) show that the use of public health clinics increased faster in PROGRESA villages than in control villages. There was also an increase in nutrition monitoring visits. The use of public hospitals probably fell because the program improved health and lowered the incidence of severe illness. Gertler (2004) finds that among children in the treatment families born during the 2-year intervention of PROGRESA or aged 0–35 months before the program, illness was significantly lower compared to control families. Program-assisted children were less likely to be anemic and grew about 1 centimeter more during the first year of the program. Rivera et al. (2004) and Behrman and Hoddinott (2005) confirm that the program was causally associated with greater height, lower rates of anemia, and higher hemoglobin values among infants. While access to health education and services might have led to the positive impacts on health, the increased income due to the cash grant could have added to the benefits (Ruiz-Arranz et al. 2002). PROGRESA households increased their total expenditure on food as well as the variety of foods consumed.

Moreover, Colombia’s Familia en Accion program increased newborns’ weights, children’s intake of protein and vegetables, the probability of vaccination, and reduced caretaker-reported diarrhea (Attanasio et al. 2005). Maluccio and Flores (2005) indicate that Nicaragua’s RPS program increased annual per capita expenditure on goods and the use of preventive health care services for children, accompanied by an improvement in the nutritional status of beneficiary children under 5. However, the distribution of iron supplements was unable to improve hemoglobin levels or lower the rates of anemia. Finally, Schady and Paxson (2007) find that Ecuador’s Bono de Desarrollo Human (BDH) program, which was implemented as an unconditional cash transfer program, had positive effects on children’s and mothers’ hemoglobin levels; children’s physical, cognitive, and behavioral test scores; and the probability of receiving parasite treatments.

2. **Subsidized insurance programs**

Financing is often a major constraint preventing poor families from using health care services and from purchasing health insurance. A number of governments set up programs offering free or
subsidized health insurance to the poor, to improve the poor’s access to health care and reduce the financial burden due to health problems. Several such programs have been assessed thoroughly in the literature.

Accumulated evidence shows that free or subsidized health insurance encourages use of health services. Viet Nam introduced a formal social insurance program in 1992 consisting of three schemes—Compulsory Health Insurance, Voluntary Health Insurance, and Health Insurance for the Poor—covering various groups of citizens. Sepehri, Sarma, and Simpson (2006) find that government-provided health insurance reduced out-of-pocket expenditure and the reduction was more pronounced among lower-income individuals. Sepehri, Simpson, and Sarma (2006) also show that the compulsory insurance scheme and the insurance scheme for the poor increased length of stay in the hospital, while the voluntary insurance scheme had minimal effect on length of stay. The compulsory scheme also increased the likelihood of hospital admission far more than the other two schemes. Similarly, Hou and Chao (2008) examine Georgia’s Medical Assistance Program, launched in June 2006 to provide health insurance to the poor, and find that the initial impacts of the program include increased use of acute surgery and inpatient services by the poor. In evaluating a subsidized health insurance program for Colombia’s low-income families, Trujillo, Portillo, and Vernon (2005) find that the program greatly increased medical care use among the country’s poor and uninsured.

The People’s Republic of China piloted in 2003 a new cooperative medical scheme called Rural Mutual Health Care or Rural Cooperative Medical Scheme. The scheme adopted insurance coverage of outpatient services from the demand side and drug policy from the supply side. A few studies looked at the impact of the scheme on the health service use, financial burden, and health status of participants. Wagstaff et al. (2009) find that the program increased outpatient and inpatient use and reduced service costs, but it did not reduce out-of-pocket expenses per outpatient visit or inpatient spell. Zhou et al. (2009) indicate that the demand for outpatient visits rose with the program and per-visit expenses decreased with the drug policy, a result seemingly contrary to that in Wagstaff et al. (2009). Wang et al. (2009) show that the Rural Mutual Health Care had a positive effect on the health status of beneficiaries and improved mobility and usual activity for those over 55.

However, there is concern that the public health insurance programs may benefit the nonpoor disproportionately. A common issue raised by a few evaluation studies is how to make public health insurance programs more pro-poor. Indonesia initiated its Social Safety Net health card program in late 1998 to protect the poor from the economic crisis. Health cards that subsidized public health care were allocated to the most vulnerable households or those severely affected by the crisis. Somanathan (2008) finds that the program resulted in less declines in the use of public sector outpatient services for children during the crisis years. Pandey et al. (2007) indicate that the program led to a net increase in the use of outpatient care for the poor and a substitution from private to public health care for the nonpoor. The benefits of the corresponding increase in supply of public services have been captured mainly by the nonpoor. Yip and Berman (2001) studied Egypt’s School Health Insurance Programme, a government subsidized health insurance system that targets school children. While the program significantly improved participants’ access to health care, as measured by visit rates and lower out-of-pocket expenditures, the differences in access between children enrolled and not enrolled in schools were enlarged. Unfortunately, those children not enrolled tended to be poor and live in rural areas.
3. **School or community-based health interventions**

Health interventions are increasingly implemented in schools or communities, in the hope of generating economies of scale and maximizing positive spillover. Evaluations of many such programs show that the interventions generally had positive effects on health and sometimes nonhealth outcomes. For example, in 1998 the Ministry of Health in Kenya carried out a mass deworming program in 75 rural primary schools, in which 90% of enrolled pupils suffered from intestinal worm infections. The interventions, randomly phased into schools, include health education on worm prevention and provision of deworming medicine. Miguel and Kremer (2004) find that the program reduced worm infections, sickness, and anemia by one quarter, which was far more cost-effective than many alternative ways of boosting school participation, although no positive effect on test scores was detected. The project had positive externalities within and across schools and reduced school absenteeism by 25%.

Anemia is among the most widespread health problems for children in developing countries. The Government of India introduced a randomized health intervention delivering iron supplements and deworming drugs to preschool children. Bobonis, Miguel, and Puri-Sharma’s (2006) evaluation of the impact of the intervention reveals that weights increased among assisted children and school absenteeism declined by one fifth. The results contribute to a growing view that school-based health programs are effective in promoting school attendance in less developed countries.

Colombia’s Hogares Communitarios de Bienestar Familiar is a large community-based program providing food and care to poor children. The program encouraged the formation of parents associations in poor neighborhoods, which then elected community mothers to take care of children aged 0–5 within the community and feed them three times a day. Ruel et al. (2003) find large positive impacts on children’s dietary intake. Indeed, Attanasio and Vera-Hernandez (2004) further indicate that the program not only had positive impacts on children’s health (such as height) and school participation and performance, but also freed mothers in the program communities to participate in the labor market.

Uganda launched a project to enhance communities’ capacity to provide public services for children, such as growth monitoring, food security projects or early childcare and education facilities, and integrated health services. A field experiment on community public health management was also carried out. Through two rounds of meetings, local NGOs encouraged communities to get more involved in providing public primary health care and strengthened the communities’ capacity to hold local health providers accountable for performance. Alderman, Siddiqi, and Britto (2007) find that the project decreased the probability of early cessation of breast-feeding, increased nutrition intake and diversity, and encouraged the use of deworming medicines as well as several other benefits on children’s behaviors. Bjorkman and Svensson (2009) find that health workers in the community-based public health care monitoring program in Uganda appeared to boost efforts to serve the community, increased use of health services substantially, and improved health outcomes as measured by child mortality and child weight.

In contrast, Das Gupta et al. (2005) present a less optimistic picture about community-based health intervention in examining the Integrated Child Development Services program launched in India in 1975. The program established an Integrated Child Development Services center in each of the selected villages to provide nutrition and health services to children under 6 and pregnant and lactating mothers. Das Gupta et al. (2005) did not find positive effects of the program on children’s height and weight; indeed, in certain regions, the program even reduced children’s weight.
4. **HIV/AIDS-targeted interventions**

Public interventions targeted at HIV/AIDS try to seek ways to cause behavioral change among potentially vulnerable populations. In this regard, a key role is played by education and information. In Kenya, an education/information campaign on HIV/AIDS was implemented in rural areas. Four interventions were randomized across schools, including training teachers in a government-designed HIV/AIDS education curriculum, asking students to debate the role of condoms and write essays about protection against the disease, providing free uniforms, and informing teenagers about the variation in HIV rates by age and gender.

Duflo et al. (2006) find that training teachers in the HIV/AIDS curriculum had little impact on students’ relevant knowledge, attitudes, and behaviors, except that girls were more likely to be married in the event of a pregnancy. The condom debates and essays increased self-reported use of condoms without increasing self-reported sexual activity. Dupas (2009) further reveals that the most effective intervention was information on HIV risks. Learning HIV risks led to a 28% decrease in teen pregnancy and substitution away from older (riskier) partners and toward protected sex with same-age partners. In contrast, the national abstinence-only HIV education curriculum had no impact on teen pregnancy.

In some areas with high rates of HIV infection, individuals have no intention to know about their HIV status. This makes efforts to curb further transmission very difficult. A field experiment in rural Malawi tried to use monetary incentives to induce people to learn their HIV test results. In a study of the impact of the intervention on condom purchases, Thornton (2008) finds that a very small incentive doubled, to about 70%, the percentage of HIV-tested individuals wanting to learn their test results. Sexually active HIV-positive individuals who learned their results were three times more likely to purchase condoms than sexually active HIV-positive individuals who did not learn their results. However, the sexually active purchased only two additional condoms. There was no significant effect of learning HIV status on the purchase of condoms for the HIV-negative individuals.

5. **Pricing and take-up of health services**

Results from several randomized experiments do not favor the idea of charging fees for some basic health products and services. Kremer and Miguel (2007) show that after the introduction of cost-sharing in the school deworming project, the take-up rate dropped from 75% in the free-treatment schools to 19% in the cost-sharing schools. Students with helminth infections did not appear any more likely to pay for the drugs, suggesting that the charging scheme did not help target the program to those who most needed it. Because charging raised administrative costs significantly, the revenues raised would only expand program coverage by 5% given a fixed budget. Two programs distributed insecticide-treated nets in Kenya and Uganda.

In the case of Kenya, Cohen and Dupas (2010) report that take-up rate declined with the price charged at an increasing rate and decreased by 60.6 percentage points at the highest nets price, which still represented a 90% subsidy. There was no evidence that prices screened out people in less need of the nets or that those paying a higher price were more likely to use the nets.

In the case of Uganda, Hoffman (2009) indicates that charging for a net increased the likelihood that it would be used by the main income earner in the household in Uganda rather than the most vulnerable members, such as young children and pregnant women. In a door-to-door
marketing campaign in Zambia offering bottles of water disinfectant, Ashraf, Berry, and Shapiro (2007) find that higher prices deterred the purchase of the water disinfectant, did not help target the disinfectant to households that could benefit from it the most, and did not induce more intensive use of the disinfectant. Holla and Kremer (2009) suggest that free distribution of water disinfectant or mosquito nets may not lead people to accumulate these products in large stocks; rather it encourages people to sell them in secondary markets or use them in alternative and inefficient ways. The evidence, however, was not sufficient to judge the likelihood of this concern.

In addition to providing affordable health services accessible for the poor, interventions need to involve incentives and other approaches to ensure high adoption. To explain the phenomenon that the poor do not exploit the benefits of some freely provided preventive measures, such as vaccines, Banerjee et al. (2010) carried out a series of field experiments in rural India, where the rate of fully immunized children under 2 was only 1% despite the availability of free vaccines at local health centers. The experiments found that people positively responded to the reduction of travel costs (transaction cost) to health centers by establishing vaccine dissemination camps at their premises. However, a much larger increase in the rate of immunized children was observed when a small incentive was given upon each immunization received by a child under 2. What this result tells us is that people who did not respond to a price reduction funded by a full subsidy that makes the cost of immunization close to zero, may strongly respond to a small and immediate incentive. This could be due to the lack of information about the long-term benefits of immunization or people’s strong preference for the present gains.

C. Infrastructure

A large proportion of development assistance has been allocated to infrastructure investment, mainly transport, energy, and water supply. Accordingly, the number of impact evaluation studies in these subsectors has increased in recent years.

1. Transport

A large share of developmental interventions focuses on building or rehabilitating transport facilities, intercity and intracity highways and railways, rural roads, and waterways, to name a few. However, most impact evaluation studies focus on rural roads because they are viewed as one of the most important poverty alleviation instruments. It is argued that rural roads can raise living standards and improve the welfare of poor rural households by increasing access to goods and services, stimulating agricultural production and diversification, and creating off-farm employment. Moreover, the influence of rural roads is thought to be largely confined to well-defined zones (such as villages), which renders it suitable for impact evaluation. In contrast, trunk roads or railways connecting cities tend to have complex impacts on the whole region. The network or spillover effects make it very hard to identify qualified counterfactuals. Thus, alternative tools and analytical frameworks are needed to measure the impacts of these large transport projects (Chambers et al. 2009).

The major difficulty in rigorously evaluating rural road impact is that road placement is not randomly determined. Roads are built in areas that have attributes correlated with the level of the outcome variables or the changes in outcomes. Researchers may not observe some of these attributes and thus they cannot be controlled in the models. The issue remains the same even if the

\[10\] For a more comprehensive and detailed review of impact evaluation of rural road projects, see van de Walle (2008).
evaluation is focused on micro entities (households, firms, etc.) because there are likely to be some unobserved community characteristics that determine project placement as well as affect individual outcomes. Moreover, it may take a while to get a comprehensive picture of the effects of a rural road improvement. An evaluation study needs to follow up after a sufficiently long period (e.g., 5 years) to capture all welfare impacts of the project or to measure the long-term effects correctly.

The DID method is commonly used for estimating rural road effects, which involves subtracting the difference in the outcome of nonproject areas before and after the intervention from the difference of the project areas before and after the intervention. The results can be attributed to the project itself if road placement is not affected by unobserved, time-varying factors that also affect outcome changes. In a set of studies (e.g., Lokshin and Yemtsov 2005, Mu and van de Walle 2007), matching methods—more specifically propensity score matching—are combined with DID to better construct the comparison group and relax functional restrictions in traditional DID, which sometimes causes specification bias. However, neither traditional DID nor combined DID/propensity score matching can handle the selection bias stemming from time-varying unobservables correlated with both project location and changes in outcome. In this case, the IV method is necessary to generate credible estimates of project impact. However, it is difficult and rare to find a qualified IV for rural road placement. If a large number of rural roads were under consideration and the rule was that those linking villages with 30% or higher poverty incidence were selected, regression discontinuity designs would offer an alternative method for impact evaluation.

Questions examined with regard to evaluating rural roads typically include how rural roads impact the availability and consumption of goods and services; how they affect input and output prices; whether they encourage market development, or increase nonagricultural income-earning opportunities; and how the effects differ across gender and between poor and rich households. The findings emerging from a number of recent studies, however, present a mixed picture.11

Escobal and Ponce (2004), using cross-sectional data, find that rural road rehabilitation enhanced nonagricultural income-earning opportunities in poor rural areas of Peru. The authors argue that higher incomes led to higher savings in the form of livestock, since road rehabilitation did not give rise to higher consumption.

Akee (2006) finds that a new road built in the Republic of Palau increased wage sector employment at the expense of self-employment in agriculture, discouraged migration overseas, and raised automobile ownership in formerly inaccessible rural areas. The findings also show that inequality declined both within and between regions. However, the impact of road construction on average household wages and income was negligible.

Among studies that take impact heterogeneity into account, Khandker, Barnes, and Samad (2009) find significant positive impacts of road investment, including raising agricultural production, wages, output prices, and schooling outcomes; and reducing input and transport costs in Bangladesh. They also find that the benefits of road investment leaned more toward the poor than the nonpoor. Lokshin and Yemtsov (2005) find that rural road rehabilitation in Georgia increased off-farm employment for nonpoor households and female wage employment for poor women only. The study finds no impact of the project on agricultural product sales.

---

11 A large body of literature examines the causal relationship between general transport infrastructure—as opposed to a particular transport project—and economic outcomes (e.g., Gibson and Rozelle 2003; Banerjee, Duflo, and Qian 2009). They can be classified into impact evaluation by a broad definition, though they are not included in this review.
Mu and van de Walle (2007) find, at the community level (communes), that rehabilitation and new construction projects increased the presence and frequency of markets; the availability of various goods and services such as fertilizer, gasoline, and hairdressing; employment for unskilled labor; and primary school completion rates. On the other hand, they do not find evidence of impact on many other outcomes such as land market development and transport services. The study documents the tendency for greater impact in poorer communities, probably due to lower initial conditions.

Urban transport infrastructure investment, such as in major intracity arteries, metro rail systems, or bus rapid transit systems, are thought to affect labor market outcomes of city residents, the location of firms seeking economic opportunity, access to services, and land and house prices (Boarnet 2007). A few studies examine their impact in developed countries (e.g., Holzer, Quigley, and Raphael 2003; Chalermpong 2004; Bowes and Ihlanfeldt 2001). But there is an absence of similar studies in developing countries, suggesting a potentially fruitful area for researchers.

2. Energy

Developmental energy projects range from the construction or rehabilitation of power generation, transmission, and distribution facilities, to rural electrification with grid or off-grid power supply, renewable energy development, and energy efficiency improvement. However, there have been very few impact evaluations of energy projects relative to the large number of projects implemented. Existing impact evaluations in energy mainly focus on rural electrification, probably because of its focus on reducing rural poverty. Some interest falls on the effects of household fuel use on indoor air quality and health outcomes.

Rural electrification is believed to improve households’ all-around welfare through multiple channels. Cheaper and better lighting allows longer study time and social activities, while replacing lighting fuels such as kerosene improves indoor air quality. Electrical appliances such as electric fans, washing machines, and television sets raise living conditions, save domestic labor, and promote learning and knowledge. Access to electricity can significantly improve public services in health care, education, and public security. And finally, access to power is expected to enhance rural agricultural and nonagricultural productivity as well as encourage small business development, leading to higher employment and income. These direct and indirect impacts may be measured in terms of household or individual income, consumption, employment, health status, education performance, etc.

Similar to transport projects, it is very difficult, if not impossible, to randomize program participation in rural electrification projects. One must rely on nonexperimental methods to obtain reliable estimates of impacts on the electrification beneficiaries. The same issue that troubles impact evaluation of other types of projects arises again. In some respects, communities with electricity are likely to be different from those without, in ways which are unobserved to the researcher and correlated with economic performance of the community or individual households. For instance, those located closer to generation or transmission facilities get connected first, and may be relatively wealthier than those farther away. Without a good measure of the difference in initial economic conditions, comparing the electrified against the unelectrified tends to produce biased estimates. While this type of selection caused by fixed cross-sectional differences can be dealt with by the DID method, correlation between project placement and the economic growth path presents selection bias that is harder to address.
The World Bank (2008) argues that selection bias is not an issue in the case of rural electrification in that the determinants of selection (income and geographical location) are observable and thus can be controlled in the model. But other researchers are more concerned. In studying the impacts of rural electrification in South Africa after apartheid, Dinkelman (2009) notes that infrastructure spending is often targeted at growth centers or at backward but politically important areas. Comparing electrified and unelectrified communities is thus unlikely to solve the problem of biased estimates. Instead, she combines DID with the IV approach and uses community land gradient as an instrument for the electrification status. The two approaches present different results.

Dinkelman (2009) presents hard evidence of the positive impact of rural electrification on domestic energy use and female employment.\textsuperscript{12} She finds that electrification raised electric lighting by 71 percentage points and cooking by 24 percentage points, and reduced wood use for cooking by 28 percentage points (the corresponding DID estimates are 23, 6, and 4.2 percentage points). When employment is examined separately for men and women, the IV estimates indicate a statistically significant 13.5 percentage point increase in female employment, while DID estimates show nearly zero effect. For male employment, IV yields a smaller (4.2 percentage points), statistically insignificant increase, while DID shows a marginally significant reduction of 1 percentage point. The study shows that electrification freed females to join the labor market and did not hurt males’ labor participation. It also finds larger employment effects for women with fewer child-care responsibilities and who initially rely more on wood use for cooking.

Barış and Ezzati (2007) studied a World Bank project to alleviate indoor air pollution in rural households through technological and behavioral interventions on energy use. A total of 5,500 households from poor rural areas in four provinces in the People’s Republic of China were divided into three groups—one receiving improved stove and ventilation systems plus health education and behavioral interventions, one receiving only health behavioral interventions, and one receiving nothing. Analysis is based on the baseline data and postproject data collected a year after the project was completed. The results indicate that heating stove improvement mitigated indoor air pollution measured as particulate matter, carbon oxide, and sulfur dioxide concentrations. The effects of cooking stove improvement are less consistent, which might be subject to user behavior variation. On the other hand, health education and behavioral interventions alone did not lead to significant indoor air quality improvement without alternative stoves. This may be because infrastructural and household economic conditions kept households from making actual behavioral changes such as switching fuels or taking up better technologies. Unfortunately, the study does not explore the effects of intervention with stove improvement only. It will be interesting to see whether health knowledge and behavioral change are important complements to the household energy intervention.\textsuperscript{13}

A low uptake of energy efficiency technology, even if it is financially viable, is often referred to in the economics literature as the energy efficiency paradox. This phenomenon, common to developed and developing economies, might be explained in a number of ways, including market failures such as asymmetric information, environmental externalities, liquidity constraints, learning-by-using spillover, as well as behavioral failures (Gillingham, Newell, and Palmer 2009). Accordingly, various policies and public programs have been tried to overcome the barriers to

\textsuperscript{12} Some studies on rural electrification (e.g., ESMAP 2003, Laxmi et al. 2003, Massé 2003, Parikh 2005) are reported in World Bank (2008). However, they do not satisfy the conditions required of a rigorous impact evaluation.

\textsuperscript{13} The health impacts of the project are not reported in the study.
energy efficiency adoption. These include, for instance, product labeling to provide greater and more reliable information about the energy efficiency of the products, industrial energy audits to help firms determine their energy saving potential, financial incentives such as tax credits to encourage energy efficiency investment, and supporting energy service companies to facilitate energy efficiency financing. However, the actual effectiveness and impacts of energy efficiency promotion interventions on improving energy efficiency and changing energy use behavior are largely unknown, especially in developing economies. Consequently, even as energy demand surges with rapid economic growth, there is little knowledge of the optimal strategies for promoting energy efficiency and conservation in developing countries. Research looking into the design of and assessing different interventions to close the gap in energy efficiency would be very useful.

3. Water supply and sanitation

Infrastructure aimed at improving household water supply, water quality, and sanitation accounts for a large proportion of developmental projects. It includes provision of water supply at the community level; connection of piped water; typically to urban households; improvement of latrine and drainage facilities; and protection of spring water sources. Because water services are seen as driving health impacts, as well as providing utility access, a number of noninfrastructure interventions for public health have also been conducted independently or jointly with water infrastructure projects. For example, some projects promote hand washing or drinking water treatment at point of use.

Health outcomes are the most important aspect of water supply and sanitation projects. In particular, project impacts on diarrhea incidence among children under 5 and child mortality have been widely examined. While many evaluations on water projects have been conducted since the 1980s, more rigorous studies emphasizing the causal effects of the projects have only emerged in recent years.

A larger variety of evaluation methods have been applied to the water supply sector than to the energy and transport sectors. Among others, Jalan and Ravallion (2003) use propensity score matching to assess whether access to piped water brings child-health benefits in India; Galiani, Gertler, and Schargrodsky (2005) combine propensity score matching with DID to measure the impact of privatized water supply on child mortality; Gamper-Rabindran, Khan, and Timmins (2008) use instrumental variables to estimate the impact of community piped water on diarrhea-caused mortality in Brazil; and Kremer et al. (2006) adopt a randomized experimental approach in evaluating the health effect of spring protection on diarrhea incidence.

A few studies find that access to piped water and sanitation (e.g., sewerage) infrastructure have reduced infant mortality (Gamper-Rabindran, Khan, and Timmins 2008, Newman et al. 2002) and the prevalence and duration of diarrhea among children under 5 (Jalan and Ravallion 2003; Kolahi, Rastegarpour, and Sohrabi 2009). In addition, how to deliver piped water and sanitation services is likely to be important in developing countries. Argentina privatized 30% of municipal water companies in the 1990s, leading to improvements in water production, water supply, sewage drainage volume, and water leakage, according to Galiani, Gertler, and Schargrodsky (2005). They estimate that child mortality fell 5%-7% in areas that allowed private firms to run piped-water service, and that the effect was largest in the poorest areas.

On the other hand, the evidence of the health effects of rural water infrastructure short of piped water is much weaker (e.g., Fewtrell et al. 2005, Esrey 1996). Furthermore, some studies
point to the failure of community-based infrastructure maintenance (e.g., Miguel and Gugerty 2005). The best maintenance scheme for rural water infrastructure remains an open question. Kremer et al. (2006) show that springwater protection in western Kenya is very effective in improving household water quality, but has little effect on diarrhea incidence and child weight and height. Some researchers believe that communal water infrastructure would probably work in combination with other interventions that improve sanitation or change hygiene behavior, such as point-of-use water treatment to improve the quality of drinking water and the promotion of hand washing. However, there is little evidence for such complementarity, despite the fact that point-of-use water treatment and hand washing have been established as an effective means of reducing diarrhea (Zwane and Kremer 2007). Moreover, the quality of such evidence found for water quality and hygiene interventions may suffer from bias, relatively short study periods, and small beneficiary samples (Waddington et al. 2009).

Although health concerns dominate the impact evaluations of water supply and sanitation projects, these projects were also expected to have positive effects on education and labor supply. The argument is twofold. First, improved health due to the intervention enables individuals to attend school or go to work more regularly. Second, the intervention would generate significant time savings in collecting and treating water that could be spent on productive activities by households. Ilahi and Grimard (2000) find a significant relationship between the proximity of water sources and market work for women in Pakistan. James (2003) shows that women used time saved through water supply improvements such as the provision of hand pumps in rural enterprises to generate additional household income. Khandker (1996) finds that drinking water from household tube wells increased school performance of boys but not girls. More studies are needed to estimate the direct causal link between water supply and sanitation projects and the labor supply of women, and/or to identify the mechanism through which household water supply affects child school performance and explain the gender differentials.

D. Microfinance

Despite rapid expansion of the microfinance industry as a means to reduce poverty over decades, the evidence showing that microfinance and microcredit in particular positively impacts the lives of the poor is still limited. Quality evidence backed by impact evaluations of microfinance operations until quite recently had been scarce due to the absence of an experimental approach.

Earlier assertions about the success of microcredit programs mostly relied on observational evidence comparing borrowers and nonborrowers. Observational evidence on microcredit programs, however, faces typical selection bias problems. Microcredit clients differ from nonclients in unobservable ways: microcredit borrowers may have special determination and entrepreneurial ability to improve their lives by starting new business activities, which nonborrowers may not have. If that is the case, the estimate of impacts by comparing borrowers and nonborrowers would overestimate the impact of credit. Similar bias also exists for the selection of the intervention area, for microcredit operations usually target areas with certain characteristics, such as the poorest villages or better-off villages.

A randomized field experiment ensures reasonably similar characteristics, observable and unobservable, between microcredit borrowers and nonborrowers, and thus addresses the selection bias issue. Randomization can be applied at the intervention placement stage. Randomly selected areas will be the treatment group, for which credit will be provided, and the rest become the
control group (Goldberg and Karlan 2008). Another possible design is to randomly select individual borrowers. Evaluation of randomized experiments at the individual level must account for potential spillover effects of the intervention.

Randomized field experiments applied to evaluating the field operations of microfinance have been rapidly gaining popularity and have started testing various aspects of conventional wisdom in microfinance. A fast-growing number of experimental studies in recent years have refuted the validity of some past evidence that was based on nonexperimental evaluation models. The results of such experiments clustered together are expected to have important policy implications.

1. Does microfinance reduce poverty?

Conventional wisdom that suggests microfinance reduces poverty has come into question. Three key impact evaluations, Pitt and Khandker (1998), Morduch (1999), and Khandker (2005), paved the way for a larger focus on microfinance among the development community, although there was some opposing evidence then. All are observational studies looking at microcredit in Bangladesh. Pitt and Khandker provide the statistical evidence for the claim that 5% of Grameen borrowers get out of poverty every year. Roodman and Morduch (2009) revisited the three most influential impact evaluations of microcredit. After re-running regressions and applying new statistical tests, the results failed to show that microcredit either increased household spending or reduced its volatility. Moreover, re-estimating Pitt and Khandker’s regression models even yield an opposite sign, which suggests that lending to women makes families poorer. The study argues that the credibility of existing academic evidence to support the notion that microcredit reduces poverty is very weak, and underscores the difficulty in inferring causality from nonexperimental data.

Rapidly emerging experimental impact evaluations provide further support to this finding. Together with Spandana, a fast-growing microfinance institution (MFI) in India, Banerjee et al. (2009) conducted a field experiment in urban slums in India to test the impact of microcredit on the welfare of the poor. Fifty-two neighborhoods were randomly selected from 104 neighborhoods for starting loans in 2005. More than 15 months later, a little less than 7,000 households covering both the treatment and control areas had been surveyed. The study compared all those surveyed in the treatment areas with all those surveyed in control areas, and found no impact on average household spending, the weight of women’s decisions in household spending, children’s illness, school enrollment, and school expenditures, although some positive business development for existing and prospective business owners was found.

A field experiment (Karlan and Zinman 2009) conducted in the Philippines produced similar results that put the fundamental rationale for microcredit in question. The study randomly selected clients of a microcredit lender in Manila from about 1,600 credit applicants, who were marginally above the lender’s creditworthiness criteria, and provided them with individual liability credits. The study tested the common assumption used for microcredit expansion that small business development is constrained by limited access to credit, thus credit expansion

---

14 Coleman (1999) evaluated the impact of credit in northern Thailand, and found that the impact estimate based on borrowers against nonborrowers would have overstated the gain from credit due to selection bias. The difference between the impact estimated by comparing randomly selected treatment and control groups and the one estimated by comparing borrowers and nonborrowers was sizable.

15 Counterarguments continue and the debate on the causality between microcredit and poverty is far from closed.

16 See next section for a detailed introduction of Banerjee et al. (2009).
should promote small business development and improve the lives of the poor. The study shows, however, that marginally creditworthy microentrepreneurs who received credit reduced the size and scope of their businesses compared to the control group. The borrowers did not invest in the businesses but in education. The results also show that microcredit seems to complement, not crowd out, other informal forms of insurance, such as emergency credit from families and friends. In sum, microcredit does not seem to have a direct impact on business expansion, but changed household investment patterns and risk-mitigation behaviors. This treatment effect was stronger among higher-income male groups, which are not typically targeted by microcredit interventions.

The above experimental studies measured credit impacts on specific segments of microcredit clients over a short period, therefore their results alone cannot directly be interpreted as the impacts of microcredit as a whole. Further replications of the evaluation studies in different country contexts, on different segments of microcredit clientele, and over different time frames are therefore essential in order to generalize the experimental results. Application of the experimental approach in evaluating the impacts of microfinance only started in recent years, and there are still missing pieces in the body of experimental research that are needed to fully understand the causality between microfinance and poverty reduction.

Dupas and Robinson (2009) were the first to evaluate the impacts of microsaving through field experiment. The study, done in rural Kenya, finds a positive impact from the interest-free saving facility on the lives of the rural poor: the formal savings products increased productive investment levels and the daily expenditure of relatively poor female entrepreneurs. Their experimental result supports the findings in the broader literature on the impacts of financial services on the poor, which is mostly nonexperimental, such as Burgess and Pande (2005).

2. Can microfinance clients afford high interest?

The effects of the high interest rates charged by some MFIs have been argued over for quite some time, including borrowers’ affordability and the need for MFIs to commercialize and become financially viable. Many argue that the poor cannot afford to pay high interest rates because their businesses do not generate high returns, meaning such credit would not improve their welfare. Karlan and Zinman (2010) in South Africa provide striking evidence to the contrary. The study randomly selected applicants to consumer loans who had not marginally met requirements and had been rejected, giving them a better chance of receiving a high interest loan. In 6–13 months after loan disbursements, the study found that the welfare of loan recipients improved: they had higher incomes, were more likely to have kept their jobs, and were less likely to experience hunger. But the study does not cover the potential impacts of high interest rates on the poorer segment of microfinance clients, who would be rated far below the creditworthy criteria.

Similarly, an experimental study by De Mel, McKenzie, and Woodruff (2008a) also demonstrated the ability of microentrepreneurs in Sri Lanka to pay high interest rates. The experiment sought to reveal how much clients were earning on the capital and whether high interest rates were affordable. Among randomly selected entrepreneurs given working capital, either in cash or in kind, the average return on additional capital was 5.7% per month, substantially higher than the market interest rate.

17 The consumer loans provided were for 4 months at 11.75% per month.
The same experiment was replicated in Mexico, with similar results. McKenzie and Woodruff (2008) randomly selected microentrepreneurs, gave them cash or capital in kind, and measured their returns to capital. The average was as high as 20% per month, well above the market rate. Returns were particularly high for credit-constrained business owners. The paper concludes that market failure prevented owners with viable business ideas from expanding their business to the optimal size.

3. Does microfinance empower women?

There is concern on the empowerment potential of microfinance operations targeted at women. It is not access to microcredit per se that empowers women, but giving control over assets to women that matter. Many microfinance operations are targeted at women on the belief that they are more creditworthy and can enhance their income by investing the capital in businesses. The findings of De Mel, McKenzie, and Woodruff and their follow-up study in Sri Lanka (2008a and 2008b), however, present striking evidence to show that the average return on capital was almost zero for female entrepreneurs, indicating that significant variance in affordability exists, depending on types of clients. Men generated higher returns to capital in general, while 59% of female-owned businesses generated negative returns, 14% generated 0%–5% per month, and only 27% earned more than 5% per month. The study found that women did not use smaller grants for business investment, and the larger grants invested in business had low returns. The results cast doubt on the effectiveness of microcredit investment that simply targets women as a way to develop small businesses and increase the income of poor households. In addition, the results in Banerjee, Duflo, and Qian (2009) suggest that microcredit does not enhance women’s role in intrahousehold decision making.

That said, an impact evaluation of commitment savings products in the Philippines reported a positive effect on women’s decision-making power within the household. Ashraf, Karlan, and Yin (2006) find that access to individually held commitment savings products led to a shift toward the purchase of female-oriented durable goods in the household, particularly for women with below median decision-making power in the baseline. The result indicates the significance of giving control over assets to women, not just targeted transfer of income, for enhancing women’s bargaining power.

4. How to ensure repayment?

A key to the success of the Grameen Bank is often thought to be its group-lending model, which is believed to ensure higher repayment rates by creating peer pressure within groups. Gine and Karlan (2006 and 2009) test this conventional wisdom by removing group liability for randomly selected microcredit clients in the Philippines, and observe no measurable drop in repayment, both over short and long periods. Gine and Karlan (2009) also find that the removal of the group liability requirement actually led to wider loan outreach.

Another key factor to successful microfinance is a weekly repayment meeting, which is believed to impose discipline and peer pressure in groups and keep each repayment amount relatively low and affordable, thereby maintaining high repayment rates. An experiment by Field and Pande (2008) finds evidence to challenge this belief. The paper studied an MFI in India that switched

---

18 A commitment savings product is an account with a prespecified goal for maturity, for which only the account holder, not other family members, has access and control.
the weekly repayment schedule to a monthly repayment for randomly selected borrowers, and found no change in the repayment rate.

Since many microcredit operators provide capital for microentrepreneurs, some types of business development and training assistance are often attached to credit. In their study of the impact of such business training provided by the Foundation for International Community Assistance (FINCA Peru), a microfinance organization, Karlan and Valdivia (2009) find evidence to show considerable improvement in client repayment and retention, and positive impact on the clients’ business income.

E. Agriculture

Rising agricultural productivity has significant implications for rural poverty in many developing countries. The World Bank (2007) reports that cereal yields have been rising steadily in the world since the 1960s. The use of fertilizer, together with irrigation and improved crop varieties, is considered a major driver of the rise. Although such productivity improvement has spread both in irrigated areas and rainfed arable lands, sub-Saharan Africa has not benefited. Even in countries where agricultural growth is strong, productivity in some regions remains low. Low-intensity input application (such as fertilizer), among other things, is suggested as a major cause of the underperformance of such areas. Some countries subsidize the use of fertilizer in order to enhance land productivity and agricultural sector growth. However, the take-up rate is still low in some areas. What works for enhancing the adoption of agricultural technologies among farmers is, therefore, of critical relevance to ongoing efforts to reduce rural poverty through improved agricultural production. Besides, there are studies examining the impacts of programs promoting export crops, construction of large dams, and land tenure on agricultural production and poverty.

Similar to evaluation of other interventions, evaluating agricultural interventions based on observational data suffers from the selection bias problem. For instance, to measure the diffusion of technology through social learning, simply observing farmers who adopted a new technology (e.g., hybrid seeds) and their neighbors’ rate of technology adoption suffers from serious biases due to unobservables. Although the correlation in adopting the technology among neighbors may be a reflection of imitation (social learning), it could also be due to common shocks all farmers face, which are however not observed by the evaluator.

A randomized social experiment that randomly provides the technology to farmers can shed light on this issue. The study finds that the degree of the omitted variable bias in an analysis using observational data is significantly large in this case (Duflo 2006). If a farmer’s adoption of fertilizer is regressed against his or her friends’ adoption rate without experimental control, the result indicates that a farmer who has one more contact that uses fertilizer is 10% more likely to use the fertilizer in a given season. However, the field experiment reveals no evidence on social learning among neighbors and friends in terms of using fertilizer.

In search of reasons to explain low fertilizer adoption among farmers, a series of field experiments has been conducted and evaluated to test several hypotheses on fertilizer adoption in western Kenya (Duflo, Kremer, and Robinson 2007, 2008, and 2009). The use of fertilizer among maize farmers was less than 15% in Kenya, despite its potential for producing much higher maize yields. The study first introduced fertilizer in randomly selected small farm plots in rural western Kenya to test actual productivity gains from using fertilizer. It is proved that the use of an appropriate
amount of fertilizer alone substantially increases returns with an average return of 36% over the season (69.5% per year) without changing other farming conditions.19

Information and measures to facilitate farmers’ ability to purchase fertilizer at the time of harvest (such as small incentives at harvest time) most effectively promote fertilizer use. When a technology for improving productivity is locally available and its effectiveness in substantially increasing return is proved by field tests in local farms, what other factors block farmers’ adoption of the technology, exploitation of its benefits, and improvement of their incomes? The answer may lie in a lack of information and difficulty in saving enough to meet input costs. Duflo, Kremer, and Robinson (2007) show that information certainly plays a role: farmers in demonstration plots were 10% more likely to use fertilizer for the next season than farmers in the control group. However, the story does not end here. An experiment was run in the same study area to provide randomly selected farmers with the option to purchase fertilizer with no subsidy but with free delivery right after the harvest, and the results were compared to those of alternative interventions such as fertilizer subsidies. Interestingly, it is found that farmers’ ability to save until the next cropping cycle is a key determinant of fertilizer usage; the no-subsidy option, with free delivery immediately after the harvest, increased fertilizer use by 17%, a higher impact than that of a 50% fertilizer subsidy offered later in the season.20 Subsequently, the study also finds that a small reduction in the fertilizer price immediately after the harvest, without free delivery, induced as much increase in fertilizer use as a 50% reduction in the fertilizer price with free delivery at the time when fertilizer top dressing needed to be done.

Besides technology adoption, some agricultural interventions aim to expand crop variety and provide marketing services. Ashraf, Karlan, and Gine (2008) studied DrumNet’s program to help Kenyan farmers switch to growing horticultural produce, which is exported to the European market. Two interventions were actually carried out randomly and evaluated: a full package of services ranging from training, inputs, credit, marketing, to transportation, and a package of all services except credit. The evaluation finds that the program was effective in changing farmers’ crop choice and increased incomes for those who grew export crops for the first time. The differences between the credit treatment group and the no-credit treatment group are, however, insignificant.

The common hypothesis that the overall impact of irrigation dams is positive and benefits will be redistributed to those harmed is not necessarily true. The need for institutional measures is highlighted to address distributional aspects of large dam construction. Duflo and Pande (2007) provide intriguing empirical evidence of the impacts of large dams on agricultural production and rural poverty, based on nonexperimental data. To address the endogenous placement problem in which the construction of dams is affected by regional wealth and the expected returns, river gradient variables were used as an instrument for presence of dams. The study finds that large dams in India have increased irrigated areas and agricultural production in downstream communities, while agricultural productivity gains in the vicinity of the dams were insignificant and agricultural production was made more vulnerable to rainfall shocks. Moreover, rural poverty levels increased significantly in the areas where dams were located, while in downstream areas poverty decreased slightly. In other words, no treatment to address adverse distributional impacts was in place for the dam investments.

Empirical evidence supports the theory suggesting that security of land tenure encourages investment for land fertility and thus increases agricultural productivity. In their study of the

---

19 Although application of fertilizer is a well-known existing technology in the study area, many farmers did not know how much fertilizer to use and the ones who used more than required tended to have low or negative returns.

20 No impact on fertilizer use was found for the free delivery option at the time of fertilizer top dressing.
causality between security of land tenure and land fertility and agricultural productivity in Ghana, Goldstein and Udry (2008) find that owners of more secure land tenure tend to invest more for land fertility and have substantially higher output: farmers with secure land tenure tend to allow sufficient time for fallowing. A large portion of agriculture in the area relies on shifting cultivation and fallowing is the most cost-effective investment for soil fertility, given the high price of fertilizer, the availability of abundant land, and low existing crop yields. The study also shows that those without secure land tenure, particularly women, tend to fallow their allocated plots less, yielding much less return. The result highlights the relevance of land title issue to agricultural development assistance, and provides important implications for relevant future investments, particularly in poor areas with difficult terrain.

F. Other Sectors

Besides the above five sectors, governance, job training, small and medium enterprise (SME) development, and gender are highly relevant to development policy and practice. Means to identify and curb corruption, deliver effective job training, promote SME growth, and enhance roles of women in community development are all critical topics in development research. While a few impact evaluations have been done, which we summarize below, a lot more empirical studies are in demand to provide solid evidence on the effectiveness of various interventions as well as insightful guidance on future moves in these fields.

1. Governance and corruption

Instead of understanding the causes of corruption and weak governance, some development economists seem to be more interested in looking for effective means for curbing corruption and strengthening governance. A recent well-known study by Olken (2007) compares road construction material actually used and material reported to be used under two monitoring schemes in Indonesia. The difference, as an objective measure of corruption, is affected by the threat of audits, but generally not by the monitoring of community meetings. The result suggests that traditional, top-down monitoring, such as a government audit, can play a role in precluding corruption, even in a highly corrupt environment.

While enforcing and enhancing institutions and laws to combat corruption appear to have a long way to go in many developing countries, it is argued that improving public access to information is a crucial part of a bottom–up anticorruption strategy. In the mid-1990s in Uganda, the capture and corruption of public funds were serious. For every dollar spent by the central government, the schools received only $0.20 on average. Concerned with this, the Government of Uganda initiated a newspaper campaign in the late 1990s, which published data on monthly transfers of capitation grants to local governments. Reinikka and Svensson (2004) studied this campaign and find striking results: capture was reduced from 80% in 1995 to less than 20% in 2001. The authors conclude that public access to information is a powerful deterrent to the capture of public funds at the local level.

2. Job training

The common question in the field is whether labor market programs in developing countries that provide job training and employment services have increased employment rates and raised participants’ earnings. Riboud, Tan, and Revenga (1994) studied the Mexican Labor Retraining
Program for unemployed and displaced workers and find that participation in the program reduced the mean duration of unemployment for both male and female trainees, and increased males’ monthly earnings. The results also show that the wage gains vary systematically with schooling attainment, with the largest earnings increase for males with 6–12 years of education. Jalan and Ravallion (2003) show that gains from the Argentinean Workfare Program were sizable, amounting to about half the gross wage, and similar for men and women. Younger workers benefited more from the program than other workers. Betcherman, Olivas, and Dar (2004) find that active labor market programs providing services like counseling, job matching, and placement assistance have positive impact on employment and the earnings of participants in developing and transition countries. Training for the unemployed, on the other hand, more effectively increases employment but not earnings.

3. **Small and medium enterprise development**

The importance of SMEs to economic growth, labor employment, and poverty reduction has been increasingly recognized. What public policies should the government consider for effectively promoting SME development? A limited number of studies seek empirical answers. Examining the effect of a business registration reform in Mexico on economic activity, Bruhn (2008) finds that the reform increased the number of registered businesses in eligible industries. This increase was due to former wage earners opening businesses. In contrast, former unregistered business owners were not more likely to register their business after the reform. Moreover, the reform increased employment in eligible industries and the competition from new entrants lowered prices and decreased the income of incumbent businesses. Considering the lack of business skills among microentrepreneurs, Karlan and Valdivia (2009) study a randomized business training program targeted at female microentrepreneurs in Peru. The results show that the program led to limited improvement in business knowledge, practices, and revenues as well as repayment and client retention rates for the microfinance institution.

4. **Gender and development**

A few well-known studies highlight the role women play in development. Pitt and Khandker (1998) find that when the participants of group-based credit programs in Bangladesh were women, the program credit had a larger effect on the behavior of poor households. For example, annual household consumption expenditure increased more in the case of female than male participation for the same amount of money borrowed from the program. In India, a 1993 constitutional amendment requires that one-third of village council leader positions be reserved for women, which significantly increased the influence of women in the provision of local public goods in rural areas. Examining the impact of this rule in West Bengal, Duflo and Chattopadhyay (2004) find that village councils headed by women were more likely to invest in public infrastructure—such as water, fuel, and roads, which are seen as directly relevant to the needs of rural women—while male leaders were inclined to invest in education. Moreover, female leadership encouraged women’s participation in village policy making.

Female empowerment programs engage women in productive activities such as grain and spice processing, handicraft production, and animals and poultry raising. What are the impacts of such programs on women’s economic, social, political, and psychological outcomes? Evaluating the Women Development Initiatives Project—which invested in women’s skills, productivity, and organizational capacity—Legovini (2006) shows that program participants on average earned more per month and that additional income increased household purchases.
This chapter provides a detailed review of evaluation studies on several development interventions, namely, the private school voucher program in Colombia, provision of piped water in rural India, rural transport project in Viet Nam, rural electrification in post-apartheid South Africa, group-based microcredit in India, and horticultural export promotion program in Kenya. The selected interventions cover the main development sectors: education, water and sanitation, transport, energy, microfinance, and agriculture sectors, respectively. More importantly, the studies represent the frontier of impact evaluations in each sector. They either employ cutting-edge methodology to produce solid evidence, provide fresh perspectives or insights for policy discussion, or examine an innovative and promising intervention. Through detailed account of these studies, some kind of benchmark may be set for future evaluations conducted by ADB.

A. Vouchers for Private Schooling in Colombia

Private sector participation in education services delivery is increasingly being considered as an effective way to improve capacity and quality of public education. In late 1991, Colombia introduced the Programa de Ampliacion de Cobertura de la Educacion Secundaria (PACES) program in order to expand school capacity and increase secondary school enrollment. The PACES program involved offering vouchers (that partly covered the cost of private secondary schools) to children aged 15 or below who were finishing primary school. Nearly half of private schools in the 10 largest cities participated in the program in 1993. These schools, mainly serving low-income populations, resemble the public schools more than the nonparticipating private schools in terms of pupil–teacher ratios, facilities, and student performance.

The PACES program set restrictive criteria on eligibility to target low-income families. The applicants had to come from neighborhoods classified into the two lowest socioeconomic strata (out of six). Furthermore, only children who had attended public primary school and were admitted to a participating private secondary school qualified. The voucher would be renewed every year as long as the recipient moved up to the next grade with satisfactory academic performance. The conditional renewal provided strong incentives for voucher recipients to work hard to achieve good grades. PACES vouchers were worth about $190 in 1998, while the average matriculation and tuition fees for the participating private schools amounted to about $340. Although voucher recipients needed to supplement the voucher with their own money, the program was still oversubscribed. Some cities and towns therefore used lotteries to randomly select applicants for awarding vouchers.
Depending on the ratio of applicants to available vouchers, voucher award rates varied considerably by city and cohort. In Bogotá, for instance, 58.8% of the applicants won vouchers in 1995 while 84.7% won in 1997. PACES awarded over 125,000 vouchers to children from low-income families, one of the largest school voucher programs to date.

Angrist et al. (2002) and Angrist, Bettinger, and Kremer (2006) obtained lists of voucher lottery winners and losers from regional offices of the Colombian Institute for Education, Credit and Training Abroad, which ran the program. From the summer of 1998, they interviewed by phone roughly 1,600 PACES applicants on the lists—who applied to the program in 1995 or 1997 in Bogotá or 1993 in Cali.21 By stratification, half of the interviewees won the lottery while half did not. They collected information on age, gender, parent age, education and wage, scholarship use, school choice, schooling, and noneducation outcomes such as marriage and work. As a necessary check for randomized experiment data, the authors compared personal characteristics across voucher status and found no statistically significant differences in age, gender, parent age, education, and wage for the Bogotá-95 and Bogotá-97 samples but not for the Cali-93 sample. Because the Bogotá 1997 cohort is too recent and the Cali-93 sample appears nonrandom, the authors focused on results from the Bogotá-95 sample and presented results from the pooled sample.

The researchers also invited a subsample of interviewees from the Bogotá 1995 cohort to take a multiple-choice achievement test. The test was developed, with the participation of Colombian educators, for native Spanish speakers in the United States and had been previously used in Colombia. Of the 473 invited, 60% (or 283) took the test. Although the response rate to test invitation was about 5% higher for voucher winners, the difference by voucher status was not statistically significant. In addition, the personal characteristics of those tested were generally similar to the sample of the Bogotá-95 cohort, and personal characteristics did not differ between test-taking voucher winners and test-taking losers. This favorable evidence suggests that the sample selected was not serious in taking the test, and comparison of test results between voucher winners and losers should produce unbiased estimates of the effects of obtaining vouchers on educational performance.

Angrist et al. (2002) find that the voucher winners were about 51 percentage points more likely than losers to use some kind of scholarship (including non-PACES scholarships) and 67 percentage points more likely to have ever used a scholarship. The former was 6–7 and 17 percentage points more likely than the latter to have begun sixth grade and seventh grade in private school, respectively, and 15–16 percentage points more likely to be in private school at the time of the survey. On the other hand, vouchers did not change decisions about school attendance, although they did affect the choice between public and private schools. However, the PACES program may have indirectly increased school enrollment by freeing up resources in overcrowded public secondary schools. Lottery winners completed more schooling (0.08–0.16 years) than losers and were less likely to repeat grades. Moreover, the effects were found to be moderately larger for girls than boys.

Private schools may have an incentive to let students with vouchers upgrade even if their performance has failed to meet standards because the vouchers would not be renewed for

21 A total of 6,156 from the three cohorts applied to the program. Interviewers approached nearly 3,000 applicants and the overall response rate was 54%. In view of the fact that winners and losers were almost equally likely to be interviewed, the authors argued that the low response rate should cause little sample selection bias.
children repeating a grade, and these children were then likely to drop out of school. It is therefore important to look into test scores and noneducation outcomes to partly address this concern. The authors find that lottery winners scored over 0.2 standard deviation more than lottery losers, with marginal significance. According to United States norms for the same test, the gain of a 0.2 standard deviation is roughly associated with one additional school year. Again, the effects for girls are larger and more precise than for boys. Furthermore, evidence suggests that the lottery winners were less likely to marry or cohabitate, tended to be working, and worked fewer hours than losers.

As some lottery losers obtained scholarships from elsewhere, while some winners did not use or retain their PACES scholarships granted by vouchers, the results summarized above are actually the impact of winning a scholarship rather than the impact of using a scholarship. Since final use of a scholarship is a highly self-selected decision, the authors employed the IV technique to estimate the impacts of scholarship use using lottery status as the IV for voucher use. They find that using a private school scholarship increased schooling and test scores, reduced repetition and likelihood of marriage or cohabitation, and had no effect on school attendance. The IV estimates are generally larger than the corresponding ordinary least squares (OLS) estimates in magnitude, except for test scores.

Finally, the authors considered the overall cost and benefits of the program. The social cost of the program was calculated at $43 per year per child (the sum of $24 of extra public education expenditure and $19 of a winning household’s net contribution). The program resulted in an additional 0.12 to 0.16 grades completed and about a 0.2 standard deviation higher on test scores. Given about 10% return to a year of schooling in Colombia and the prospect of $3,000 annual earnings for the participant, the wage gain attributed to PACES ranges from $36 to $300 per year. Thus, the net present value of the benefits easily outweighed the social cost of the program.

While Angrist et al. (2002) focused on the short-term impacts of PACES, Angrist, Bettinger, and Kremer (2006) gained access a few years later to the administrative records of Colombia’s centralized college entrance examinations. This enabled a follow-up study to evaluate the long-term impacts of the program. Since 90% of all graduating high school seniors take the exam in Colombia, registration with the Instituto Colombiano para la Evaluación de la Educación (ICFES) is a good proxy for high school graduation. When PACES applicants of the 1994 cohort were matched to 1999–2001 ICFES records, those who were not identified in the ICFES records could be treated as not graduating from high school to a large extent. But the issue of matching quality arises. If voucher winners’ identification (ID) numbers were better recorded, they would be more likely to be matched with ICFES records. To address this concern, the authors restricted their sample to those with valid ID and age information. Moreover, they tried four different match definitions: exact ID match; ID and city match; ID and name match; and ID, city, and name match. Essentially, what this does is vary the proxy for high school graduation and test the robustness of the estimated effects of vouchers on high school graduation. They also estimated the effects of PACES vouchers on exam performance, which is contained in ICFES records.

Across various match definitions, 32%–35% of PACES applicants were matched to ICFES records. The results show that winning vouchers raised ICFES registration rates, and probably high school graduation, by 5–7 percentage points relative to a base rate of 25%–30%. The effects are robust to changes in sample, specification, and definition of a match. Estimating the effects on test scores of a college entrance exam is not straightforward because winning a PACES voucher significantly affects test-taking probability. The authors tried different econometric methods to
control this selection bias and conclude that the voucher program has led to a substantial gain in academic achievement as well.

The findings of these two studies suggest that a demand-side subsidy program like PACES can be a cost-effective way to increase private participation in public education and lead to educational attainment and academic performance, especially in developing countries where public education is weak and private schools are relatively well developed.

B. Piped Water Access and Diarrhea among Children in India

Lack of access to clean drinking water is a significant challenge in developing countries. The World Health Organization estimates that diarrhea is the second cause of death in children under 5, and is responsible for killing 1.5 million children every year. Contaminated food and water sources are regarded as the major cause of high diarrhea incidence in developing countries (WHO 2009).

The challenge has motivated many public programs to expand piped water access to improve drinking water quality of the poor. However, Jalan and Ravallion (2003) argue that expanding piped water is not sufficient for improving child health because private behavior, especially that of parents, plays an important role in preventing diarrhea. It is the combination of public provision of piped water and proper private inputs such as hygienic water storage and boiling water that can fight diarrhea effectively. However, a poor or poorly educated parent may not make these necessary complementary efforts. As a result, the health gains from access to piped water may not be realized or maximized for children from poor families even when the programs aim at the poor.

Jalan and Ravallion (2003) utilize cross-section data of Indian rural households to quantify the child health benefits of piped water in terms of reduced diarrhea incidence. The data, collected through a nationally representative survey in 1993–1994, contain detailed information on education and health status of 33,000 rural households from 16 states of India. The data, with household weights applied, show that 24.8% of households had piped water; among which 7.6% is inside and 17.3% outside the house. The outcome variables of interest are the incidence of diarrhea among children under 5 and the reported illness duration. The focus of the paper is how the impact of access to piped water on child diarrhea varies with household income per person and the highest education level of any female in the household.

The authors use the propensity score matching method to estimate the causal effects of piped water on child health. As opposed to the regression method, matching estimation does not impose arbitrary assumptions about functional forms and error distributions. First of all, the propensity score is estimated for each household from a Logit model, which equals the predicted probability of the household having piped water in their house based on the observed characteristics of the household and the village in which it is located. Second, households without piped water are matched to households with piped water through the nearest neighbor approach. The average outcome of the closest five households without piped water, measured based on the estimated propensity scores, is taken as the counterfactual for each household with piped water. The average impact of access to piped water is the mean difference of the outcome of the households with piped water and their counterfactuals.

The Logit model indicates that households living in larger villages with better infrastructure and education and commercial facilities are more likely to have access to piped water. Wealthier
households, electrified households, and female-headed households are more likely to have piped water. Other household characteristics such as religion and tribe backgrounds also play important roles. After matching, the mean propensity scores of households with piped water and those without are indifferent.22

The matching estimates indicate that among households with piped water, diarrhea prevalence would be 21% higher and illness duration would be 29% higher without piped water, suggesting significant positive effects of piped water on child health. When the sample is stratified by household income per capita, however, no child-health gains are found among the poorest two quintiles. Similarly, health impacts of piped water are found larger and more significant in families with better educated women.

The authors further stratified the sample by both income and education. They find that even in the bottom two income quintiles, a household with a woman having more than primary schooling can obtain the benefits of piped water in terms of lower prevalence and duration of diarrhea among children. On the contrary, low-income households with female members’ education no more than primary school do not extract the health benefits. For households from the upper quintiles, the gains of access to piped water to child health are significant irrespective of the female members’ education levels. These results suggest that the education of women plays a particularly important role for the poorer household in realizing the health benefits from piped water.

There might be differential effects of accessing piped water via a tap within the household’s premises or via a public tap nearby. To address this issue, the study compared households with a private tap to those relying on public tap for drinking water. The main findings are that overall private tap access reduces illness duration significantly, but not diarrhea prevalence, as opposed to public tap; for households where the female member is uneducated, the private tap is much more effective in lowering both the prevalence and duration of diarrhea than the public tap.

The study delivers some fresh messages for policy. While the mean impacts of access to piped water on child health are statistically and quantitatively significant, parents’ knowledge and behavior are necessary complementary inputs for realizing the benefits in poorer families. The evidence that pro-poor infrastructure placement may not end up favoring the poor calls for a combination of public investments in infrastructure provision with other interventions in education and income–poverty reduction.

C. Rural Roads in Viet Nam

In 1997, the World Bank financed the Rural Transport Project 1 in Viet Nam. The project was intended to reduce poverty in poor communes by enhancing transport links between communes and markets (World Bank 1996). Communes are the lowest level of government in Viet Nam and an average commune observed in the study had about 1,300 households living in 10 or so hamlets. Above the communes are districts and provinces. The project, with a budget of about $61 million, was designed to rehabilitate 3,500 kilometers (km) of district roads and 1,500 km of communal roads in 18 poor provinces from 1997 to 2001. The road links selected for rehabilitation were mostly impassable by motor vehicles and rehabilitation was aimed at a minimum-cost engineering solution that ensured

---

22 A more sophisticated test of how well the matching process works should compare the means of all the observed characteristics of the two groups. The authors did not mention these results in the paper, however.
that roads were passable at a minimal level for motorized vehicles. The project stipulated that no new road construction should be considered under the project.

Although it is commonly believed that rural roads can alleviate poverty in rural areas by increasing access to goods and services, enhancing agricultural production, and stimulating diversified income-earning opportunities, there is little hard evidence actually demonstrating these benefits. The impact evaluation conducted on this project was the first comprehensive attempt to rigorously assess the effects of rural roads on a poor rural area. The study assumed the commune as the project’s zone of influence and thereby focused on the road impacts at the communal level. Note that as the road links were selected based primarily on the population and cost criteria, i.e., average investment costs no more than $15,000 per km and at least 300 people per km being served, it is implausible to do a randomized trial for rehabilitation assignment. The evaluators mainly relied on data of a sample of project and nonproject communes, collected before and after the intervention, to estimate the impacts of the project.

Three data sources were used in the evaluation. First, data on 200 project and nonproject communes were collected through a survey designed specifically for the project. Six of the 18 project participating provinces were randomly selected for the survey from the country’s north, center, and south regions, with two provinces each. A random sample of 100 communes was drawn from all project communes in these six provinces, while another 100 communes were randomly selected from all nonproject communes that were located in the same districts as the sampled project communes. Second, a household questionnaire was administered to 15 households in each sampled commune. Instead of being evaluated directly, the household information was used to create additional commune level variables. Finally, a project level database was constructed, which detailed project location, timing, costs, the condition of the road preproject and postproject, kilometers rehabilitated, and whether bridges were rehabilitated.

The baseline was collected before the project started in June 1997. The subsequent rounds followed in the summers of 1999, 2001, and 2003. Thus, the 2001 and 2003 waves were done after the completion of the project. The complete data set, consisting of a four-period panel of 200 communes, was quite rich in commune information. Nearly 100 variables appeared in the analysis encompassing a number of outcome variables measuring good availability of food; daily use and services; medicine and fertilizer; as well as community development with respect to transportation, markets, employment in agriculture and nonagriculture, and health and school services. World Bank staff designed the survey and supervised data collection carried out by the Institute of Economics in Ha Noi. Excluding World Bank staff time and travel expenses, the evaluation cost about $200,000 in total (Baker 2000).

In addition to the DID method routinely applied to the panel data, the researchers combined propensity score (PS) matching and propensity score weighting with DID to construct a better counterfactual. First, they predicted the probability of participating in the project (namely, PS) for each commune as a function of observed commune characteristics in 1997. In the PS-matching case, a comparison group comprising nonproject communes with the closest PS was assigned to each project commune, which was followed by nonparametric DID to obtain the average treatment-on-the-treated effect. In the PS-weighting case, the observations were weighted with

23 Priority was given to mountainous areas with concentrations of ethnic minorities, which failed to meet these criteria.

24 Interested readers are referred to various tables in van de Walle and Mu (2007) and Mu and van de Walle (2007) for a full list of covariates and outcome variables.
the PS, and a DID-type regression was applied to the weighted sample. The method was able to remove selection bias stemming from road placement determined by the initial baseline conditions that also influenced the outcomes of interest. However, in the presence of unobserved, time-varying factors correlated with both placement and changes in outcomes, the estimates will still be biased.25

Prior to studying the effects of the project on rural development, the researchers examined whether the donor’s aid financed what was intended in the context of Viet Nam’s project (van de Walle and Mu 2007). In other words, was there fungibility or a flypaper effect with regard to the project aid?26 This is not a question conventionally answered by the impact evaluation, but it was quite important and interesting to both donors and researchers. Comparing the kilometers of rehabilitated and newly built roads during the 2 years before 1997 and before 2001 for the project and nonproject communes, the authors find that the aid resources had little impact on rehabilitated road kilometers while more roads were built in project areas. The findings tell a story of partial fungibility as the project clearly prohibits building new roads under the project, and partial flypaper effect in that the money largely stuck to the transport sector.

Mu and van de Walle (2007) looked at the impacts of rural road improvements by 2001 and 2003. Because of their findings, the reported impacts were attributed to both road rehabilitation and new road construction. They examined two broad categories of outcomes, i.e., local community development and goods availability. In particular, they find that by 2001, the road project increased employment opportunities for unskilled labor; the number of upper secondary schools, and the primary school completion rate. By 2003, more markets became newly available and market frequency increased in the project communes than in the nonproject communes. There was evidence by 2003 that a share of households switched from agriculture to the service sector. In terms of goods availability, consumer goods and production inputs seemed to respond at different speeds to the road improvements. A number of food items showed up in the project communes by 2001, but the differences vanished 2 years later. On the contrary, significant positive impacts on availability of some nonfood goods, such as fertilizer and gasoline; and services, such as hairdressing and tailoring, were discovered by 2003 but not by 2001. Finally, there seems to have been little evidence of road impacts on many other outcomes, including land market development and transport services.27

Besides the average treatment effects, the researchers also find enormous heterogeneity in the effects of the project. While the poorer communes, often located in mountainous areas, tended to have higher impacts on many indicators of market development, a high concentration of ethnic minorities and high illiteracy rates tended to lessen the impacts, holding other characteristics constant. Several commune attributes, such as distance to the nearest city, initial accessibility, and distance to the closest market strengthened impacts on some outcomes while lowering impacts on others.

A major limitation of the study is that the evaluation was focused on the impacts of road improvements on communes where the road links pass through. The project could well have

25 Mu and van de Walle (2007) also use a Multiple Indicator Multiple Cause model developed by Jöreskog and Goldberger (1975) to explore heterogeneous impacts of various covariates on a vector of summary outcome variables.
26 The flypaper effect is somewhat opposite to fungibility, referring to the fact that external funding stimulates higher spending of the intended type by the government. Thus, the money “sticks” as flies on flypaper.
27 All the results should be considered with caution since estimates are sensitive to the choice of estimation method.
generated spillover effects to neighboring areas, which the study could not say much about. It also reflects the difficulty in evaluating transport projects. It is conceivably even harder to construct the counterfactuals for projects building intercity highways or intracity arteries. Van de Walle (2008, 2) even claims in her comprehensive review that “highways and long distance trunk roads joining up regions, that tend to have macroeconomic effects and are not assigned to specific communities—so that a counterfactual cannot be identified based on observable data, and classic evaluation tools are useless—are not considered in depth.”

D. Rural Electrification in South Africa

A significant share of energy development projects has aimed to electrify rural areas where people collected wood for fuel and used kerosene lamps for lighting before the project. A variety of direct and indirect benefits are associated with rural electrification (World Bank 2008). But very little concrete evidence has ever been available, especially with respect to the indirect benefits such as promoting employment. We introduce here a recent study (Dinkelman 2009) evaluating the employment effects of a rural electrification campaign in post-apartheid South Africa.

In 1993, a year before the end of apartheid, over two-thirds of households in South Africa were without electricity. Most were black African households settled in so-called homeland areas and denied access to basic services including electricity. All homelands were reintegrated into the country after the 1994 democratic elections, and the new government set up the National Electrification Program, making this its development priority. As part of the National Electrification Program, Eskom, South Africa’s national electricity utility, committed to electrify 300,000 households annually from 1995 onward. A key feature of the rollout program was that it focused on, and did reach, poor households. Both local political pressures and connection costs played important roles in the prioritization of communities for electrification. While political factors are difficult to measure given the available data, the bulk of connection costs fell on laying distribution lines from the substations to households, which largely depended on three measurable factors, i.e., proximity to substations, settlement density, and land gradient and terrain. Eskom had strong incentives to allocate projects first to the areas with lowest average cost per household connection. From 1993 to 2003, more than 470,000 households were electrified under the roll-out in KwaZulu-Natal province alone, the focus of the study.

Dinkelman assembled community-level data from two publicly available census surveys in 1996 and 2001, as well as Eskom project data and geographic data. Geographical information system software was used to calculate the average land gradient and other special variables for each community and to merge census communities with Eskom project areas. The final data set includes 1,992 rural ex-homeland communities in KwaZulu-Natal, which were not exposed to the electrification project before 1996. A total of 391 communities had their first electrification projects from 1996 to 2001, while the rest (1,601) did not. The author examined the impacts of electrification on household energy sources and gender-specific employment. The information used as covariates in the analysis covered a community’s household density; poverty rate; adult sex ratio; fractions of female-headed households; Indian and white adults; men with high school education; women with high school education; distances to the nearest town, road, and grid by 1996; and changes in access to water and toilet facilities.

28 Electrification of a community is defined as a community having experienced a spike in household electricity connections according to Eskom’s data.
Given the 2-year panel data containing both treated and untreated communities, a straightforward design for impact evaluation would be the DID method. If electrified communities differ from unelectrified communities in some time-invariant aspects, which are also correlated with levels of outcomes, the DID approach can solve the selection bias problem. However, as the author argued, the electrification project in this case and in many other infrastructure projects all over the world were not randomly deployed. They are often targeted at either growth centers or lagging areas. These are likely to have growth paths distinct from those areas less favored for the infrastructure investment. In other words, placement of the projects may be determined in a nonrandom way, which is correlated with changes in outcomes. If this is true, the DID estimators will be biased.

Concerning the potential bias of DID estimation, Dinkelman proposed to use a community’s average land gradient as an instrument for the electrification status. The observation is that land gradient determines the connection costs in part, which is a primary factor in prioritizing areas for electrification. The less steep the land is, the cheaper it is to lay power lines, and the more likely a community will be electrified earlier than the rest. To have this instrumental variable work, the assumption must hold that conditional on the covariates, such as distance to town and road, the land gradient does not affect employment growth through any other channel.

First of all, the results clearly demonstrate that the land gradient significantly affected the probability of a community being electrified during 1996–2001. One standard deviation increase in average gradient (about 10 degrees) reduced the probability of electrification by 4 percentage points. Greater gradient also delayed the time it took to be electrified and decreased the fraction of households electrified. One channel through which electrification causes changes in employment patterns is that households with electricity switch out of traditional fuels and spend less time on home production. To manifest this point, the author looked at the effects of electrification on energy sources for lighting and cooking first. The DID estimates show that electrification raised electric lighting and cooking by 23 and 6 percentage points, respectively, and lowered reliance on wood for cooking by 4.2 percentage points. The IV estimates were 71 percentage points higher in electric lighting, 24 percentage points higher in electric cooking, and 28 percentage points lower in wood use, which are all substantially larger than the DID estimates.

The above results imply the effects of electrification on employment could be large as electricity saves time for collecting wood and cooking. The DID and IV estimates exhibit distinct implications for the employment responses to electrification. The IV estimates indicate a statistically significant 13.5 percentage point increase in female employment, while the DID estimates show nearly zero effect. For male employment, the IV estimate yields a smaller (4.2 percentage points), statistically insignificant increase, while the DID estimate shows a marginally significant reduction of 1 percentage point. The IV results are consistent with the story in which females, as the primary home workers, were freed up to join market production as a result of electrification, while electrification did not hurt males’ labor participation. The author further explored the heterogeneity in the impacts on female employment and finds that women from the middle-poor and the second-richest communities that initially relied heavily on wood for cooking responded more to electricity access. Moreover, effects were larger for women in their 30s and 40s who have fewer child-care responsibilities.

This study represents some of the first hard evidence on the impacts of energy infrastructure in a developing country. It illustrates the critical role played by diligent data collection and synthesis as well as the significant differences resulting from applying correct methods in impact evaluation. It
also highlights the importance of including employment measurement in designing monitoring and evaluation frameworks for infrastructure projects. The study, however, is not without problems. As the rollout was not random and the sample comes from a truncated period, it may not represent the overall population. The validity of the instrumental variable may need to be closely scrutinized. If the land gradient were to affect some employment-related unobservables (e.g., commuting costs) the estimates would be biased. Finally, it may be better to have a similar analysis at the household level, where measures of treatment and outcomes can be more accurate, than the community level.

E. Group-Based Microcredit in India

Advocates of microfinance, in particular microcredit, claim that evidence mounts to show that microfinance contributed to eradication of poverty and hunger, universal primary education, gender equality, reduction in child mortality, and improvement in maternal health. Skeptics, on the other hand, are concerned with possible negative effects of microfinance, such as overborrowing leading to greater long-term poverty, and are not satisfied with the evidence available. It is of first-order importance to evaluate the impact of microfinance on individual welfare and local development, especially given the presence of other competitive antipoverty measures. Nevertheless, unbiased estimation of its impact is quite a big challenge when only observational data are available. The issue is that clients of microfinance are highly self-selected and thus dissimilar to nonclients in both observable characteristics, such as education, health, preloan income, and unobservable ones, such as entrepreneurial ability or innate motivation (Bauchet and Dalal 2009). Moreover, MFIs do not open their offices randomly.

Banerjee, Duflo, Glennerster, and Kinnan (2010) report on the first randomized evaluation of the effect of the group-lending microcredit model. Two randomizing approaches may be adopted in a microcredit project. One is to randomly select individuals among eligible applicants to receive credit. The other approach is to randomly select areas to avail of microcredit services. Considering possible spillover effects, it is more desirable to randomize at region rather than individual level.

In 2005, Spandana surveyed 120 neighborhoods, referred to as “slums” for permanent settlements of the poor, in Hyderabad, India, and decided that 104 neighborhoods are suitable for opening branches. These are areas having no preexisting microfinance facilities, having poor but not the poorest residents who are potential borrowers, and containing fewer migrant workers. These 104 neighborhoods were paired by per capita consumption, fraction of households with debt, and fraction of households who had a business. One of each pair was then assigned to the treatment, meaning that Spandana opened a branch and began progressively operating in the area between 2006 and 2007, and the other area was left without Spandana’s engagement.

A household survey was conducted for each of the 104 areas between August 2007 and April 2008. For 52 treatment areas, the survey began generally 15 to 18 months after the disbursement of loans from Spandana. Because of low rates of MFI borrowing, the survey oversampled households whose characteristics suggested high borrowing propensity and Spandana borrowers. However, the oversampling was corrected in the econometric analysis. Data including over 6,000 households show that households are statistically indifferent between treatment and comparison areas in terms of literacy, the likelihood that the wife of the household head works for a wage, the size of the household, number of women aged 18–45, the percentage who operate a business opened a year or more ago, and the likelihood of owning land. Although other MFIs also started operating both in
treatment and comparison areas, households in treatment areas are still more likely to borrow from Spandana than other MFIs.

Simply comparing averages across households in treatment and comparison areas, the study estimates the impact of microcredit becoming available in an area on business operation, household spending, education, health, and women's empowerment. The results suggest that microcredit has a significant positive impact on creation of new businesses as well as on the profitability of preexisting businesses. It also increased expenditure on durables, although it had no effect on average monthly consumption per capita. Households in treatment areas did not spend more on children’s health and education, and women in these households were no more likely to be making decisions about household spending, investment, savings, or education than those in comparison areas, at least over a short term.

The authors also explored the heterogeneous effects of microcredit availability. They find that households with high business propensity started more businesses in treatment than in comparison areas. Households who owned a business before the start of the program, or who were likely to start a new business, showed a significant positive treatment effect on durables spending. The latter also reduced spending on nondurables and temptation goods, probably to finance the fixed cost of starting a new business. In contrast, those households who have the lowest propensity to start a business increased consumption of nondurables and temptation goods. The results further suggest that microcredit opportunity led to a positive increase in business profits for existing business owners, although the estimate varies with the sample used.

Banerjee, Duflo, Glennerster, and Kinnan (2010, 31) conclude that in the short term at least, “microcredit does not appear to be a recipe for changing education, health, or women’s decision-making. Microcredit therefore may not be the ‘miracle’ that is sometimes claimed on its behalf, but it does allow households to borrow, invest, and create and expand businesses.”

F. Horticultural Export in Kenya

Although Kenya exports a significant amount of horticultural produce to Europe, the contribution of smallholder farmers to total horticultural exports is limited. Several stylized constraints make it difficult for smallholder farmers to further participate in this market. In general, smallholder farmers lack information on pricing and exporting opportunities, lack reliable production contracts with large brokers or exporters, and lack access to credit and transportation services. As a result, they continue to grow crops for local markets, foregoing the export market, which is conceivably more profitable.

To overcome these barriers, DrumNet, a Kenyan NGO, designed a program that provides a package of services to help farmers adopt, finance, and market export crops—mainly French beans and baby corn. A farmer who is a member of a registered farmers group (known as self-help group or SHG), has irrigated land, and meets a minimal financial requirement can participate in the program. For its clients, DrumNet offered an orientation course, distributed seeds and other inputs, monitored planting, negotiated price with exporters, organized transportation, and closed transactions. As

29 Note that this impact is different from the impact of actually getting microcredit. The former is called intent to treat (ITT) estimates in the evaluation literature. With the positive effect of availability of microcredit on the likelihood of getting microcredit, as in this case study, the two impacts differ in magnitude but not in direction.
A variant, some clients were also given group-liable loans. The program is similar to a typical out-grower scheme common in horticulture and other export crops industries except that DrumNet acted as a third neutral party to promote trust between farmers and exporters.

The study by Ashraf, Karlan, and Gine (2008) on the impact of DrumNet’s program on farmers’ crop production and household income represents the first randomized evaluation of a program of this kind. As the program includes two treatments, one with a full package of services including credit and one with all services except credit, it is interesting to learn whether there are additional benefits of providing microcredit along with the service package. One concern with the evaluation of the program is that individuals who are more entrepreneurial or better motivated are more likely to participate and hence those who did not participate are not a good comparison group for the participants. To deal with likely selection issues, Ashraf and her colleagues conducted a randomized controlled experiment with DrumNet to evaluate the program’s impact.

After screening, 36 registered SHGs in Kirinyaga, Kenya were chosen and randomly assigned into three experimental groups, i.e., treatment with credit, treatment without credit, and control. Each group consists of 370 or so farmers from 12 SHGs. Baseline and follow-up surveys were conducted in April 2004 and May 2005, respectively. Data show that SHGs and individual farmers in the three groups were balanced on most variables except for infrastructure and remoteness of SHGs and individual borrowing and household income. Farmers in the sample grew subsistence crops (beans, maize, potatoes, and kale) half the time, and cash crops such as coffee and bananas 34% of the time. Only 12% of the farmers already grew French beans, and none grew baby corn.

While 41% of SHG members from treatment-with-credit group joined DrumNet, only 27% from treatment-without-credit group did so. However, regressions controlling for other households characteristics show that the impact of the credit component is positive but not statistically significant in explaining program participation. Regression results also indicate that literacy and leadership are positively correlated with joining DrumNet. There is an inverted U-shape relationship between income and program participation, suggesting that neither the wealthiest nor the poorest farmers signed up for DrumNet.

The authors estimated the intent to treat (ITT) effect, measuring the average impact of offering the DrumNet program, as well as the average treatment-on-the-treated (ATT) effect, measuring average impact of participating in the program. The former is obtained by simply comparing individuals in the treatment groups and the control group, while the latter calls for the use of the IV technique (using random assignment to treatment as an instrument for program participation) to address the selection bias. Given availability of baseline information, a panel model including a dummy for each SHG and individual baseline characteristics is utilized to improve causal estimates in the presence of imperfect randomization.

First of all, the study finds that the program had a strong and significant effect on crop choice. Farmers in the treatment SHGs (both with-credit and without-credit) were 19.2 percentage points more likely to grow an export crop than those in the control SHGs. Likewise, a greater proportion of land was dedicated to cash crops in the treatment groups, whereas expenditure on inputs did not

---

30 For IV estimates to be valid, one needs to assume that there is no within-group externality. In this study, it means that farmers who were affiliated with the treatment SHGs but did not participate in the DrumNet program were not influenced by those participating. The assumption is a concern for the study, so the authors suggest taking the ITT estimates more seriously than the ATT estimates.
increase. Second, large increases in production in kilograms were found for baby corn, but not for French beans. Third, the program was found to have a positive though not statistically significant impact on household income. Finally, the program seems to encourage households to obtain loans from formal sources (other than DrumNet). A somewhat surprising finding is that no significant differences were found between credit treatment groups and no-credit treatment groups.

When the sample is split between farmers who started growing export crops before DrumNet and those who did not, heterogeneous impacts exist for the two subsamples. It is the first-time growers of export crops who devoted more land to cash crops and increased production of French beans. More importantly, the program increased household income by 32% for first-time growers in the treatment groups as opposed to the control group. No similar impact on income was observed for the prior growers. The benefits of the program for prior growers concentrated on a reduction in marketing costs.

Although DrumNet succeeded in turning farmers to more profitable markets and convincing buyers to purchase farmers’ produce, the program unfortunately did not last long when the exporter in 2006 stopped buying from DrumNet farmers, who failed to meet new European Union certifications. DrumNet lost money as farmers defaulted on the loans. Farmers had little choice but to return to growing what they had been growing before. This experience may tell why farmers had not grown export crops after DrumNet.
VI. Key Areas for Practical Consideration

A n impact evaluation study normally involves project initiation, team composition, evaluation design, data collection, data analysis, and publication and dissemination of results. In addition to large financial and intellectual resources demanded, a number of practical issues need to be considered and addressed in order to facilitate the study. Below is a summary of some experiences and lessons learned from existing practice in four key areas.

A. Evaluation Planning and Design

Once the need for an impact evaluation is recognized, the project manager should assess up front whether the financial requirement for the study can be met. Budget availability significantly affects the whole evaluation design, data collection, the ability to use appropriate methodology, and the kind or level of expertise that can be engaged. Most cases of successful implementation of evaluation design had strong financial commitment from donors and/or client partners. Donors are usually the most likely funding source, while the host countries may sometimes be willing to share some costs. Third-party funding is also possible with organizations such as CGD, International Initiative for Impact Evaluation (3ie), etc. In some cases, the evaluation design may be made to fit the limited budget and still achieve reliable results. But in other cases, without sufficient funding, the evaluation has to be cancelled.

In the meantime, the decision to proceed with an evaluation should largely be contingent on the political willingness of the client government (Bedi et al. 2006). The project manager needs to obtain support for the evaluation from the government by addressing potential political concerns. Many governments do not see the value of evaluating projects, not to mention investing resources in it. There may also be reluctance to allow an independent evaluation that may produce results unfavorable to government policy. In such cases, considerable effort is needed to convince the government of the benefits of the evaluation in terms of confirming and demonstrating positive policy effects and strengthening future policies. Political concerns about the experimental approach for evaluating programs may be dismissed on the basis that randomization is a transparent and fair way to assign treatment given limited resources.

When the impact evaluation is perceived to be financially and politically feasible, the next step is to select the lead evaluator who will be tasked to ensure quality. The key qualification a lead evaluator needs is probably knowledge of different program evaluation methodologies and experience in applying those methods. Experience with sector projects and survey administration
in development contexts is also expected for an evaluation expert to lead the study. Next, with the lead evaluator, the project manager should identify and recruit qualified individuals to form the evaluation team. A survey expert is probably the most important one if primary data need to be collected through surveys. Other members may include someone familiar with the project sector, such as a health specialist in evaluating a water supply and sanitation project. Local organizations or individuals are usually hired for data collection. In this case, a local manager among them may be designated to coordinate the evaluation with the lead evaluator.

Together, the evaluation team comes up with an evaluation design. This should have, but not be confined to, the following features. First, the evaluation question should be clear, specific, and straightforward. A question like “What kind of projects improves child health most effectively?” is too broad to answer through one impact evaluation. Second, what is to be measured should be clear in terms of both the intervention and outcomes. The required information should be available in the existing data or obtainable via survey. Third, a wide range of methodological options should be assessed and the most appropriate methodology chosen. Each method has its weaknesses and strengths, and no single method is perfect. The primary criterion in choosing among the evaluation methods is one which best constructs the counterfactual. The choice of methodology is also affected by the budget available, project type, evaluation question, data availability, etc. Fourth, adequate data should be collected in a cost-effective way to answer the evaluation question. The data should cover a variety of groups for establishing the counterfactual and exploring heterogeneous impacts, have sufficiently large sample size to achieve statistical power, and contain the full range of variables to be used as controls. The subsequent discussion will further elaborate on lessons learned about collecting data.

Advance planning and careful design of an evaluation study are essential to a successful impact evaluation. The District Primary Education Program (DPEP) in India, the largest World Bank–sponsored education program, is an example of a large program with potentially very interesting evaluations that failed to produce impact evidence due to a lack of planning. DPEP was a comprehensive program involving teacher training, inputs, and classrooms that seeks to improve the performance of public education. Despite the apparent commitment to a careful evaluation of the program, several features made a convincing impact evaluation of DPEP impossible. First, the treatment districts were selected according to two criteria: low level of achievement and high potential for improvement. In particular, the first districts chosen to receive the program were selected based on “their ability to show success in a reasonable time frame” (Pandey 2000, 14). The first selection criterion made clear that any comparison between the levels of achievement of DPEP districts and non-DPEP districts would probably be biased downward, while the second criterion would result in upward bias in any comparison between improvements of achievement between DPEP and non-DPEP districts (difference-in-differences). This did not prevent the DPEP from putting enormous resources into monitoring and evaluation: large amounts of data were collected, and numerous reports were commissioned. However, the data collection process was conducted only in DPEP districts. These data can only be used to do before–after comparisons, which clearly do not make sense in an economy undergoing rapid growth and transformation.

B. Collecting Data

As data collection often represents more than half of the cost of an evaluation, exploring what data exist and can be accessed is the first important step before launching any new data collection effort. Existing surveys, census, and administrative records from government agencies, schools, hospitals, local markets, etc. may provide estimates of pre-intervention conditions as well as
While using secondary data can save significant amounts of money, it is important to assess how well the data sets satisfy the need of the evaluation in terms of timing, completeness, adequacy, and potential bias. Ideally, secondary data should be collected close to the start of the project, cover a sufficiently large sample of all groups of interest, contain information on all key project variables and outcome indicators, and measure potential control variables such as demographics and expenditure. When secondary data are used as the baseline, one has to figure out how postproject data can be linked to it at the level of the analysis unit (village, household, etc.). Caution is also warranted due to systematic biases in some administrative records. For instance, the number of children attending schools tends to be overreported and the number of crimes underreported.

If new collection of data is necessary, the evaluation team should carry out the baseline before project implementation and the postproject survey after project completion, with sufficient time for the impacts to take shape. The process should be guided by the lead evaluator and the survey expert. It is advisable to plan the before and after surveys all at once, taking into account issues such as what information would be collected in each survey, how to efficiently allocate the budget between the two rounds, how to ensure the same sample units will be followed through time, etc. Designing survey instruments requires expertise in statistical sampling and questionnaire design and knowledge of the project, sector, and local culture. Thorough consideration needs to be given to sample size, sample representativeness, and information adequacy and accuracy, among other things. Pretesting of the questionnaire is recommended to make sure that the survey will work in the field. The data collectors or interviewers must be trained specifically for the surveys, even though they are experienced.

Data costs can be reduced if a proper sampling framework is created and the questionnaire is targeted precisely at the right information. The experience of the evaluators will play a critical role in deciding what information needs to be obtained and avoiding replication of information between the two rounds of surveys. As discussed in randomized experiment design, if the intervention is much more expensive than data collection, it is recommended to have a large comparison group and to conduct a baseline survey for the randomized experiment, which is not necessary but has extra benefits for the analysis. On the contrary, if the intervention is cheap and data collection is costly, then one may consider experimenting on a larger scale and collect only post-intervention data with a relatively big experimental group. Finally, survey costs may be lowered by replacing face-to-face interviews with telephone or e-mail interviews. The possibility of switching to cheaper alternative survey modes is rising with increasing access to the telephone and internet in developing countries. Respondents without phones can take calls in a community center.

C. Building Partnerships and Capacity

Working in partnership with the project team, research institutions, international consultants, local governments, nongovernment organizations (NGOs), and researchers is always recommended. Impact evaluations are quite demanding in terms of the skills needed to conduct them. An evaluation team needs to have a strong combination of high-level scientific and professional expertise in economics, statistics and surveys, and the project sector, etc. At present, such expertise exists to a large extent only among a limited number of institutions and independent consultants, and is particularly lacking in developing countries (Foresti et al. 2007, Levine and Savedoff 2006). In setting up a partnership

---

31 Experience from World Bank evaluations shows that it is often possible to find high quality secondary data covering most baseline information needs.
with researchers, long-term relationships may have higher return than one-off evaluation, where the researcher is asked to evaluate what the project implementing organization has decided to evaluate. In a long-term partnership, researchers are given the option of defining what is to be evaluated. The collaboration would turn into a continuous process of learning and questions of mutual interest would be answered more efficiently.

Local participation is critical to the success of an evaluation study. It would be easier for the government to recognize the evaluation as legitimate if it is well motivated and engaged at the beginning of the evaluation. The needs of the government relevant to their policy and development strategy could also help shape the evaluation questions. In many cases, local government support is key to acquiring the data for evaluation. In considering whether there is local capacity to implement the evaluation, it may be worth getting in touch with local NGOs in the project sector. In addition to assessing their own projects, NGOs are generally more willing to assist in impact evaluation of donors’ projects, and they often can make up for shortfalls in various types of technical expertise. Some recent examples include the collaborations between the Massachusetts Institute of Technology (MIT), ICS Africa (the Kenyan NGO), and Pratham (the Indian NGO). Letting international researchers and local people work together also enhances the capacity of local institutions and individuals, which may lead to more rigorous evaluation being incorporated into local projects.

D. Disseminating Results

Countries and development agencies should have their impact evaluations go through peer review and published regardless of the conclusions. It may initially be difficult to externally publicize results pointing to problems with some programs. However, organizations doing so could benefit from showing their seriousness about learning from the evaluation and acting on good evidence. Furthermore, an organization’s reports will gain greater credibility, with both favorable and unfavorable reports equally published and debated. Part of this dissemination should include making primary data publicly available for reanalysis, and encouraging the production of systematic reviews of evidence.

Evaluation reports are usually sent to ministry staff or government officials. However, policy makers have little time to read the findings in full detail. Therefore, the evaluation team is encouraged to publish the results in the form of policy briefs. The results could also be deposited into a searchable database, such as those available at the World Bank. Donors may also want to make the evaluation findings more widely accessible among government and nongovernment practitioners, and the public.

E. Summing Up

To conclude, there is no single well-developed evaluation design that can universally be applied to all projects. The optimal design for the project considered varies across and within sectors. While evaluations with creative design and data analysis are desirable, it is advisable to replicate the project evaluation in different contexts and/or using different methods, experimental and nonexperimental. It is worthwhile to exert every effort to obtain the right data, which is a prerequisite of a quality impact evaluation. Another essential component of the evaluation strategy is to build capacity and establish a strong partnership with leading research institutions and NGOs. Finally, the quality of the evaluation and the reliability of the results improve considerably if the impact evaluation is carefully planned and properly designed at an early stage, often at the outset of a new program or project. This can also lower cost substantially.
Impact evaluation focuses on changes in outcome that are attributable to development interventions. It is therefore different from project monitoring or project progress evaluation. Not only does impact evaluation tell us whether the intervention works or not, it should also inform in what context the intervention does or does not work, and to what extent the impacts vary between different groups of beneficiaries. Good impact evaluations will be very useful for guiding future allocation of resources for development.

The core challenge of an impact evaluation lies in credible attribution of the observed changes in the outcome variables to the intervention being evaluated. The word “rigorous” is often used to emphasize the seriousness an impact evaluation takes on the attribution. The difficulty of attribution arises because the intervention is not randomly assigned or individuals do not participate in the program in a random manner. Simply comparing those treated by the intervention and the untreated will result in biases in evaluation results and lead to misleading conclusions about the impact of the intervention.

Different methods based on observational data have been developed to overcome this attribution challenge in impact evaluation. Among them, the difference-in-differences approach is probably the most commonly used due to its reliance on relatively fewer assumptions and flexibility in application. Concerned with the limitations of observational data, scholars and practitioners have increasingly adopted randomized controlled trials for impact evaluation in recent years. This approach proactively uses experimental data collected from the field for impact estimation. While randomized evaluation represents a powerful tool to obtain credible impact estimates, many practical issues such as sample attribution, partial compliance, and spillover need to be looked at and addressed when the evaluation is designed and carried out. In short, all approaches and methods have their merits and limitations. Selection of the most appropriate methodology needs to be done carefully at the very beginning of a project and account for the features and context of the intervention.

The impact evaluation literature has been growing explosively. While a remarkable share of existing studies focus on education and health, other sectors such as infrastructure, agriculture, and microfinance have seen a rapid increase in both number and rigor of evaluations. Some studies confirm while others cast serious doubt on conventional wisdom on the effectiveness of evaluated
interventions. Both high-quality individual studies and surveys constitute valuable knowledge shared by the international development community. As one of the key players in this community, ADB could contribute to this global public good by replicating existing studies in the regional context, implementing and evaluating innovative interventions, disseminating the established findings to its developing member countries, and assisting them in evaluating their own programs and policies.
The standard approach to the analysis of program impact is the Rubin Causal Model, in which the causal inference is formulated as comparisons of potential outcomes (Imbens and Wooldridge 2008). Suppose we have observations for \( N \) units, with some participating in the program and others are not. Let \( Y_i \) be a random variable measuring the outcome of interest for unit \( i \). \( T_i \) indicates the participation status of \( i \). \( T_i = 1 \) if unit \( i \) participates in the program or gets “treated,” and \( T_i = 0 \) if not participating or untreated. Thus, there exist two possible outcomes, \( Y_i(T_i = 1) \) and \( Y_i(T_i = 0) \), for one being treated and not being treated, respectively. For simplicity, we write these as \( Y_i(1) \) and \( Y_i(0) \). \( Y_i(1) \) and \( Y_i(0) \) are counterfactuals for each other. The gain from participating in the program for \( i \) is \( Y_i(1) - Y_i(0) \), which is referred to as program impact or treatment effect for \( i \).

Usually, the potential outcomes are further modeled as linear functions of a vector of covariates, denoted \( X_i \), with additive error terms:

\[
Y_i(0) = X_i \beta_0 + u_{i0} \quad \text{for} \quad T_i = 0, \quad \text{(A1.1)}
\]

\[
Y_i(1) = X_i \beta_1 + u_{i1} \quad \text{for} \quad T_i = 1, \quad \text{(A1.2)}
\]

where \( u_i \) captures the unobservable parts of the outcomes, and \( u_{i1} - u_{i0} \) can be interpreted as the idiosyncratic gain for unit \( i \). \( u_{i1} - u_{i0} \) may be known to the unit although it is not observed by the researcher. Assume \( E(u_{i0}|X) = E(u_{i1}|X) = 0 \).

The evaluation problem is that we do not observe both \( Y_i(1) \) and \( Y_i(0) \) at the same time. The observed outcome can be written as

\[
Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0) = X_i \beta_0 + T_i [X_i \beta_1 - X_i \beta_0 + u_{i1} - u_{i0}] + u_{i0}, \quad \text{(A1.3)}
\]

so \( Y_i \) is either \( Y_i(1) \) or \( Y_i(0) \). The second equality is obtained by plugging equations (A1.1) and (A1.2) into (A1.3). In this case, one must use observations from the group not exposed to the program to form a counterfactual for the treatment group and estimate some kind of average treatment effect. Two parameters given most attention are the average treatment effect (ATE) and the average treatment effect on the treated (ATT), as follows:
\[ \text{ATE} = E[Y(1) - Y(0)|X] \]
\[ = X(\beta_1 - \beta_0) + E(u_1 - u_0|X) \]
\[ = X(\beta_1 - \beta_0) \]  

and

\[ \text{ATT} = E[Y(1) - Y(0)|T = 1, X] \]
\[ = X(\beta_1 - \beta_0) + E(u_1 - u_0|T = 1, X) \]
\[ = \text{ATE} + E(u_1 - u_0|T = 1, X) \]  

Equations (A1.4) and (A1.5) show that ATE and ATT coincide when \( E(u_1 - u_0|T = 1, X) = 0 \), which may be satisfied in two special cases. The first is that the idiosyncratic gain from the program is zero for everyone, i.e., \( u_{i1} - u_{i0} = 0 \) for all \( i \). As a result, the whole model reduces to the common treatment effect model. The other case is that people do not know or act on the idiosyncratic gain, namely

\[ E(u_1 - u_0|T = 1, X) = E(u_1 - u_0|X) = 0. \]  

In a regression of \( Y \) on \( T \) and \( X \), the coefficient before \( T \) is equivalent to the mean difference between the outcomes of participants and nonparticipants conditional on \( X \) (Heckman 1997), which is

\[ E[Y(1)|T = 1, X] - E[Y(0)|T = 0, X] = \text{ATT} + E(u_0|T = 1, X) - E(u_0|T = 0, X) \]
\[ = \text{ATE} + E[u_0|T = 1, X] - E[u_0|T = 0, X] + E(u_1 - u_0|T = 1, X) \]  

The first and second equalities show that the common source of bias for estimating ATT and ATE this way is the correlation between \( T \) and \( u_0 \), which results in \( E(u_0|T = 1, X) \neq E(u_0|T = 0, X) \). This is the case if a unit chooses to participate in the program based systematically on some unobservables that exist in the absence of the program. An additional source of bias arises for estimating ATE when \( T \) is dependent on \( u_1 - u_0 \) or \( E(u_1 - u_0|T = 1, X) \neq 0 \), which implies that the expectations of the idiosyncratic gains are different between participants and nonparticipants.
Assume we have a sample of $N$ units, indexed by $i$. The observed outcome, treatment status, and vector of covariates for each unit are denoted by $Y_i$, $D_i$, and $X_i$ respectively. $D_i = 1$ if the unit is treated, and $D_i = 0$ otherwise. $u_i$ is used to denote the unobservables. In cases where we observe the same unit for more than one period, a subscript $t$ is added to indicate the time dimension. Under a common two-period scenario, $t = 0$ for the preprogram period and $t = 1$ for the postprogram period.

**Before–After Comparison**

In one case, all observed units are treated at the same time point by the program and data are available for both preprogram and postprogram periods. We use $D_t$ to denote the treatment status at time $t$ with $D_0 = 0$ and $D_1 = 1$. Note that $D_t$ does not vary across units. The before–after comparison can be modeled by the regression equation

$$
Y_{it} = \alpha + \theta D_t + X_i \beta + \mu_i + u_{it} \quad (i = 1,...,N, \ t = 0, 1),
$$

where $\theta$ is the average treatment effect (assuming universal treatment) of the program on the outcome, $\mu_i$ is a unit-specific dummy, often called fixed effect, that summarizes all time-invariant characteristics and $\alpha$ is a constant. The key assumption for consistently (unbiased) estimating $\theta$ is that conditional on $X$ and $\mu$, $D$ is mean independent with $u$, or $E(u|D, X, \mu) = E(u|X, \mu) = 0$.

**With–Without Comparison**

There are generally two approaches for with–without comparison: the regression approach and matching approach. The regression approach is modeled as

$$
Y_i = \alpha + \theta D_i + X_i \beta + u_i \quad (i = 1,...,N).
$$

The key assumption for consistently estimating $\theta$, referred to as the unconfoundedness assumption, is $E(u|D, X) = E(u|X) = 0$.

Matching compares treated units with the matched untreated units. One of the most popular matching methods is propensity score matching. The propensity score, denoted by $p_i$, where $p_i = p(D_i = 1|X_i)$, measures the probability of being treated given the covariates. It can be estimated
with a nonlinear binary model such as logit or probit. The propensity score matching estimator of $\theta$, the average treatment effect, is

$$\theta = \frac{1}{N} \sum (D_i - 1) Y_i / (1 - p_i), \ (i = 1, \ldots, N). \quad (A2.3)$$

In addition to the unconfoundedness assumption, the overlap assumption needs to hold for matching, which states that it is possible to observe both the treated and untreated units for all possible values of $X$. Formally, it is $0 < p(D_i = 1 | X_i) < 1$ for all $i$.

**Difference-in-Differences Method**

The sample contains both before- and after-program observations for the treated as well as untreated units. Let $\mu_i$ and $\lambda_i$ be the fixed individual effect and time effect, respectively. $D_i = 1$ if $i$ is treated and $t = 1$, and $D_i = 0$ otherwise. The approach can be formulated as the equation

$$Y_{it} = \alpha + \theta D_{it} + X_{it} \beta + \mu_i + \lambda_t + u_{it} \ (i = 1, \ldots, N, t = 0, 1), \quad (A2.4)$$

where $\theta$ is consistently estimated if $E(u|D, X, \mu, \lambda) = I(u|X, \mu, \lambda) = 0$.

**Regression Discontinuity Design**

The method is plausible when there is a forcing variable $s$, which (largely) determines treatment at a cutoff point $s_0$. That is,

$$D_i = 1 \text{ if } s_i \geq s_0 \text{ and } D_i = 0 \text{ if } s_i < s_0 \ (\text{sharp design}), \text{ and}$$

$$Pr(D_i = 1 | s_i \geq s_0) \gg Pr(D_i = 1 | s_i < s_0) \ (\text{fuzzy design}).$$

In this case, the program impact can be identified and nonparametrically estimated. Readers are referred to Imbens and Lemieux (2008) for a detailed discussion of RDD estimation and the necessary assumptions for the estimators to be valid.

**Instrumental Variables**

Assume instrumental variable $z$ is available such that

$$E(u|z, X) = E(u|X) = 0, \text{ and}$$

$$Pr(D = 1|z, X) \neq Pr(= 1|X).$$

A consistent estimator of $\theta$ can be obtained in a two-step procedure. First, estimate

$$Pr(D_i = 1 | z_i, X_i) = \hat{G}(z_i, X_i),$$

and then plug $G(z, X)$ into the outcome equation and estimate

$$Y_i = \alpha + \theta \hat{G}(z_i, X_i) + X \beta + u_i \ (i = 1, \ldots, N). \quad (A2.5)$$

For a more formal discussion of the econometrics of impact evaluation, see Imbens and Wooldridge (2008).
References


A Review of Recent Developments in Impact Evaluation

Impact evaluation aims to answer whether and to what extent a development intervention has delivered its intended effects, thus enabling evidence-based policy making. The desire for more hard evidence of the effectiveness of development interventions has fueled a growing interest in rigorous impact evaluation in the international development community.

This report discusses the fundamental challenge of impact evaluation, which is to credibly attribute the impact, if any, to the intervention concerned. It then discusses the merits and limitations of various impact evaluation methods. It also presents a survey of recent applications of impact evaluation, focusing on the typical evaluation problems looked at, methods used, and key findings. The report includes six case studies and outlines practical steps in implementing an impact evaluation.

About the Asian Development Bank

ADB’s vision is an Asia and Pacific region free of poverty. Its mission is to help its developing member countries reduce poverty and improve the quality of life of their people. Despite the region’s many successes, it remains home to two-thirds of the world’s poor: 1.8 billion people who live on less than $2 a day, with 903 million struggling on less than $1.25 a day. ADB is committed to reducing poverty through inclusive economic growth, environmentally sustainable growth, and regional integration.

Based in Manila, ADB is owned by 67 members, including 48 from the region. Its main instruments for helping its developing member countries are policy dialogue, loans, equity investments, guarantees, grants, and technical assistance.