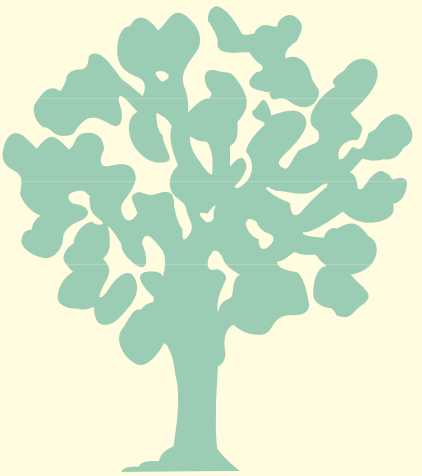
WORKING PAPER

Rough Guide to Impact Evaluation of Environmental and Development Programs

No. 40-09

Subhrendu K. Pattanayak



SANDEE

South Asian Network for Development and Environmental Economics

March 2009

Rough Guide to Impact Evaluation of Environmental and Development Programs

SUBHRENDU K. PATTANAYAK

Associate Professor, Duke University and Fellow, SANDEE

March 2009

South Asian Network for Development and Environmental Economics (SANDEE)
PO Box 8975, EPC 1056
Kathmandu, Nepal

SANDEE Working Paper No. 40-09

Published by the South Asian Network for Development and Environmental Economics (SANDEE)
PO Box 8975, EPC 1056 Kathmandu, Nepal.
Telephone: 977-1-552 8761, 552 6391 Fax: 977-1-553 6786

SANDEE research reports are the output of research projects supported by the South Asian Network for Development and Environmental Economics. The reports have been peer reviewed and edited. A summary of the findings of SANDEE reports are also available as SANDEE Policy Briefs.

National Library of Nepal Catalogue Service:

Subhrendu K. Pattanayak
Rough Guide to Impact Evaluation of Environmental and Development Programs
(SANDEE Working Papers, ISSN 1893-1891; 2009- WP 40)

ISBN: 978 - 9937 - 8093 - 6 - 8

Key words:

1. Program Evaluation
2. Randomized Social Experiments
3. Propensity Score Matching
4. Natural and Quasi Experiments
5. Instrumental Variables
6. Panel DID
7. Theory Based-evaluations

The South Asian Network for Development and Environmental Economics

The South Asian Network for Development and Environmental Economics (SANDEE) is a regional network that brings together analysts from different countries in South Asia to address environment-development problems. SANDEE's activities include research support, training, and information dissemination. SANDEE is supported by contributions from international donors and its members. Please see **www.sandeeonline.org** for further information about SANDEE.

This work was carried out with the aid of grants from several sponsors, including the International Development Research Center, Canada, the Swedish International Development Cooperation Agency and the Norwegian Agency for Development Cooperation. The views expressed in this publication are those of the author and do not represent the views of the South Asian Network for Development and Environmental Economics or its sponsors.

Comments to be sent to Subhrendu K. Pattanayak, Associate Professor, Duke University, Email: subhrendu.pattanayak@duke.edu

TABLE OF CONTENTS

1. INTRODUCTION 1

2. DEALING WITH COUNTERFACTUALS AND CONFOUNDERS 2

2.1 THE CURIOUS CASE OF COUNTERFACTUALS AND COUNFOUNDERS 2

2.2 PROGRAM THEORY AND 'LOGIC MODELS' 3

2.3 STEPS IN AN IMPACT EVALUATION 5

3. DESIGN AND ANALYSIS OF IMPACT EVALUATIONS 7

3.1 RANDOMIZED EXPERIMENTS 8

3.2 PROPENSITY SCORE MATCHING (PSM) 9

3.3 NATURAL EXPERIMENTS AND INSTRUMENTAL VARIABLES (IVM) 11

3.3.1 BASIC LOGIC OF IV 11

3.3.2 THE TROUBLE WITH INSTRUMENTS! 11

3.4 PANEL DATA AND 'DIFFERENCE-IN-DIFFERENCE' ESTIMATORS 13

3.5 SUMMARY 14

4. LEARNING BY DOING: A SOUTH ASIAN CASE STUDY 14

5. CONCLUSIONS AND BROADER CONSIDERATIONS 16

5.1 MEASURING EQUITY AND HETEROGENEITY 17

5.2 MEASURING SUSTAINABILITY 17

5.3 A CALL FOR RIGOROUS IMPACT EVALUATIONS 18

6. ACKNOWLEDGEMENTS 18

7. REFERENCES 19

APPENDIX I. GLOSSARY OF TERMS 26

APPENDIX 2. STATA LOG FILE FOR PSM DATA EXERCISE. 27

LIST OF TABLES

Table 1:	Examples of indicators for a sanitation promotion program.	23

LIST OF FIGURES

Figure 1:	Generic Model of a Program	24
Figure 2a:	Evaluation of decentralized management of forests in Nepal (Edmonds, 2002)	24
Figure 2b:	Evaluation of social mobilization for sanitation in India (Pattanayak et al., forthcoming)	24
Figure 2c:	Evaluation of payments to landowners in Costa Rica for provision of ecosystem services (Pfaff et al., 2008)	25
Figure 3:	Difference-in-Difference Estimator	25

Abstract

Abstract: Evaluation of programs, either before they are designed or after they are implemented, are increasingly viewed as a critical for learning and improving accountability of public policies. Unfortunately, resource and environmental economists in developing countries have little or no training or guidance on how to conduct such evaluations using sound and rigorous empirical methods. This paper is a "rough guide" for evaluation of programs, projects and policies in the environment and development arena. First, we provide a general overview of the what, how, and why of program evaluation, with particular emphasis on the role of control groups, pre-&-post measurement, and covariate data to define counterfactual scenarios (including formal definition of all terms). Second, we review detailed examples of the four main methods for evaluation – randomized experiments, natural experiments, matching methods, and panel-based DID estimators – with a description of the pros and cons of each method. Third, the guide provides one detailed case study of a SANDEE funded project from South Asia that allows the reader to learn by using data and econometric code to practice and appreciate some of the challenges of impact evaluations. Finally, we conclude by placing the econometric evaluations within the broader context – how can we move beyond estimation of average treatment effects; what do we do under time, resource and data constraints; and when and where should we rely on theory-based evaluations.

Keywords: Program Evaluation, Randomized Social Experiments, Propensity Score Matching, Natural and Quasi Experiments, Instrumental Variables, Panel DID, Theory Based-evaluations.

Rough Guide to Impact Evaluation of Environmental and Development Programs

Subhrendu K. Pattanayak

1. Introduction

Evaluation of programs, either before they are designed or after they are implemented, are increasingly viewed as a critical for learning and improving accountability of public policies. Duflo et al. (2007) argue for an expanded role for rigorous impact evaluation of development projects, programs, and policies and the development community increasingly recognizes the need for evidence on effectiveness (e.g., Managing for Development Results, and the Development Impact Evaluation Initiative). Unfortunately, resource and environmental economists have little or no training or guidance on how to conduct such evaluations using sound and rigorous empirical methods. Ferraro and Pattanayak (2006), Frondel and Schmidt (2005) and Pattanayak et al. (forthcoming-a) discuss this gap in the context of biodiversity conservation, environmental protection, and water and sanitation. For example, see a series of calls (at the rate of one or more publication per year) for evidence-based conservation (Kleiman et al., 2000; Pullin and Knight., 2001, Sutherland et al., 2004; Saterson et al., 2004; and Stem et al., 2005). While there are many texts explaining the how and why of impact evaluations (WB-OED 2004; Kusek and Rist 2004; Prenusshi et al. 2000), these ideas have not been mapped into the 'environment and development' (ENVDEV) context and terminology. This paper is a "rough guide" for evaluation of programs, projects and policies in the environment and development arena.

The fields of medicine & public health, education, crime, and poverty reduction have taken tremendous strides by relying on 'evidence-based practice'. It is important to rigorously evaluate public programs and policies for many reasons (CDC, 1999). First, demonstrating that a particular program yields environment, health, and poverty reduction benefits can be used to set priorities and mobilize support to expand or modify programs. Second, even though specific programs show great promise, they might not work under all field conditions. Program outcomes can be highly variable, with some interventions and programs in some settings showing little impact. Good evaluations can identify why this might happen and what adjustments can be made to correct it. Third, if small-scale projects are to make an important contribution to government policy, they need to be expanded or "scaled up". It is important to know what aspects of these projects lead to greater or less success. Finally, impact evaluations are often conducted in close collaboration with regional governments, local NGOs, and partners in academia and community organizations. Thus, good evaluations affect stakeholders by stimulating dialogue and raise awareness about effective policy tools. Yet, to date we have few or no rigorous scientific impact evaluations showing that environment and development policies are effective in delivering many of the desired outcomes.

To understand why we make this claim, consider two criteria that are commonly used. First, a rigorous scientific evaluation must utilize some mix of control groups, baselines, and covariates to establish the counterfactual scenario and permit the estimation of impacts. Second, the example should be from within the sector, producing evidence on environment and development outcomes. Previous reviews of the literature suggests that there are few completed studies and on-going

evaluations that have evaluated the impacts of environment and development policies (Ferraro and Pattanayak, 2006; Pattanayak et al., forthcoming). In the concluding section, we speculate on possible reasons that there have been so few rigorous evaluations in this sub-field.

The guide is organized as follows. First, we provide a general overview of the what, how, and why of program evaluation, with particular emphasis on the role of control groups, pre-&-post measurement, and covariate data to define counterfactual scenarios (including formal definition of all terms). Second, we review detailed examples of the three main methods for evaluation - randomized experiments, natural experiments, and matching methods - with a clear description of the pros and cons of each method. Third, we place econometric evaluations within the broader context - how can we move beyond estimation of average treatment effects; what do we do under time, resource and data constraints; when and where should we rely on theory-based evaluations; and how can simulation-based evaluations complement and/or substitute for econometric evaluations. Finally, the guide provides detailed case studies that will allow the reader to use data and econometric code to practice the three main methods discussed in the guidelines. The guide also provides a detailed reference list and recommended readings.

2. Dealing with counterfactuals and confounders

Evaluation means different things to different people and often we confuse a monitoring study and or process assessment with an impact evaluation - a much narrower, and often more rigorous type of study. Baker (2000) defines a comprehensive evaluation as one that includes monitoring, process evaluation, economic evaluation, and impact evaluation. She also summarizes the different purposes each type of evaluation. Monitoring is used to assess whether a program is being implemented as was planned. Process evaluation assesses how the program operates and focuses on problems in service delivery. Economic evaluation (cost-benefit or cost-effectiveness) assesses program costs and benefits. Impact evaluation, the focus of this paper, measures the impacts of the program on individuals, households, and individuals, and determines whether the program caused these impacts (Baker 2000; WB-OED 2004). This section draws extensively from Poulos et al. (2006) and uses examples from the water and sanitation sector (WSS). The overall framework readily applies to environment and development programs without loss of generality. Appendix I provides a glossary of impact evaluation terminology to guide the first-time reader.

2.1 The curious case of counterfactuals and counfounders

The fact that impact evaluation is concerned with the results that are caused by the program distinguishes it from program and process evaluations. Process evaluation is focused on how well the program is operating, and relies mainly on qualitative analyses to identify bottlenecks in program implementation or service distribution, deviations from the project plan, user satisfaction, as well as conflicts or transaction costs. As described, these are vital complements to an impact evaluation in gaining a thorough understanding of what works and why.

To measure final impact, an impact evaluation must determine what would have happened in the absence of the program - this is known as the counterfactual. This is complicated by the fact that the counterfactual is naturally unobservable - we can never know what change would have occurred in program participants (treatment group) if the program was not implemented. For example, consider a decentralization policy to involve communities in the management of their forests. We often assume that people would have collected forest products and or degraded the

forests at the same rate without local forest management as before the policy change. This can be misleading because there could be a general trend towards lower use of forest products, for example, because of improving incomes. Impact evaluations must therefore rely on control or comparison groups, as well as a number of statistical and econometric techniques to estimate this counterfactual. These tools help the analyst control for factors or events (called confounders) that are correlated with the outcomes but are not caused by the project.

Confounders are correlated with the intervention and may affect the outcomes, masking the intervention's effect. Examples of confounders include socio-cultural behaviors (e.g. collective action to improve access to community resources sources), institutional factors (e.g., other programs promoted by other government departments, non-governmental, or donor organizations), bio-physical characteristics (e.g., water table, climate, soils and geology), and general trends (e.g., macro-economy, prices, inflation). Failing to account for the influence of confounders introduces a source of bias-omitted variable bias. The identification and measurement of the counterfactual, comparison, or control and the careful consideration of confounders is the primary distinguishing feature between process evaluations and impact evaluations.

Treatment groups are usually different than untreated groups for political or economic reasons. For example, communities targeted by a WSS intervention may be worse off than other communities, because the intervention was targeted to poor communities with inadequate WSS conditions. In addition to observable differences, there are often unobservable differences between the treated and untreated groups. These differences can exist in their ability to participate in the program and their motivation to implement the program. When groups are not comparable, the difference between the groups can be attributed to two sources: pre-existing differences and the impact of the program. The former can cause selection bias in the measurement of program impact.

The key focus of impact evaluation is its ability to measure the causes of outcomes. In general, impact evaluation use either randomized trials or, when interventions are not randomly assigned, appropriate quasi-experimental methods. An experimental design, in theory, eliminates all sources of selection bias. However, experimental designs are often not feasible for political or logistical reasons and these designs have rarely been used for ENVDEV (see next section for "information" treatments). Thus, we rely on quasi-experimental designs that employ a battery of purposive sampling and econometric estimation techniques to control for selection on observables and unobservables (Shadish et al., 2002).

2.2 Program theory and 'Logicmodels'

Figure 1 illustrates a generic program. The first two boxes (Program Resources and Program Activities) represent the planned work for the program and the other three boxes (Program Outputs, Program Outcomes, and Program Impacts) show the intended results. In the simplest terms, an impact evaluation is concerned with the two right-most boxes in this diagram. However, an impact evaluation must also understand what happens on the left side, i.e. the resources, activities and outputs. For example, it is possible that a program's intended impacts are not met because the program activities were not implemented as planned. Also, knowing the inputs will help determine whether the intended final outcomes and impacts of a program are feasible. Monitoring studies and process evaluations can help provide information on inputs when they are combined with impact evaluation. Figures 2a - 2c provide concrete examples that map the

concept to specific case studies; two of these are discussed in detail in Section 3. These include evaluations of decentralized forest management in Nepal (Fig. 2a), social mobilization for sanitation in India (Fig 2b), and payments to landowners in Costa Rica to provide ecosystem services (2c). For the remainder of this section, we focus on the second example to flesh out the framework, while emphasizing that this logic model can be easily generalized to any ENVDEV policy or project.

An impact evaluation measures a program's progress by tracking indicators of the program's inputs and results. An indicator is any direct and unambiguous measure of progress toward the intended goals of a program. Prenusshi et al. (2000) define a good indicator as: (1) relevant to program objectives (e.g., incidence of diarrheal diseases); (2) varying across areas, groups, over time, and sensitive to changes in policies, programs, and institutions (e.g., medical and time costs imposed by diarrheal diseases); (3) not easily diverted or manipulated (e.g., presence of an individual household latrine - IHL); and (4) able to be tracked (e.g., observing how many households have an IHL next to their residence). During the evaluation process, it is important to monitor program resources and activities by tracking what can be called "intermediate" indicators provide information on activities and outputs and thus provide valuable information on whether a program was implemented successfully (Prenusshi et al. 2000). Given the length of the causal chain from resources to impacts, it is critical to know (e.g., by using process evaluations) if things fell apart before the program had an opportunity to have any final impacts. Outcomes and impacts are tracked through "final indicators".

Program resources are the available financial, human, social, and institutional capital for the program. These include funds from donors, government, and matched funds from communities. It includes the human capital (from the government, nongovernmental organizations, and communities) that contributes to operating and maintaining the system and partnerships that facilitate system operations. Finally, formal institutions (laws, regulations, economy) and informal institutions (custom, norms, social capital) that support or constrain the system are also program resources.

Program activities are the actions and processes carried out by the program to bring about the intended goals. An impact evaluation should focus on inputs that are explicitly allocated to the program, and not the broader conditions that are necessary for program success, such as hydrological or governance conditions. While these conditions may be essential to the program, since they are not allocated explicitly to the program, we call them "external factors" and discuss them below. Intermediate indicators of the social mobilization program for sanitation can include the number of FTEs assigned to implement the program, funds disbursed, completion of community action plans, and the types of mobilization activities carried out in each community (see Pattanayak et al. [forthcoming-b] for specific details).

Program outputs, program outcomes, and program impacts constitute the program results. A program output is any direct product of program activities that program providers have direct control over. Outputs include the type of products and levels of service delivered to participants, such as the installation of sanitation infrastructure (e.g., the number and type of IHLs) or hygiene training. These outputs in turn can also be viewed as "interventions" that generate impacts as discussed next.

Program outcomes are the changes in behaviors, knowledge, and actions among participants as a result of the program. Continuing our sanitation mobilization example (figure 2b), this would

include use of and satisfaction with IHLs, and changes in community water quality. A program can have both short term outcomes (attainable in 1-3 years) and long term outcomes (attainable in 4-6 years). While program outcomes reflect access to, use, or satisfaction with the intervention, program impacts are the fundamental change experienced by program beneficiaries as a result of the program. The fundamental changes are improvements in beneficiaries' well-being measured (e.g., health or income). In the sanitation example, the impacts would include children's development (e.g., anthropometrics, school attendance, non-cognitive skills) and household wellbeing (e.g., time and money savings).

A good evaluation should track any "external" indicators, which measure factors exogenous to the program that could influence the program's ability to achieve its intended results (Prenusshi et al. 2000). As discussed previously, ignoring these exogenous factors can introduce confounding bias into the evaluation. For example, in the sanitation mobilization campaign, governments may target poor communities or villages with concentrations of socially marginal populations (scheduled castes and tribes). Poor or socially marginal communities may have worse WSS conditions, health, and incomes at baseline. Failing to account for the differences between the treated and untreated groups in these external factors, which are correlated with both the intervention and the impacts of interest, would lead to downwardly biased estimates of impacts. The experimental evaluation described by Pattanayak et al. (forthcoming-b) minimizes such biases through randomized assignment of the mobilization campaign and double-difference strategy that uses the panel data set.

The logic model clarifies the relationship among all program elements and helps to keep the focus on outcomes and impacts, achievement of which is the raison-d'être of a development program. While there is relatively greater tracking of activities and outputs, it is not enough to just measure outcomes and impacts. If we do not see an expected result, we will not be able to explain the lack of an impact without tracking the related activities and outputs. Measurement all along the causal chain is critical.

2.3 Steps in an impact evaluation

Baker (2000) describes key steps in designing and implementing impact evaluations. The first step is to determine whether or not to carry out an evaluation because impact evaluation can be complex and expensive. Baker (2000) and Ferraro and Pattanayak (2006) suggest a number of criteria to determine whether an impact evaluation is required. One is to compare the likely costs and benefits of the impact evaluation. The benefits of an evaluation are likely to be higher when the project is innovative (e.g., testing new technology, new delivery mechanisms, or new organizational structure); is scalable, replicable, and likely to be expanded to other settings; involves substantial resource allocations; and has well-defined interventions. On the other hand, the benefits of impact evaluation are likely to be low when a program's outcomes are non-generalizable because of the population, institutions, systems, program, or environmental setting is peculiar. If the project is experimental and likely to be revised over time, it will difficult to conduct an impact evaluation.

Also, impact evaluations are more likely to be beneficial when the outcomes are a matter of debate, and given the paucity of rigorous impact evaluation in the environment and development, there are a number of unresolved issues. First, we are aware of only a countably few evaluations that demonstrate the impacts of programs for biodiversity conservation, natural resource management, control of pollution, or water and sanitation, particularly in terms of socio-economic

outcomes such as incomes (or consumption), human capital (e.g., education and health), or gender and ethnic inclusion. Second, the impact of some of the main ENVDEV policies such as private sector participation in water and sanitation delivery or decentralized management of environment and natural resources, need additional study. Finally, related to the previous point, we have little or no evidence on the relative impacts of community-level interventions compared to more individual or household level programs and policies, e.g., source water treatment vs. point of use water treatment (see for example Clasen et al. 2006). For example, Mansuri and Rao (2004) review studies that evaluate community-driven economic development programs in a number of sectors, including WSS, labor, agriculture, and others. They contend that effectiveness and sustainability of community-based and community-driven projects depends on a number of factors, including the heterogeneity of the community's population, the level of social capital, the role of external agents, and the design of the program itself. They do not identify any studies that establish a causal relationship between any outcome and the participatory elements of the project. Another criterion for determining whether to do an impact evaluation is the presence of strong political and financial support. Without the support of the leadership in the sector, programs, and communities, analysts are unlikely to gain entrée to the information needed for a rigorous impact evaluation and supporting monitoring, program, and process studies.

The second step is to clarify the objectives of the evaluation. This should be done early in the program during identification and preparation. Clear objectives reveal the core issues that will be the focus of the evaluation and inform the selection of measures, data sources, and evaluation design. Environmental programs typically have multiple objectives, some relating to results in the sector (e.g., increase per capita water consumption, increase in species protected) and others related to results outside the sector, including health outcomes and income.

The third and fourth steps, which are interrelated and may be completed interactively, are to explore data availability and design the evaluation. Qualitative and/or quantitative measures of intermediate and final indicators (i.e., program resources, activities, outputs, outcomes, and impacts) are necessary for the impact evaluation and these may be acquired through the collection of secondary or primary data. These indicators help define the context in which the program is implemented. Table 1 provides examples of each type of indicator for a hypothetical public water supply project. The evaluation design, whether experimental or quasi-experimental, is determined by the project objectives and the available data (see Section 4).

After forming an evaluation team, the next steps are to design data collection procedures, the remaining steps, which are accomplished during project implementation, are data collection, data analysis, synthesis and reporting results to stakeholders, and incorporating findings into design of projects, programs and policies.

Bamberger et al. (2004; 2006) have developed a modified impact evaluation framework specifically for those cases in which analysts must conduct impact evaluations under budget, time, and data constraints. These may occur when the evaluation is begun well after the program design and implementation or when baseline data is unavailable because of budget or political realities. Their framework offers a structured approach to addressing the constraints in order to ensure the highest quality evaluation possible.

3. Design and Analysis of Impact Evaluations

Data will need to be collected on intermediate and final indicators (as well as external factors) to carry out the impact evaluation. The most scientifically rigorous impact evaluations are those that use a combination of data sources to triangulate and verify information. While many of the indicators require quantitative data, qualitative data also plays an important role in assessing factors such as participant satisfaction or implementation of the program. Indicators should be SMART: Specific: - what is intended to be measured; Measurable - clear and unambiguous; Attributable - to the program; Realistic - reasonable cost and frequency of data collection; and Targeted - about the relevant or target population. Moreover, these should not easily diverted or manipulated (Prenusshi et al. 2000).

Whether primary or secondary data are used, sample size and power calculations should be performed to determine the sample sizes necessary to detect differences under a range of conditions at a given significance level and power. The significance level (typically called the size or a) represents the probability of a type I error, which is the rejection of the null hypothesis when it is true (i.e., the false positive rate). The power of a test is represented by 1- ß, where ß is the probability of a type II error, which is the failure to reject the null hypothesis when it is false (i.e., the false negative rate). Here we are talking about the significance and power of a statistically estimated impact. Power calculations can be adjusted for response rates, expected variation in the sample, and any expected attrition (e.g., less people signed up for the program than anticipated). Practically speaking, sample sizes will often be governed by the budget (and in some cases time) available to conduct evaluations. Budget constraints implies tradeoffs such that the evaluator has to yield by accepting less precise estimates of impacts or answer a narrower question. Sample sizes can be reduced by relaxing/lowering the significance level of the test (i.e., significance testing at the 10 percent level requires a smaller sample than significance testing at the 5 percent level). Alternatively, sample sizes can be reduced by reducing the power of the test (See Bamberger et al. 2004). However, it is important to recognize that the numbers produced by these formulas are mere guides and not foolproof standards for any study because of uncertainties related to the design effects and expected impact sizes.

We do not discuss data collection in this guide and point the interested reader to an excellent guide for data collection for impact evaluation by Wassenich and Munoz (2007). The authors discuss data types, data sources, and data quality from a practical and feasibility perspective. Data quality is an underestimated challenge for impact evaluation - with too much attention and fascination with econometrics, and too little attention to the quality and reliability of the data itself (Heckman et al., 1999). Data collection for impact evaluations can be efficient compared to general data collection because the nature of the evaluation exercise forces the evaluator to precisely define the outcome and intervention, as well as critical confounding factors.
In this section, we review detailed examples of the three main methods for evaluation - randomized experiments, natural experiments, and matching methods. We summarize the pros and cons of each method with the help of one or more case study. To develop a deeper understanding of these approaches, the interested reader is encouraged to review the specific papers referenced in this section.

3.1 Randomized experiments

Randomized experiments are one of the best known designs for impact evaluations. Experiments identify the effect of a program or policy by randomly distributing alternative causes over experimental conditions. If implemented appropriately, this design ensures that potential confounders are balanced across program (intervention) and control units and therefore any differences in the outcomes between the two can be attributed to the program. True random experimental designs are nonexistent in the field of environment and natural resources policy, but their absence has nothing to do with characteristics of the field. If development economics can use field experiments to test the effects of micro-credit on household welfare and child de-worming on school performance, there is no reason that environmental practitioners cannot implement randomized experiments to examine the effectiveness of interventions like payments for environmental services or conservation education.

Although randomized experiments are possible, they can be challenging to implement and evaluate in many circumstances. For example, it would be hard to randomly regulate some communities and not others and, for political reasons, it can be difficult to create a program that provides benefits randomly rather than to areas that most need them. Where such obstacles are overcome by simplifying or compromising the program, we might end up inducing various kinds of randomization bias such that the participants no longer represent an average person, having adjusted their behavior to benefit from the project. Heckman and Smith (1995) also worry about substitution bias arising because control groups gain access to 'substitute' resources and activities that are available from overall programs. In some cases, these are offered by program administrators to compensate the control group for being excluded. Another problem often encountered is the scale of a program - a program implemented at the national level cannot be randomized. However, if a national program is rolled out region-by-region, it may be possible to select the regions randomly and use the other regions as controls. Finally, the renewed emphasis on bottom-up, community-driven programs limits the use of randomized design, as a random assignment would contradict the overall intention of a community-led program.

Information campaigns, on the other hand, are amenable to randomized designs. Information, education and communication (IEC) is widely viewed as an alternative to economic or moral incentives for promoting and stimulating behavior change. One of the attractive features of IEC type instruments is the ability to test its effectiveness through randomized assignment at relatively low costs. Luby et al. (2004), Jalan and Somanathan (2008) and Pattanayak et al. (forthcoming-b) all use randomized designs to evaluate how information influences water and sanitation behaviors. Because water and sanitation programs and policies have public good characteristics (i.e., they provide benefits that are spread over or targeted to defined groups of individuals), group randomization is more often the relevant approach in this context. That is, treatments are often implemented at the group (community or region) level. In some cases, the impacts are also measured at the group level. As discussed next, the analysis of the data from a randomized experiment involves a simple and relatively straightforward comparison of outcomes in exposed and control groups.

Pattanayak et al. (forthcoming-b) describe a randomized evaluation of an information, education, and communication (IEC) campaign in rural Orissa (Bhadrak district) and its impacts on use of individual household latrines (IHLs). Twenty villages are randomly assigned to the IEC campaign, while 20 control observationally similar villages did not receive the treatment. The pre-intervention survey showed that households in the treatment and control groups are similar in socio-demographic

characteristics, assets and income, education, health literacy and hygiene behaviors. Through simple luck of draw, treatment villages had more IHLs than control villages in the baseline. The authors account for these pre-existing differences using a DID estimator that measures changes (see section below). The authors find that the IEC campaign substantially increased IHL uptake by about 30%. This increase in IHL, in turn, led to improvements in child health, measured as decrease in child diarrhea and increases in arm circumference.

3.2 Propensity score matching (PSM)

The method of matching is a non-experimental approach for pairing treatment units (areas or households) with "very similar" control units using statistical modeling. Perhaps the best known and most used matching method is propensity score matching (Ravallion, 2007; Todd, 2007). Propensity scores represent the probability of participation in a program, typically estimated from a statistical model of participation as a function of ecological, socio-economic, institutional and geographic factors. PSM controls for observable selection bias by ensuring that treatment and control groups are comparable in all aspects except that they have not received the intervention. This method calculates the probability (i.e., propensity score) that participants and non-participants would participate in the intervention based on a set of observed characteristics, identified by the researcher. The statistical model (e.g., logit or probit) allows calculation of a score for everyone, and then participants and non-participants are matched according to this propensity score. As constructed, the probability score is a weighted index of several variables that allows the evaluator to avoid the curse of dimensionality and not have to non-parametrically match on several variables. In the water and sanitation sector, it has been used in evaluations of social investment funds in Bolivia (Newman et al. 2002), and in evaluations of privatization in Argentina (Galiani et al. 2005). It has also been used to evaluate the effects of forest disturbance on forest amenities (Pattanayak, 2004), decentralized management on forest cover (Somanathan et al., 2009), the Endangered Species Act on species recovery (Ferraro et al., 2007) and park establishment on deforestation (Andam et al., 2008).

Pipeline matching: a second type of matching controls for observable selection bias by identifying program participants (individuals or communities) who are in the 'pipeline'. In this case the control group is constructed from communities or households that have applied to the program and are eligible, but have not yet been selected to receive the intervention. Pipeline comparison, in theory, ensures that that the treatment and control (pipeline) groups are comparable in all aspects except that they have not received the intervention. Pipeline matching has been used in evaluations of SIFs in Armenia (Chase 2002).

To better understand PSM, consider Jalan and Ravallion's (2003) study on whether household access to piped water reduces child diarrhea. PSM helps overcome any potential selection bias arising, for example, from households choosing piped water only if their children have high diarrhea rates. The authors use household survey data from rural India to build a binary-choice model of who has piped water at home as a function of approximately 90 variables, including village-level characteristics such as agricultural modernization and educational and social infrastructure and household characteristics such as demographics, education, religion, ethnicity, assets, and housing conditions. Predicted probabilities from this first stage regression are used to pair households who have taps with those who do not have taps but have the same predicted probability of having taps. A comparison of means suggests that households without pipes experience 21% higher prevalence, and 29% greater duration of diarrhea compared to "observationally" equal households having piped water.

PSM is not without its potentially problematic assumptions and implementation challenges. First, PSM requires large amounts of data both on the universe of variables that could potentially confound the relationship between outcome and intervention, and on large numbers of observations to maximize efficiency of the search. Second, related to the previous point we can never be entirely sure that we have actually included all relevant covariates in the first stage of the matching model and effectively satisfied the conditional independence assumption (CIA). However, at least the method forces the analyst to be explicit about the 'observable' characteristics that are used to compare intervention and control units, and which might explain self-selection (by participants) or targeting (by program administrators). In practice, often the first stage results are not reported or discussed (e.g., as why the covariates may explain targeting or volunteering), not unlike the practice with the first stage of an IV estimation process (see next). Furthermore, PSM is non-parametric: we do not make any functional form assumptions regarding the average differences in the outcome. Although the first stage involves specification choices - e.g., functional form (logit, probit, scobit, etc.) - empirical analyses tend to find impact estimates that are reasonably robust to different functional forms. Third, the 'stable unit treatment value assumption' is perhaps the most contentious feature of the methodology: essentially the method assumes there are no spillovers across the units of analysis and treatment is homogenous. While there are some ways to examine heterogeneity, in general the maintained assumptions of treatment heterogeneity, no spillovers and conditional independence are no different from other evaluation methods, including randomized trials and other quasi-experimental approaches. We will return the treatment heterogeneity question in Section 5.

3.3 Natural experiments and instrumental variables (IVM)

Closest in spirit to randomization is the method of natural experiments, which refer to situations where nature (or chance) creates "treatment" and "control" units. For example, to test policy effectiveness, one may have data from regions exposed to weather events such as hurricanes, floods, fires or landslides, which create natural barriers that differentially protect or expose forests to social pressures. Areas on either side of these natural barriers provide comparable sites for evaluations. Nature, rather than people, selects the units on the basis of chance and therefore eliminates selection bias. In conservation biology, islands are the best known examples of natural experiments and have been important in testing ecological theories. In the water sector, geology or bio-physical characteristics might result in a natural experiment.

In the absence of random events, researchers can apply instrumental variable methods (IVM). "Archimedes said, Give me the place to stand, and a lever long enough, and I will move the Earth. Economists have their own powerful lever: the instrumental variable estimator. The instrumental variable estimator can avoid the bias that ordinary least squares suffers when an explanatory variable in a regression is correlated with the regression's disturbance term. But, like Archimedes' lever, instrumental variable estimation requires both a valid instrument on which to stand and an instrument that isn't too short (or too weak)." [Murray, 2006]

Edmonds' (2002) study of forest user groups and fuelwood extraction in Nepal provides an example of the instrumental variable method (IVM). The question he is interested in answering is whether formation of community forest user in Nepal groups reduces forest degradation. The evaluation problem is simply that communities facing high levels of forest degradation might (a) chose to form forest user groups to address the degradation problem, or (b) be targeted by government agencies to build user groups. If that is the case, then simple comparison of forest conditions in communities with and without FUGs may find small impacts (the estimate is biased

downwards). As discussed previously, instrumental variables (IV) provide reliable independent source of data on FUG formation; i.e., independent of the forest degradation concerns. Edmunds suggest that the presence of other extension programs, range posts, and agricultural extension agencies can serve as a proxy for village accessibility, which he contends is a major exogenous (independent) determinant of FUG formation. Analysis using the IV strategy confirms that FUG reduces forest degradation by 15%.

3.3.1 Basic Logic of IV

IV are exogenous data that explain some variation in endogenous D, which is the dummy variable for program. An IV model is typically implemented as a two-stage least squares (2SLS), where in stage 1 we estimate 'endogenous' D as a function of IV and predict Dhat. In stage 2, we model the outcome as a function of Dhat. This implies some loss in efficiency, but researchers are often willing to do that to reduce bias. Consider two recent examples in environment and development. Duflo and Pande (2007) are interested in the question of whether dams impact poverty and worry about endogeneity because dams may be placed in poor (or rich) places. They contend that the gradient of river basin is a good IV for dam placement. Pitt et al. (2005) are interested in understanding how exposure to indoor air pollution impacts respiratory infections, a relationship that might also be subject to endogeneity (reverse causality) concerns. They suggest that the household members' relation to household head serves as an IV for endogenous exposure.
Strength, Validity & Exclusion Restrictions

An exclusion restriction is a central concept to the strategy (Ravallion, 2007). Recall, IVs explain D, but is (are) not correlated with Y independent of their impact via D. Thus, IVs are correctly 'excluded' from the model of Y. Putting this in Murray's terminology, an instrument is strong as long as the IVs are strongly correlated with D. Similarly, an instrument is valid as long as IVs are uncorrelated with error. The critical point is that good instruments are both strong and valid.
Returning to the Edmunds example introduced above, he has strong instruments: he confirms the strong correlation between the IVs and the likelihood of a community having an FUG through the jointly significance of IVs and model goodness-of-fit measure. His instruments are also valid: following standard over-identification tests, he shows that the instruments are individually and jointly insignificant in a regression of residuals from second stage of regression on instruments. What this suggests is that if instruments are not correlated with error (when you include other controls), then instruments impacting fuelwood use only through FUG. He also checks for validity by regressing fuelwood use in control communities on the on same instruments and finding that coefficients and individually and jointly insignificant. This second check confirms the identifying assumption that the IVs have no direct impact on fuelwood collection because the control communities are by definition not affected by FUG.

3.3.2 The trouble with instruments!

'Natural experiments' don't happen often or conveniently. In general, good instrumental variables are hard to find. Using IVM typically requires a mix of clear theoretical intuition, good quality secondary data and a solid grasp of field conditions. The main tasks for the evaluator using IVM are to: (a) avoid invalid IV, and (b) cope with weak IV (Murray, 2006). Validity is an untested (and some cases un-testable) assumption that must be subject to intuitive, empirical and theoretical scrutiny. As discussed above, Edmunds tests for over-identification restrictions and checks for correlation with Y. Furthermore, we must ensure no omitted variables and check robustness by seeing if alternative instruments give the same result. Weakness must be nipped by pre-testing

candidate instruments, using robust procedures (e.g., Fuller, conditional LR), and ensuring the use of valid instruments.

Ultimately, the justification and implementation of IVM lies outside quantitative data employed in the evaluation, with evaluators employing their understanding of theory, common sense, and knowledge (qualitative) of particular program & policy. For starters, consider five sources of instruments discussed in Ravallion (2007).

1. Imperfect compliance (in experiments): Pattanayak et al. (forthcoming-b) are interested in measuring the impacts of toilet use on diarrhea. Their experimental information and education campaign convinced some households to install and use toilets. They confirm that the campaign did not impact other key risk modifying behaviors such as hand washing and treating and handling drinking water. Thus, the randomized assignment is an appropriate instrument: it is strongly correlated with the problematic ("potentially endogenous") instrument, but uncorrelated with the outcome.
2. Geography: Attanasio and Vera-Hernandez (2004) are interested in the impacts of the nutrition program on child care and food intake, and worry about endogeneity of program uptake. They suggest that the location of house, measured in terms of distance to the community center, is exogenous because survey respondents who have moved recently never identified the desire to move closer to a community center as one of the reasons for choosing their location (even though this was one of the options). Tellingly, they also note that if their results were in fact driven by endogeneity of their IV then they would find (spurious) effects on variables that should not have any effect on a priori grounds, such as child birth weight.
3. Political economy: political characteristics of program assignment can help identification, as in an evaluation of how social funds impact schooling in Peru (Paxson and Schady, 2002). The authors argue that the geographic allocation of social fund spending would be used in part to "buy back" voters that had switched against the government in the last election. Their first stage regressions confirm that an indicator of if the communities had switched explained social fund spending. This switching was uncorrelated with schooling outcomes.
4. Exogenous shocks: An example of this approach can be found in Chay and Greenstone (2003) who examined the impact on infant mortality in the US of air pollution (total suspended particulates) and found a sizeable positive impact. Their identification rests on the fact that the recession in 1980-82 caused a sharp decline in TSP in some counties and no decline in TSP in other counties. They also show that these sharp differences in TSP concentrations have no corresponding discontinuities for the usual suspects such as income, education, and racial composition that might confound this relationship.
5. Discontinuities in policy design: Chay and Greenstone (2003) provide an example of the impacts of air quality on housing value. For identification they exploited the fact that US EPA declares a county to be in non-attainment and applies stringent regulations if TSP concentrations exceed a threshold level. Thus, there is a sharp discontinuity around the threshold and the non-attainment status is an IV for potentially endogenous air quality. They also show that the non-attainment status is uncorrelated with other potential confounders such as income, unemployment, demographics, population, and housing quality. The authors find sizeable capitalization in housing markets of improvements in air quality.

3.4 Panel data and 'difference-in-difference' estimators

The most common impact evaluation utilizes panel data sets with two features: (a) the sample observations can be sorted into treatment (program) and control groups, and (b) at least one wave of the panel from before and one from after the 'treatment'. With this basic 2x2 design (before-after & control-treatment), program impacts are estimated by calculating the difference in outcomes between treatment and control groups after program implementation minus the difference in outcomes between treatment and control groups prior to the implementation. Often, we refer to this double difference or this simple 'comparison-in-means' as the difference-in-difference (DID) estimator. Figure 3 illustrates the DID estimator.

$$DID = \{E[Y_{1t} \mid X] - E[Y_{1c} \mid X]\} - \{E[Y_{0t} \mid X] - E[Y_{0c} \mid X]\}$$

Equation above describes a DID estimator, where Y is the outcome with subscript 1 and 0 for post-treatment and pre-treatment levels, and subscripts t and c for intervention and control unit outcomes. E is the expectations operator suggesting that this is the expected treatment effect across all treatment units. The simplest estimator is not conditional on any other variable (X). However, as discussed below, it is possible to estimate conditional DIDs by including relevant covariates regressors.

DID estimators are often implemented in a regression framework by regressing the outcome on three dummy variables - the treatment category, the treatment period, and an interaction variable for the treatment category and period. The estimated coefficient on the interaction term measures the pre-to-post change in the outcome for the treatment (program) households relative to pre-to-post change in the outcome for the control households. Todd (2007) and Ravallion (2007) present further econometric details on this evaluation technique.

Given the thinness of the literature on impacts of natural resource policies and program, we do not have a rich repository of case studies on 'panel based DID'. Frankenberg et al. (2005) present perhaps the best example where they report on the impacts of the S.E. Asian forest fires and the associated haze on health. They utilize two waves of the Indonesian Family Life Surveys from 16 Indonesian provinces and more than 300 enumeration units in 1993 (before the fires) and 1997 (after the fires). Approximately 25% of the IFLS sample that comprised households in Sumatra, Kalimantan, Lampung and West Nusa Tenggara were exposed to the haze in the summer of 1997, based on aerosol indices exceeding a 1.5 for three consecutive summer days (a common air pollution standard). Health was measured using three indicators: ability to carry heavy loads, respiratory illnesses (e.g., coughs), and general health symptoms. The authors find that the fire and haze negatively impacted health, particularly of the elderly.

The strength of the panel-based-DID estimator comes from its intuitive appeal and simplicity: we derive an estimate of the impact by comparing the treatment and control groups using the post-treatment data (second difference), after we use the pre-treatment data to equate treatment and control groups (first difference).

However, there are at least two disadvantages that relate to the very simplicity of such a panel based impact assessment. First, constructing panel data sets can be expensive, time consuming, and logistically challenging particularly because we need to collect baseline and follow-up data that straddle the implementation of a program. This seems like a formidable obstacle in a field that has shown reluctance to evaluate programs using careful data collection (e.g., in program

and non-program areas). Second, the design assumes that the potential selection bias (i.e., due to administrative targeting or volunteering) is linear and time invariant such that it can be subtracted off in the first differencing. If this is not the case, we need some intuition for what other factors might be causing differential trends in the treatment and control groups. As Frankenberg et al. (2005) show, further threats of confounding can be addressed by including empirical surrogates of such factors as additional covariates (observables). If covariates are introduced, some of the attractive properties of the simple DID estimator are lost because now we need to worry about specification challenges of introducing additional X variables. Additionally, fixed effects can be estimated to account for time-invariant unobservables at the household, community or other relevant unit level. Finally, robustness can be further evaluated by estimating a semi-parametric DID model, which essentially uses inverse probability weights that are function of covariates that we might be concerned about (Abadie, 2005).

3.4 Summary

Typically, experimental approaches are internally valid for answering a narrow question precisely. This precision and focus comes at the price - experimental analyses are limited to the extent that the results can be generalized or scaled up, thus reducing their policy value. The evaluator must be extra vigilant in the design stage to ensure the integrity of the experiment and can usually work with "smaller" samples (compared to quasi-experimental studies). In contrast, the quasi-experimental methods are internally weaker, but potentially stronger in terms of some extrapolation and insights for scaling up. To do this, they typically require much larger data sets, additional structure (and or assumptions) and therefore more work in the analysis stage. However, the extent to which quasi-experimental methods mimic experimental approaches by narrowing the focus and estimating effects around a cut-off for example, they acquire the limitations of experiments. Deaton (2009) presents a recent critical review of experimental and natural experimental evaluations.

Within the quasi-experimental methods - IV, PSM and DID - no approach necessarily dominates all others (Ravallion, 2007; Todd, 2007). Each is appropriate for a particular context, policy and most specifically the type and quality of data that is available. Under some settings, a combination of methods can help further reduce selection and endogeneity biases such as in the case of PSM & DID. Furthermore, the simple versions (e.g., non-stratified) of all three methods as well as randomized experiments are equally limited in addressing heterogeneity of treatment and impacts, and potential spillovers. We return to this issue in the conclusion.

4. Learning by doing: A South Asian Case Study

From a learning perspective, there is no substitute to personal involvement in the design and implementation of an evaluation, especially at the design stage. Unfortunately, this is a luxury for most readers and users of this document. Thus, in this short section, we work with data from a SANDEE funded project to develop some understanding of the analysis for a pilot program and (through that experience) to identify some of the fundamental practical challenges of impact evaluations.

Jalan and Somanathan (2008) have evaluated an information program in which a random sub-sample of 500 households (from a total of 1000 households) in Gurgaon (outside New Delhi) are informed about the fecal contamination of their drinking water. The pre-intervention survey showed that households in the treatment and control groups are similar in education, health

literacy and hygiene behaviors. Moreover, households that received a negative test result were not significantly different in water treatment behavior than control households. Seven weeks after the provision of the information, informed households were 11% more likely to begin some form of water purification compared to uninformed households. Mean purification expenditure in the sample increased by 10 percent among households that were told that their drinking water had tested positive. By way of comparison, an additional year of schooling of the most educated person in the household raises the probability of (initial) purification by 4.4 percentage points while a move from one wealth quartile to the next raises it by 15 percentage points.

Professor Jalan has provided this data as a STATA file strictly for learning purposes, which is available as an Appendix to this guide. The first task is to familiarize ourselves with the data set particularly by identifying the variables that correspond with the outcome, treatment, and key covariates. All of this will make more sense if we have read the Jalan and Somanathan paper carefully. As part of the familiarization exercise, we should calculate the means and standard deviations of the key outcome variables in the treatment and control sub-samples. Next, we should compare the means and test if these are statistically significant (based on the size of the t-statistic).

Our next goal is to replicate the main results of the Jalan & Somanathan (2008) paper based on the randomized design. Before we test for the averting behavior change induced by the treatment (information provision), let us confirm the fundamental tenets of randomization - potential confounding is eliminated / reduced by the act of randomization. We can do this by calculating means and standard deviations for several key socio-economic, health and hygiene covariates for each sub-sample (treatment and control) and checking if they are statistically significantly different. Note, the errors are correlated because several observations are included from each enumeration block. Thus, the t-statistics should account for this variance inflation, e.g., through the cluster option in STATA. Additionally, to ensure population representativeness of the sample, use the probability weights provided by the authors.

Now we can turn to the main 'averting behavior' change result. We can do this in a regression context by including an interaction variable for the study condition (e.g., dummy for treatment status) and for the treatment period. The estimated coefficient on the interaction term measures the pre-to-post change in the outcome for the affected households relative to pre-to-post change in the outcome for the unaffected households. Because all the outcomes considered by the authors are binary, we should use logit or probit regressions for these analyses.

Although this study does not use matching estimators, we can use this data set to 'practice' how we might implement a matching estimator - a very common impact evaluation tool. The statistical code for running PSM is readily downloadable from the internet (see Appendix II for instructions for downloading the code onto your computer). Consider just the baseline (pre-treatment) data and the question - does home purification cause the drinking water to be less polluted? With just this pre-treatment cross-sectional data we cannot simply compare the percentage of households with polluted water in the home purifiers group and see if they are statistically and significantly smaller than the percentage of households with polluted water in the non-purifiers group. Besides the obvious endogeneity implied by the main Jalan & Somanathan result, these households may be different across many other characteristics. PSM can address these concerns by comparing purifiers with observationally equivalent non-purifiers based on their propensity to purify. We can proceed through the following steps (e.g., the psmatch2.ado command will implement all four steps):

1) estimate "propensity score model" by estimating a logit (or probit) model of purification as a function of various household characteristics (e.g., education, gender of household head, family size, number of small children, wealth, health literacy)

2) store probability score as a separate variable and match on propensity score

3) verify that the matched groups - i.e., purifying households for whom we found matched observationally equivalent non-purifying households - are similar across all relevant household characteristics (differences are statistically insignificantly or the unmatched differences are significantly reduced)

4) calculate the mean differences in matched pairs on the outcome of interest - drinking water contamination; this is the average treatment effect.

It is helpful to compare the ATE result from this exercise to (a) simple difference in means of drinking water contamination in the purifiers and non-purifiers, and (b) a simple regression of water contamination on purification dummy and various household covariates. Appendix II provides the log file of working through steps 1 - 4 in STATA. The reader should attempt to replicate the results to test their understanding of the PSM method.

In conclusion, after re-reading Jalan and Somanathan (2008) we should develop a short write-up that summarizes the impact evaluation plan. The narrative should include the following:

1. Justification: explain why this is an important ENVDEV policy question

2. Challenges: how would you define the counterfactual? What are the potential confounders?

3. Mechanics: please describe and justify several features of this evaluation:

a. Summarize the results framework (including identification of indicators)

b. Justify the study design

c. Describe key data collection activities

d. Present your main analysis plan

4. Policy dissemination: briefly identify the stakeholders, present a communication plan, and describe your expected outcomes

5. Conclusions and broader considerations

In this section, we discuss some broader issues for impact evaluation. Some of the more recent criticisms of the typical program evaluation work from both academics (e.g., Heckman, 2001; Deaton, 2009) and leading practitioners (e.g., Ravallion, 2007) are that most applications do not consider participant or program heterogeneity in any substantive way. These critics contend that conditional mean impacts (e.g., ATE = average treatment effect) provide a very limited set of lessons for policy design and implementation. Another criticism is that ATE is a black box - you don't fully understand why you did see that particular impact. Finally, it is unclear if the methods offer any insights on scalability of the policy, particularly if there are general equilibrium effects and or spillovers of other kinds. Ravallion (2008) presents a thorough discussion of practical nuts and bolts of impact evaluation for policy support.

5.1 Measuring equity and heterogeneity

A deeper understanding of the program is a critical first step towards "opening the black box of the conditional mean impact" by recognizing heterogeneity in program delivery, acceptance and impacts (Ravallion, 2007). For example, if our target population had chosen and received different environmental and development packages at the baseline, we should be studying multiple interventions (Lechner, 2002). Quantile treatment effects represent a semi-parametric way to examine treatment heterogeneity (Gamper-Rabindran et al., 2008)

Consider the water sector, for example. Equitable access to safe water and adequate sanitation for all members of society, regardless of age or sex and regardless of social, cultural, religious, or ethnic status is a key element of WSS policies. If programs are systematically excluding sections of the population such as the poor, the policy recommendation is for some form of poverty targeting. For example, one of the biggest concerns surrounding privatization of water delivery is that the private operator will exclude the poor because of operational decisions surrounding tariff and network design. Some would argue that ensuring equity in service provision is necessary for social sustainability and for preventing outcomes such as the cancellation of privatized water service delivery in Bolivia, allegedly because of riots and civic protests around the equity issues. From an analytical perspective, equity can be examined by looking at the distribution of access across subpopulations including socioeconomic strata and ethnic minorities. Sub-group estimation of program impacts in routine in program evaluation: Galiani et al. (2005) and Jalan and Ravallion (2003) report estimates of water supply impacts on child health by income and literacy categories. Alternatively, the equity of access can be assessed by comparing prices across geographic areas, water/sanitation sources, and subpopulations. If markets are thin or missing such that prices are either not available or are not determined by the market, shadow pricing should be used to estimate the full price of services to beneficiaries.

5.2 Measuring sustainability

The sustainability of capital investments and the outcomes they generate are critical dimensions of the effectiveness of environmental programs and policies. In many ways, environmental and resource policies generate outcomes in the long term. There are countless anecdotes and field stories the regarding enthusiastic starts that subsequently were not sustained and completely abandoned within 2-5 years of starting up. However, measuring the durability of impacts is a challenge since there are often no mechanisms to support continuing monitoring after the project cycle has ended.

Carvalho and White's (2004) suggestion to use program theory to focus on sustainability risks and test this through sensitivity analysis (a series of what ifs) allows us to work within the project cycle to deal with sustainability. They suggests that the risks to outcomes be identified using a program theory that articulates how the program causes the intended or observed outcomes. Program elements on the causal path are candidates for risks. For example, the soundness of the technical design of a water system is a necessary condition for sustainable outcomes. Thus, the risk of design flaws should be assessed as part of the sustainability analysis.

For example, the evaluations of social fund investments use this approach to examine maintenance issues (see Rawlings et al. [2004] for a summary). They use surveys of more than 1,200 schools, health centers, and water and sewerage facilities to measure inputs such as staff, materials, and

maintenance. They find improvements in the quality of design and operations, staffing, administrative capacity, maintenance, cost recovery, and community training in project communities contrasted with similar indicators in non-project (matched control) communities. To the extent possible, many environmental impact evaluations should adopt this approach in order to build a body of evidence about sustainable projects.

5.3 A call for rigorous impact evaluations

We have presented guidelines for evaluations of environment and development programs that focus on establishing the counterfactual - what would have happened without the program - and estimate program impacts by using a mix of controls, baselines and covariate measurement. Unfortunately, these suggestions are based on a very thin scientifically-validated literature of published and on-going evaluations in the sector (even after using somewhat liberal interpretations of what is within the sector). Thus, we have also borrowed extensively from the general development program evaluation literature in writing these guidelines because many issues related to design, measurement, analysis and interpretation are transferable across sectors. Nevertheless, it would be better to make our case around a wealth of empirical case studies within the ENVDEV field. Therefore, we conclude with a call for building a rich repository of empirical examples.
We are certainly not calling for every ENVDEV program and project to be evaluated with an experimental or quasi-experimental design. Instead, our appeal is based on the observation that there are simply too few applications to ENVDEV, thereby impeding our ability to evaluate policies and more critically learn how to evaluate policies. These will not be trivial investments because good evaluations require systematic and comprehensive data collection on outcomes and covariates from treatment and control units before and after the intervention. Ultimately, such investments (or a series of examples) can help fill the large gap in our knowledge about the effectiveness of environmental and development policies.

6. Acknowledgements

The idea for this guide came naturally from Priya Shyamsundar and the SANDEE secretariat, based on a series of training workshops conducted by the author for the US Forest Service International Programs at Belem (Brazil) in June 2006, SANDEE at Kathmandu (Nepal) on July, 2007 and January, 2008 at Bangkok (Thailand). Vic Adamowicz provided helpful comments on an earlier draft. The concepts and methods presented in this guide draws liberally on research conducted by the author in close collaboration with Rodrigo Arriagada, Dave Butry, Paul Ferraro, Sumeet Patil, Christy Poulos, Erin Sills, and Jui-Chen Yang. The author alone is responsible for all errors in this paper.

References

Abadie, A., 2005. Semi-parametric difference-in-difference estimators. *Review of Economic Studies*. 72: 1–19

Andam, KS, PJ Ferraro, A Pfaff, GA Sanchez-Azofeifa, and J Robalino. 2008. Measuring the Effectiveness of Protected Area Networks in Reducing Deforestation. *Proceedings of the National Academy of Sciences* 105(42): 16089-16094.

Angrist, Joshua and Victor Lavy, 1999, "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics* 114(2): 533-575.

Attanasio, O. and AM Vera-Hernandez. 2004. "Medium and Long Run Effects of Nutrition and Child Care: Evaluation of a Community Nursery Programme in Rural Colombia," Working Paper EWP04/06, Centre for the Evaluation of Development Policies, Institute of Fiscal Studies London.
Baker, J.L. 2000. Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners. The World Bank, Washington.

Bamberger, M., J. Rugh, M. Church, and L. Fort. 2004. Shoestring Evaluation: Designing Impact Evaluations under Budget, Time, and Data Constraints. *American Journal of Evaluation* 25 (1): 5-37.

Bamberger, M. 2006. *Conducting Quality Impact Evaluations under Budget, Time, and Data Constraints.* World Bank Independent Evaluation Group Evaluation Capacity Development Publication.

Carvalho, S. and H. White. 2004. Theory-Based Evaluation: The Case of Social Funds. *American Journal of Evaluation* 25 (2): 141-160.

CDC (Centers for Disease Control and Prevention), 1999. Framework for program evaluation in public health. *Morbidity and Mortality Weekly Report*. 48 (RR-11): 1-40.

Chase, R.S. 2002. Supporting Communities in Transition: The Impact of the Armenian Social Investment Fund. *The World Bank Economic Review* 16 (2): 219-240.

Chay, K. and M. Greenstone. 2005. Does Air Quality Matter? Evidence from the Housing Market. *Journal of Political Economy* 118(3): 1121-1167.

Chay, K. and M. Greenstone. 2003. The Impact of Air Pollution on Infant Mortality: Evidence from Geographic Variation in Pollution Shocks Induced by a Recession. *Quarterly Journal of Economics* 113(2): 376-424.

Jalan, J. and M. Ravallion. 2003. Does Piped Water Reduce Diarrhea for Children in Rural India? *Journal of Econometrics* Vol. 112(1): 153-173.

Jalan, J. & Somanathan, E. 2008. The Importance of Being Informed: Experimental Evidence on Demand for Environmental Quality. *Journal of Development Economics*. 87 (1): 14-28.

Kleiman, D.G., R.P. Reading, BJ Miller, TW Clark, J.M. Scott, J. Robinson, R. Wallace, R.J. Cabin, and F. Felleman. 2000. Improving the evaluation of conservation programs. *Conservation Biology* 14(2) 356-365.

Kusek, J.Z. and R.C. Rist. 2004. *Ten Steps to a Results-Based Monitoring and Evaluation System*. The World Bank, Washington, D.C.

Lecher, M. 2002 Program Heterogeneity and Propensity Score Matching. *Review of Economics and Statistics* **84**(2), 205-220.

Luby, S., M. Agboatwalla, J. Painter, A. Altaf, W. Billhimer, R. Hoekstra. 2004. Effect of Intensive Handwashing Promotion on Childhood Diarrhea in High-Risk Communities in Pakistan: A Randomized Controlled Trial. *Journal of the American Medical Association* 291(21): 2547-2554.

Mansuri, G., and V. Rao. 2004. "Community-Based and –Driven Development: A Critical Review." *The World Bank Research Observer* 19(1):1–39

Murray, M. P. (2006), "Avoiding Invalid Instruments and Coping with Weak Instruments," *Journal of Economic Perspectives* 20.4: 111–132.

Newman, J., M. Pradhan, L.B. Rawlings, G. Ridder, R. Coa, and J.L. Evia. 2002. An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Investment Fund. *The World Bank Economic Review* 16 (2): 241-274.

Pattanayak, SK., C. Poulos, J-C. Yang, S.R.Patil and K.J. Wendland. Forthcoming-a. Of Taps and Toilets: Quasi-experimental protocols for evaluating community-demand driven projects. *Journal of Water and Health*.

Pattanayak, S.K., J-C. Yang, K. L. Dickinson, C. Poulos, S.R. Patil, R. Mallick, J. Blitstein and P. Praharaj. Forthcoming-b. Shame or subsidy revisited: Cluster randomized evaluation of social mobilization for sanitation in Orissa, India. *Bulletin of the World Health Organization*.

Pattanayak, S.K. 2004. Forest amenities and aesthetics: An econometric evaluation using North Carolina FIA data. Unpublished Manuscript. RTI International, Research Triangle Park, NC.

Paxson, C. and NR Schady, 2002, **"**The Allocation and Impact of Social Funds: Spending on School Infrastructure in Peru," *World Bank Economic Review* 16: 297-319.

Pfaff, A., J. Robilano, and G.A. Sanchez-Azofeifa. 2008. Payments for Environmental Services: Empirical analysis for Costa Rica. Working Paper SAN08-05. Sanford Institute of Public Policy, Duke University. 27 pages.

Pitt, M.M., M.R. Rosenzweig and M. Nazmul Hassan. 2005. "Sharing the Burden of Disease: Gender, the Household Division of Labor and the Health Effects of Indoor Air Pollution," CID Working Paper No. 119, March 2005, Center for International Development, Harvard University, MA, USA.

Poulos, C., S.K. Pattanayak, and K. Jones. Guidelines for Impact Evaluations in the Water and Sanitation Sector. Doing Impact Evaluations. No. 4. World Bank, Washington, D.C. December 2006.

Pradhan, M. and L.B. Rawlings. 2002. The Impact and Targeting of Social Infrastructure Investments: Lessons from the Nicaraguan Social Fund. *The World Bank Economic Review* 16 (2): 275-295.

Prennushi, G., G. Rubio, and K. Subbarao. 2000. Monitoring and Evaluation. Chapter 3 in *A Sourcebook for Poverty reduction Strategies* Volume 1, pages 105-130. Washington, D.C.: World Bank.

Pullin AS, Knight TM (2001) Effectiveness in conservation practice: pointers from medicine and public health. Conservation Biology 15(1):50-54.

Ravallion, M. 2008. Evaluation in the practice of development. Policy Research Working Paper 4547. World Bank.

Ravallion, M. 2007. Evaluating Anti-Poverty Programs. In T. P. Schultz and J. Strauss (eds.), *Handbook of Development Economics*. 4: 3787-3846.

Rawlings, L.B., L. Sherburne-Benz, and J.V. Domelen. 2004. *Evaluating Social Funds: A Cross-Country Analysis of Community Investments*. The World Bank, Washington, D.C.

Saterson KA, Christensen NL, Jackson RB, Kramer RA, Pimm SL, et al. (2004) Effectiveness in conservation practice: Pointers from medicine and public health. *Conservation Biology* 18:597–599.

Shadish, W.R., Cook, T.D. & Campbell, D.T. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Infrerence*. Houghton Mifflin Company, Boston.

Somanathan, E., R. Prabhakar, and B.S. Mehta. 2009. Does Decentralization Work? Forest Conservation in the Himalayas. *Proceedings of the National Academy of Sciences*.

Stem C, Margoluis R, Salfasky N, Brown M (2005) Monitoring and evaluation in conservation: a review of trends and approaches. *Conservation Biology* 19(2):295-309.

Sutherland WJ, Pullin AS, Dolman PM, Knight TM (2004) The need for evidence-based conservation. *Trends in Ecology and Evolution* 19(6):305-308.

Todd, P.E. 2007. Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated. In T. P. Schultz and J. Strauss (eds.), *Handbook of Development Economics*. 4: 3847-3894.

Wassenich, P. and J. Munoz. 2007 Data for Impact Evaluation. Doing Impact Evaluation Series No. 6. The World Bank.

World Bank. 2005. "The Development IMpact Evaluation (DIME) Initiative: Coordinating Impact Evaluation Work At The World Bank", Draft Report, World Bank: Washington, DC.

The World Bank – Operations Evaluation Department. 2004. Monitoring and Evaluation: Some tools, methods, and approaches. The World Bank, Washington, D.C.

Hyde, W.F. and G. Kohlin (2000), 'Social forestry reconsidered', Silva Fennica 34 (3): 285-314.

Leach, G. (1992), 'The energy transition', Energy Policy 20 (2): 116-123.

Masera, O.R., B.D. Saatkamp and D.M. Kammen (2000), 'From linear switching to multiple cooking strategies: a critique and alternative to the energy ladder model', World Development, 28 (12): 2083-2103.

Narain, U., S. Gupta and K. van't Veld (2008), 'Poverty and resource dependence in rural India', Ecological Economics 66 (1): 161-176.

NSSO (2001), 'Energy Used by Indian Households: NSS 55th Round, July 1999 – June 2000', Report No. 464, National Sample Survey Organization, Ministry of Statistics and Programme Implementation, Government of India, New Delhi.

NSSO (2007), 'Household Consumption of Various Goods and Services in India, 2004/05, Vol 1: Major States and All-India, NSS 61st Round, July 2004 – June 2005', Report No. 509, National Sample Survey Organization, Ministry of Statistics and Programme Implementation, Government of India, New Delhi.

Pachauri, S. and L. Jiang (2008), 'The household energy transition in India and China', Energy Policy doi:10.1016/j.enpol.2008.06.016

Pyatt, G., C.N. Chen and J. Fei (1980), 'The distribution of income by factor components', Quarterly Journal of Economics, 1: 157-192.

Sadoulet, E. and A. de Janvry (1995), Quantitative Development Policy Analysis, The John Hopkins University Press, Baltimore.

TERI (2008), 'TERI Energy Data Directory & Yearbook (TEDDY), 2007', The Energy & Resources Institute, New Delhi.

UNDP/ESMAP (2003), 'Access of the poor to clean household fuels in India', Joint United Nations Development Programme (UNDP)/ World Bank Energy Sector Management Assistance Programme (ESMAP), South Asia Environment and Social Development Department, The World Bank, Washington D.C.

Veld, V.K., U. Narain, S. Gupta, N. Chopra and S. Singh (2006), 'India's firewood crisis re-examined', RFF DP 06-25, Resources For the Future, Washington D.C.

WHO (2002), The World Health Report 2002, World Health Organization, Geneva.

TABLES

Table 1: Examples of indicators for a sanitation promotion program.

Program Component	Indicator	Example
Resources & Activities	Intermediate	Funds and FTEs for mobilization campaign Focus groups, awareness building, visits by officials
Outputs	Intermediate	Number of individual latrines built
Outcomes	Final	Use of and satisfaction with latrines
Impact	Final	School attendance, incomes
Other	External	Cultural norms, bio-physical characteristics

Source: Modified from Prenusshi et al. (2000). Pattanayak et al. (forthcoming) provide details on the specific program.

FIGURES

Figure 1: Generic Model of a Program

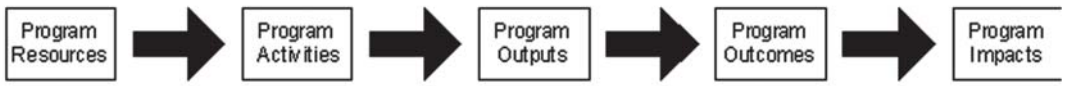


Figure 2a. Evaluation of decentralized management of forests in Nepal (Edmonds, 2002)

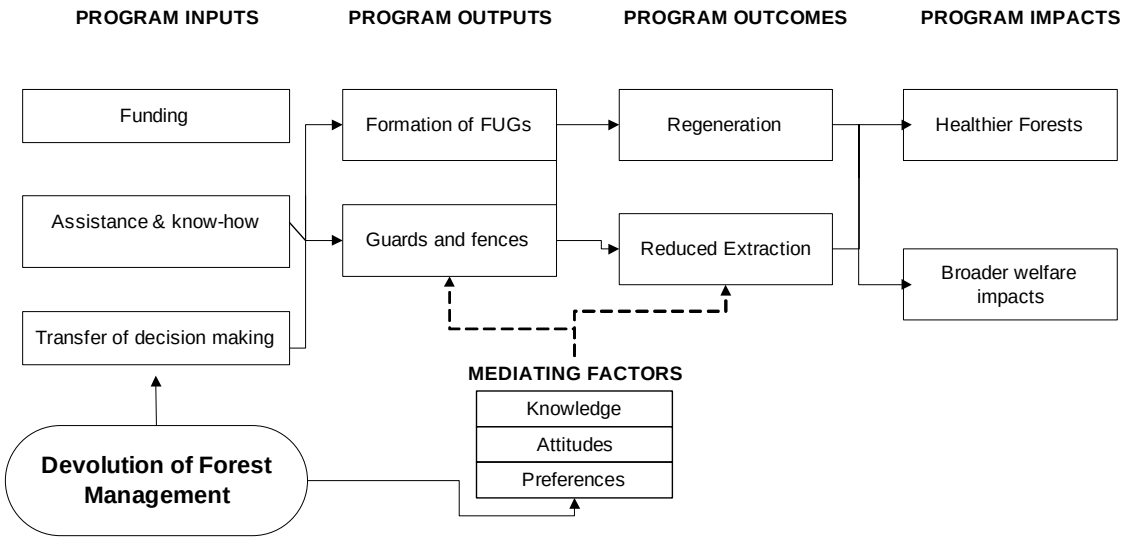


Figure 2b. Evaluation of social mobilization for sanitation in India (Pattanayak et al., forthcoming)

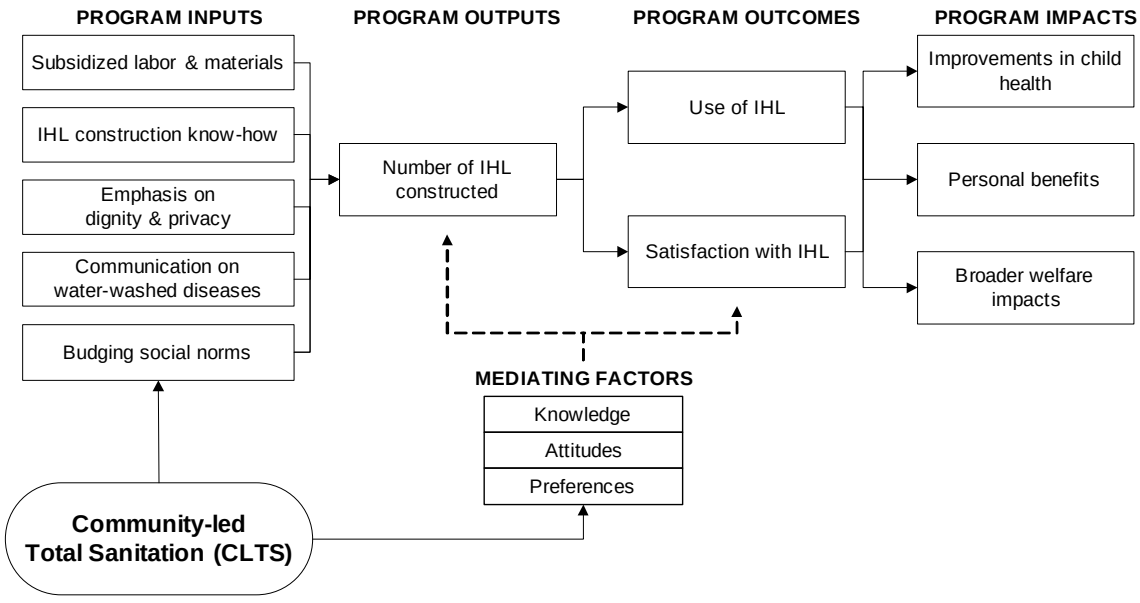


Figure 2c. Evaluation of payments to landowners in Costa Rica for provision of ecosystem services (Pfaff et al., 2008)

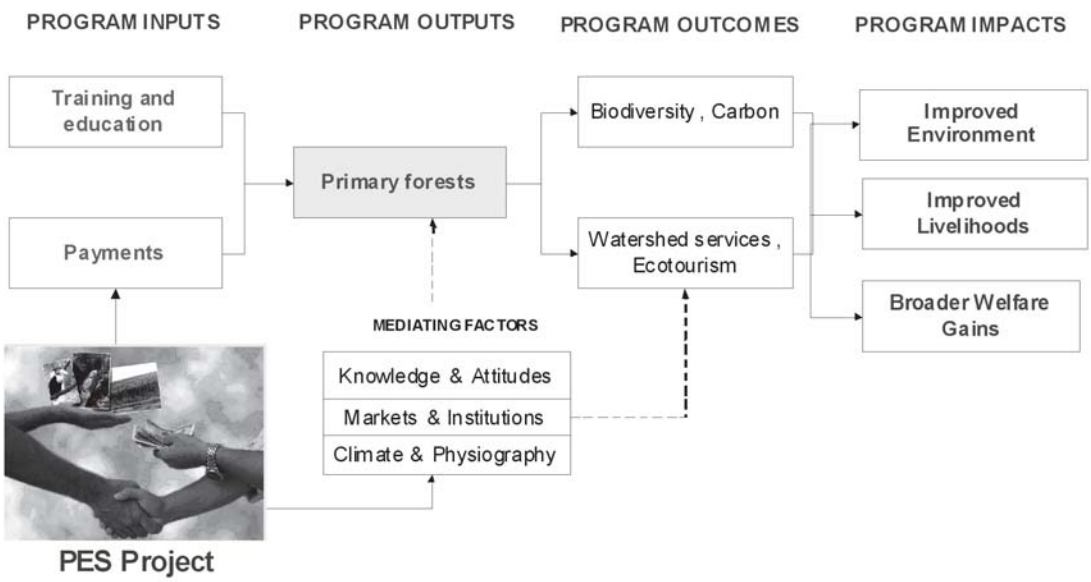
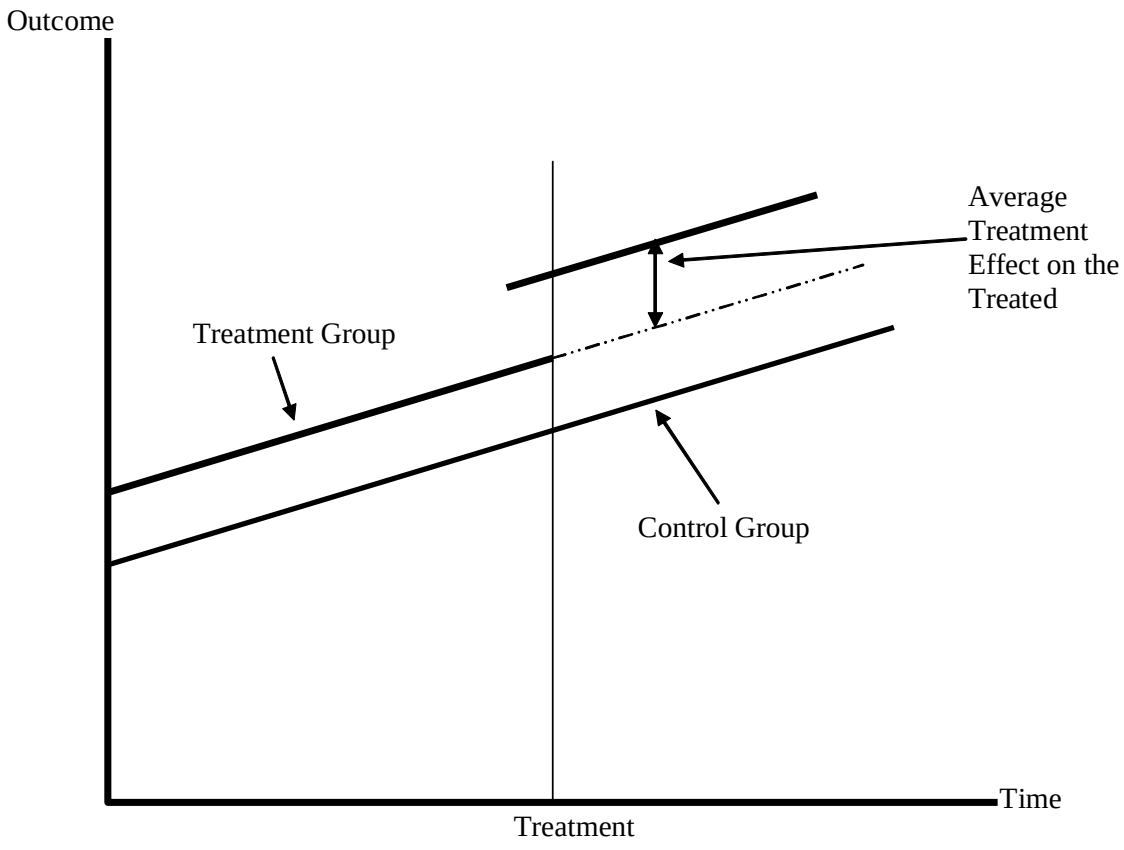


Figure 3. Difference-in-Difference Estimator.



Appendix I. Glossary of Terms

Activities: any intentional actions and processes carried out by the program, these are the components used to bring about the intended program goals

Beneficiaries: the intended recipients of the program

Counterfactual: what would have occurred (the impact) without a program or policy

Difference-in-difference: an estimator that compares impacts between control and treatment groups (first difference) before and after the intervention (second difference)

Difference in means: an estimator that compares impacts between control and treatment groups after an intervention

Experimental design: randomly assigns participants to control and treatment groups

Goals: the overall purpose of implementing the program

Impacts: the fundamental intended change occurring as a result of the program, impacts should be attainable in 7-10 years

Instrumental variables: a method that identifies exogenous variation in outcomes by using variables that determine participation in a program but would not affect outcomes as controls

Intervening factors: the external factors that interrupt the link between program outcomes and impacts

Mediating factors: the external factors that interrupt the link between program activities and output

Multivariate regression analysis: a method that uses multivariate regressions to control for observable differences in control and treatment groups

Outcomes: the specific changes in project participant's behavior, knowledge, and actions; short-term outcomes should be present in 1-3 years, long-term outcomes in 4-6 years

Outputs: the direct products of the program activities, includes types, levels, and targets of services to be delivered

Pipeline comparison: a type of matching that constructs the control group from households that have applied to the program and are eligible, but have not yet been selected to receive the intervention

Quasi-experimental design: uses a variety of statistical and econometric techniques to assign control and treatment groups

Propensity score matching: a method used in quasi-experimental design to controls for observable selection bias, calculates the probability that participants and non-participants would participate in the intervention based on a set of observable characteristics

Resources: the available human, financial, organizational, and community resources at the disposal of the program

Simulated counterfactual: an estimator that constructs a counterfactual using a theoretical model and information on the situation prior to the interventions.

Appendix 2. STATA log file for PSM data exercise.

With the data obtained by contacting the SANDEE secretariat or the author, you can practice using PSM code in STATA. Below, I insert some comments (preceded by *) in between each command (which is preceded by a .). To run this in STATA you must be able to use a pre-programmed .ado command. match.ado and psmatch2.ado are two common codes; I use psmatch2 for this illustration. If you do not have this .ado file in your STATA routines, please type in ssc install psmatch2 in your command window if you are connected to the internet. In this case, the psmatch2 routine (a) matches treatment and control households based on propensity scores, and (b) computes the average difference in outcomes (in our case drinking water quality) between the matched treatment and their control units (in our case households).

* We begin by obtaining descriptive statistics for the main 'treatment' variable – does the household purify its drinking water?

```
. tabstat drink_water, stats(N mean sd) by(flagpurify)
Summary for variables: drink_water
     by categories of: flagpurify (Purification dummy)

flagpurify |         N      mean        sd
-----------+------------------------------
         0 |       569   .659051  .4744454
         1 |       396  .5454545  .4985595
-----------+------------------------------
     Total |       965  .6124352   .487447
-------------------------------------------
```

* Next we check if the water quality is dirtier (it tested positive for microbial contamination) in households that do or don't purify their drinking water using a simple t-test. We find that it is about 11% more likely to be dirty in the non-purifying households.

```
. ttest drink_water, by (flagpurify)
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     569     .659051    .0198898    .4744454    .6199845    .6981175
       1 |     396    .5454545    .0250536    .4985595    .4961996    .5947095
---------+--------------------------------------------------------------------
combined |     965    .6124352    .0156915     .487447    .5816418    .6432286
---------+--------------------------------------------------------------------
    diff |            .1135964    .0317057                .0513762    .1758166
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   3.5828
Ho: diff = 0                                     degrees of freedom =      963

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9998         Pr(|T| > |t|) = 0.0004          Pr(T > t) = 0.0002
```

* We confirm the previous result using a simple logit regression, correcting the standard errors for clustering because of the way the sample was created.

```
. logit drink_water flagpurify, cluster (enum_block_no)
Logistic regression                                     Number of obs   =        965
                                                        Wald chi2(1)    =      10.83
                                                        Prob > chi2     =     0.0010
Log pseudolikelihood = -637.95388                       Pseudo R2       =     0.0098
                      (Std. Err. adjusted for 205 clusters in enum_block_no)
------------------------------------------------------------------------------
             |               Robust
 drink_water |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  flagpurify |  -.4767463   .1448859    -3.29   0.001    -.7607175   -.1927751
       _cons |   .6590679    .097242     6.78   0.000      .468477    .8496588
------------------------------------------------------------------------------
```

* We further test this difference by adding in other covariates, including (a) education of female head, (b) education of male head, (c) whether household has a child under 3, (d) an index of awareness of water quality issues, (e) a wealth index, (f) family size, (g) sex of household head, (h) age of household head.

```
. logit drink_water flagpurify maxfedu maxmedu u3 au wu no_members sex_head age_head,
cluster (enum_block_no)
Logistic regression                                     Number of obs   =        965
                                                        Wald chi2(9)    =      21.44
                                                        Prob > chi2     =     0.0108
Log pseudolikelihood = -633.14301                       Pseudo R2       =     0.0173
                      (Std. Err. adjusted for 205 clusters in enum_block_no)
------------------------------------------------------------------------------
             |               Robust
 drink_water |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  flagpurify |  -.5384147   .1664782    -3.23   0.001     -.864706   -.2121235
     maxfedu |   .0275263   .0181682     1.52   0.130    -.0080828    .0631354
     maxmedu |  -.0250417   .0246728    -1.01   0.310    -.0733996    .0233161
          u3 |   .0385366   .1751897     0.22   0.826     -.304829    .3819022
          au |   .2415559   .1343581     1.80   0.072    -.0217811     .504893
          wu |   .0747283   .1588533     0.47   0.638    -.2366185     .386075
  no_members |   .0266442   .0302043     0.88   0.378    -.0325551    .0858435
    sex_head |  -.2150275   .1882522    -1.14   0.253     -.583995     .15394
    age_head |  -.0041148   .0044817    -0.92   0.359    -.0128987    .0046691
       _cons |   .7734102   .3625896     2.13   0.033     .0627477    1.484073
------------------------------------------------------------------------------
```

* Now we predict whether a household purifies its drinking water as a function of several household characteristics with a simple logit model.

```
. logit flagpurify maxfedu maxmedu u3 au wu no_members sex_head age_head
Logistic regression                                     Number of obs   =        965
                                                        LR chi2(8)      =     215.34
                                                        Prob > chi2     =     0.0000
Log likelihood = -545.62666                             Pseudo R2       =     0.1648
------------------------------------------------------------------------------
  flagpurify |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     maxfedu |   .0608039   .0233081     2.61   0.009     .0151208     .106487
     maxmedu |   .1931885   .0376512     5.13   0.000     .1193935    .2669836
          u3 |   .2790982   .1858021     1.50   0.133    -.0850672    .6432637
          au |   .1734969   .1491218     1.16   0.245    -.1187764    .4657703
          wu |   1.033577   .1646235     6.28   0.000     .7109211    1.356233
  no_members |  -.1666146   .0349253    -4.77   0.000     -.235067   -.0981623
    sex_head |  -.2193012   .2037548    -1.08   0.282    -.6186533     .180051
    age_head |    .012914   .0052009     2.48   0.013     .0027204    .0231076
       _cons |  -3.783283   .5546931    -6.82   0.000    -4.870462   -2.696105
------------------------------------------------------------------------------
```

* Now we use the logit regression coefficients to predict a propensity score for purification for each household – *purifyhat.*

. predict purifyhat
(option pr assumed; Pr(flagpurify))

* Next we compute the average impact estimate by running the pre-programmed command titled psmatch2.ado. This average difference is reported as ATT. Note, S.E. for ATT does not take into account that the propensity score is estimated. You will have to run a bootstrapping routine to check this. You will also see that 19 purifying households are "off-support", which means that there is no non-purifying household that has similar propensity scores.

```
. psmatch2 flagpurify, out (drink_water) pscore (purifyhat) common trim(5)
------------------------------------------------------------------------------        Variable
Sample |    Treated     Controls   Difference         S.E.   T-stat
---------------------------+--------------------------------------------------------------
     drink_water  Unmatched | .545454545   .659050967  -.113596421   .031705692    -3.58
                        ATT | .551724138   .633952255  -.082228117   .051112448    -1.61
---------------------------+--------------------------------------------------------------
Note: S.E. for ATT does not take into account that the propensity score is estimated.

 psmatch2: |   psmatch2: Common
 Treatment |         support
assignment | Off suppo  On suppor |     Total
-----------+----------------------+----------
 Untreated |         0        569 |       569
   Treated |        19        377 |       396
-----------+----------------------+----------
     Total |        19        946 |       965
```

* The next two commands simply re-define purifying households to include only those on-support.

```
. gen purify2 = _treated==1 & _support==1
. tab purify2

    purify2 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        588       60.93       60.93
          1 |        377       39.07      100.00
------------+-----------------------------------
      Total |        965      100.00
```

* Finally, we test that the matching procedure reduced the average difference in covariates. That is, in the matched sub-sample the purifiers and non-purifiers were more likely to be similar to each other in terms of many household characteristics. There were no longer any statistical differences between them, except in the case of awareness and wealth indices. So some differences still persist.

```
. pstest maxfedu maxmedu u3 au wu no_members sex_head age_head, sum
t(purify2)
-------------------------------------------------------------------------------
                         |       Mean                      %reduct |     t-test
    Variable     Sample | Treated Control    %bias  |bias| |    t    p>|t|
-------------------------+--------------------------------+----------------
     maxfedu  Unmatched | 12.361   10.017     57.2         |   8.39  0.000
                Matched | 12.361   12.183      4.3    92.4 |   0.74  0.459
                         |                                 |
     maxmedu  Unmatched | 13.175    11.19     66.2         |   9.49  0.000
                Matched | 13.175   13.008      5.6    91.6 |   1.13  0.258
                         |                                 |
          u3  Unmatched | .24668   .22789      4.4         |   0.67  0.502
                Matched | .24668   .24138      1.2    71.8 |   0.17  0.866
                         |                                 |
          au  Unmatched | .56764   .45068     23.5         |   3.57  0.000
                Matched | .56764   .47745     18.1    22.9 |   2.49  0.013
                         |                                 |
          wu  Unmatched | .69761   .36565     70.5         |  10.63  0.000
                Matched | .69761   .75597    -12.4    82.4 |  -1.80  0.072
                         |                                 |
  no_members  Unmatched | 5.2042   5.3537     -5.9         |  -0.88  0.377
                Matched | 5.2042   5.1936      0.4    92.9 |   0.07  0.948
                         |                                 |
    sex_head  Unmatched |  .8435   .81293      8.1         |   1.22  0.223
                Matched |  .8435   .83024      3.5    56.6 |   0.49  0.623
                         |                                 |
    age_head  Unmatched | 53.584   51.408     13.9         |   2.08  0.037
                Matched | 53.584    53.18      2.6    81.5 |   0.37  0.709
                         |                                 |
```

Summary of the distribution of the abs(bias)

BEFORE MATCHING

Percentiles		Smallest		
1%	4.413798	4.413798		
5%	4.413798	5.933442		
10%	4.413798	8.104161	Obs	8
25%	7.018801	13.89144	Sum of Wgt.	8
50%	18.71163		Mean	31.22434
		Largest	Std. Dev.	28.52021
75%	61.72629	23.53183		
90%	70.46751	57.22801	Variance	813.4022
95%	70.46751	66.22457	Skewness	.4438761
99%	70.46751	70.46751	Kurtosis	1.390491

AFTER MATCHING

Percentiles		Smallest		
1%	.421106	.421106		
5%	.421106	1.24595		
10%	.421106	2.574613	Obs	8
25%	1.910282	3.515231	Sum of Wgt.	8
50%	3.927334		Mean	6.025631
		Largest	Std. Dev.	6.127983
75%	8.9818	4.339437		
90%	18.14511	5.576318	Variance	37.55217
95%	18.14511	12.38728	Skewness	1.121949
99%	18.14511	18.14511	Kurtosis	2.884055

Sample	Pseudo R2	LR chi2	p>chi2
Unmatched	0.126	163.23	0.000
Matched	0.014	14.56	0.068

[1] Managing for Development Results (MfDR) is a management strategy of the OECD/DAC-MDB Joint Venture that focuses on development performance and on sustainable improvements in country outcomes. It provides a coherent framework for development effectiveness in which performance information is used for improved decision making, and it includes practical tools for strategic planning, risk management, progress monitoring, and outcome evaluation. This strategy was devised at the 2004 Marrakech Roundtable on Results, 2004 under the auspices of the DAC-OECD Working Party on Aid Effectiveness and Donor Practices. The World Bank identified several bottlenecks that limit its ability to conduct impact evaluations at the necessary scale and with the needed continuity: insufficient resources, inadequate incentives, and, in some cases, lack of knowledge and understanding (World Bank 2005). To address these bottlenecks, the Development IMpact Evaluation (DIME) Initiative is a Bank-wide collaborative effort under the leadership of the Bank's Chief Economist that is oriented at: (1) increasing the number of Bank projects with impact evaluation components, particularly in strategic areas and themes; (2) increasing the ability of staff to design and carry out such evaluations, and (3) building a process of systematic learning on effective development interventions based on lessons learned from completed evaluations.

South Asian Network for Development
and Environmental Economics

SANDEE
PO Box 8975 EPC 1056
Kathmandu, Nepal

Tel: 977-1-552 8761
Fax: 977-1-553 6786
E-mail: info@sandeeonline.org
Website: www.sandeeonline.org

SANDEE Sponsors

Swedish International Development
Cooperation Agency

International Development
Research Centre

Centre de recherches pour le
développement international