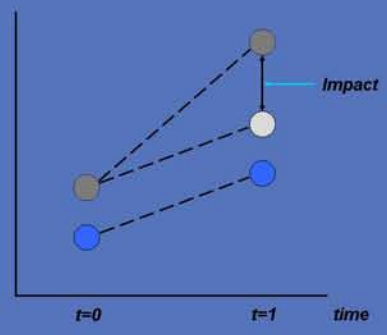


Doing Impact Evaluation Series No. 4

A Guide to Water and Sanitation Sector Impact Evaluations

Thematic Group on Poverty Analysis, Monitoring and Impact Evaluation

THE WORLD BANK
PREM Poverty Reduction Group

A Guide to Water and Sanitation Sector Impact Evaluations

December 2006

Acknowledgement

This paper was written by Christine Poulos, Subhrendu K. Pattanayak, and Kelly Jones.[1] The work was task managed by Caroline Van Den Berg and Markus Goldstein. The paper benefited greatly from comments by Aline Coudouel, Carlos Velez, Frank Drees and participants at a World Bank seminar on this topic. This work program was financed by the Trust Fund for Environmentally & Socially Sustainable Development and by the Bank-Netherlands Partnership Programs fund.

[1] Address correspondence to Subhrendu K. Pattanayak, Fellow, Environment, Health and Development Economics at RTI International (subhrendu@rti.org) and Associate Professor at North Carolina State University. Christine Poulos (cpoulos@rti.org) is a Senior Economist at RTI International. Kelly Jones (k.jones@conservation.org) was an Associate Economist at RTI International while contributing to this document. The opinions reflected in this paper are the opinions of the authors and not opinions of their institution.

TABLE OF CONTENTS

Introduction

Sustainable and equitable access to safe water and adequate sanitation are widely acknowledged as important development goals. Mechanisms to achieve these goals are broad and varied in terms of the types of services (water supply, drinking water quality, sanitation, sewerage, and hygiene); the setting (urban, peri-urban, rural); and the typology of delivery (public or private interventions, decentralized delivery, expansion or rehabilitation). In addition, there is a wide spectrum of possible socio-cultural, economic, environmental, political and legal conditions in which services are delivered. The impacts of water supply and sanitation (WSS) policies and programs range from greater efficiency in the utilities sector, improved access to higher quality services, health improvements, increased incomes and consumption, social and gender inclusion, and education improvements. Yet, to date we have few or no rigorous scientific impact evaluations showing that WSS policies are effective in delivering many of the desired outcomes, except for health.[2]

To understand why we make this claim, consider two criteria that are commonly used. First, a rigorous scientific impact 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 WSS service outcomes. Our thorough review of the literature suggests that there are essentially three sets of completed studies and a few other on-going evaluations that have evaluated the impacts of WSS policies including privatization, decentralized delivery, and information campaigns (see Section III C for a summary of study results). In Section IV E, we speculate on possible reasons why there have been so few rigorous evaluations in the WSS sector.

Duflo and Kremer (2003) 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, as demonstrated by increasing support for strategies such as Managing for Development Results[3] in the development community and the World Bank's (WB) Development Impact Evaluation Initiative.[4]

---

[2] For examples of rigorous evaluations of health impacts of WSS, see Fewtrell et al. 2005, Kaufmann 2005, Esrey et al. 1991, Esrey 1996, Curtis and Cairncross 2003.

[3] 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.

[4] 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

It is important to rigorously evaluate WSS programs and policies for four reasons. First, demonstrating that a particular WSS program yields health, socioeconomic, and poverty reduction benefits can be used to build support for program expansion or modification. Second, even though specific WSS 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 WSS 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, disseminating results of WSS outcomes will contribute to the economic development community's broader understanding of water and sanitation service delivery tools. The purpose of this paper is to serve as a guide for conducting impact evaluations in the water and sanitation sector. 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 texts do not deal exclusively with the issues and situations that can define impact evaluation within a particular sector. Thus, this document highlights three issues that require consideration in designing and implementing impact evaluations for WSS programs.

First, the main outcome of interest for WSS programs is providing people with efficient and sustainable access to safe drinking water and/or basic sanitation services. Thus, we will focus on how impact evaluations can measure changes in access. Second, we view the evaluation through the lenses of a manager in the WSS sector (for example, a WB Task Team Leaders, TTL) as someone who contributes to the design, implementation, supervision and ultimately evaluation of policies and reforms and the construction (and/or rehabilitation) of water, sewerage and sanitation systems. We consider the implications of evaluating both types of interventions. Third, we identify unique issues in data collection for WSS impact evaluations and describe appropriate and useful data collection procedures. As a consequence, we revisit the issue of defining appropriate WSS indicators, which can demonstrate the achievement of WSS targets.

In the next section we define impact evaluation and explain how it varies from other types of evaluation. In Section II we discuss the goals and impacts that are of primary interest in water and sanitation programs, and the interventions most frequently used to achieve program goals. In Section III we summarize the methodological considerations for designing an impact evaluation in the WSS sector, including evaluation design, estimation methods, and data sources and measurement. These issues are treated in greater detail elsewhere (For an example, see Ravallion 2001). We provide examples from several rigorous impact evaluations dealing with water and sanitation sector outcomes. Finally, in Section IV, we discuss some of the major challenges and practical solutions in impacts evaluations of WSS projects and policies.

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.

I. Impact Evaluations: Dealing with Counterfactuals and Confounders

Evaluation means different things to different people and often the manager confuses a general monitoring 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 document, measures the impacts of the program on individuals, households, or other groups such as firms, and determines whether the program *caused these impacts* (Baker 2000; WB-OED 2004).

The fact that impact evaluation is concerned with the results that are *caused* by the program distinguishes it from 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, we often assume that people would have access to the piped network at the same rate without the intervention (e.g., under private sector participation) as before the intervention. This can be misleading because there could be a general trend towards more access, for example, because of improving economic conditions. Impact evaluations must therefore rely on control (or comparison) groups, as well as a number of statistical and econometric techniques to estimate this counterfactual (see Section III for details). 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 in the WSS include socio-cultural behaviors (*e.g.* collective action to improve access to community sources), institutional factors (*e.g.*, other programs promoted by other government departments, non-governmental, or donor organizations), bio-physical characteristics (*e.g.*, water table and geology), and general trends. 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

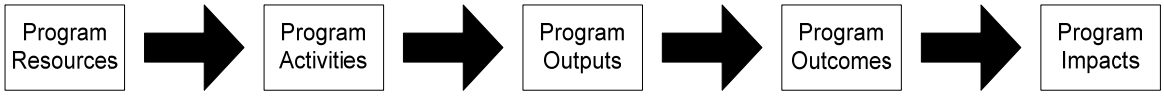
of confounders is the primary distinguishing feature between process evaluations and impact evaluations.

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. Bamberger et al. (2004) describes several ways to reduce the threat of selection bias, including statistically controlling for differences between treatment and control groups – but keep in mind that these controls can only account for observable, not unobservable differences between treated and untreated groups.

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 in WSS (see Section III 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, Cook, and Leviton, 1991). Most WSS impact evaluations use these designs. Both data and design issues are discussed in greater detail in Section III.

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.

Figure 1: Generic Model of a Program



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.

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., per capita water consumption); (2) varying across areas, groups, over time, and sensitive to changes in policies, programs, and institutions (e.g., hours of water supply); (3) not easily diverted or manipulated (e.g., presence of a pit latrine); and (4) able to be tracked (e.g., functionality of public stand pipes). During the evaluation process, it is important to monitor program inputs though what are called "intermediate" indicators provide information on activities and outputs and thus provide valuable information on whether a program was implemented successfully (Bosch et al. 2000; Prenusshi et al. 2000). Outcomes and impacts are tracked through "final indicators".[5]

Program resources and program activities constitute the program inputs. 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.

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 (e.g., funds allocated and disbursed to capital improvements, FTEs assigned to implement 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 a WSS program activities can include the number of communities selected for system improvements, funds disbursed, completion of planning processes and documents.

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 and rehabilitation of public infrastructure (e.g., standpipes, length of distribution pipe) or hygiene trainings (Bosch et al. 2000). For WSS evaluations, these outputs can also be viewed as "interventions" that affect outcomes of interest and generate impacts. See Section III C (and Appendix III) for examples of privatization, decentralization and information interventions.

[5] Note that this terminology differs slightly from the one used by the one that is used in the Results Framework of the Project Appraisal Documents, which currently uses intermediate and final outcomes as the projects' results for IDA14 reporting requirements.

Program outcomes are the changes in behaviors, knowledge, and actions among participants as a result of the program. A program can have both short term outcomes (attainable in 1-3 years) and long term outcomes (attainable in 4-6 years). Program impacts are the fundamental change experienced by program beneficiaries as a result of the program. An outcome indicator will measure access to, use, or satisfaction with the intervention. These are not the fundamental changes intended by the program, but are closely related. The fundamental changes are improvements in beneficiaries' well-being measured, for example, in terms of health or income. See Section III C (and Appendix III) for examples of water access, service quality and health impacts.

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, rural WSS programs may initially target poor communities that are located closer to water sources because of the cost advantages of serving these communities relative to more distant communities. Due to their proximity to water sources, targeted communities may have better 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 upwardly biased estimates of impacts.

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. Since impact evaluation can be complex and expensive, Baker (2000) and Ferraro and Pattanayak (2005) 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 cannot be generalized because of certain peculiar characteristics of the population, institutions, systems, program, or environmental setting. If the project is experimental and likely to be revised over time, it could be difficult to conduct an impact evaluation. However, if the evaluation is integrated with a well planned experimental project, it is possible that the evaluation could provide an answer on which intervention to scale up *within* a given project (for example by trying multiple delivery mechanisms or interventions on a pilot basis).

Also, impact evaluations are more likely to be beneficial when the outcomes are a matter of debate. Given the paucity of rigorous impact evaluation in the WSS sector, there are a number of unresolved issues. First, we are aware of no evaluations that demonstrate the impacts of WSS programs on poverty, including income, consumption

levels, education, or gender and ethnic inclusion. Second, there is insufficient evidence on the impacts of community-driven projects (Mansuri and Rao, 2004).[6] There are no studies that establish a causal relationship between any outcome and the participatory elements of the project. Third, merits of household-level versus community-level interventions, e.g., point of use water treatment vs. source water treatment, are also unclear. While the most recent systematic evaluations of the epidemiologic literature suggest water treatment at the household level is more effective in preventing enteric disease than improvements at the source, the studies of this topic show that the effectiveness varies by setting and some studies have methodological flaws that leave them vulnerable to bias and limit their comparability (Clasen et al., 2006). Finally, the impact of privatization and small-scale providers on households, as well as utility performance, remain areas needing additional study.

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. WSS programs typically have multiple objectives, some relating to results in the WSS sector (e.g., increase per capita water consumption, increases access, improve water utility performance) and others related to results outside the sector, including health outcomes. For example, in Section II A we discuss how the World Bank decided to focus on the provision of efficient, equitable, and sustainable access to WSS services and the indicators to measure such provision.

Table 1: Examples of Indicators for a Hypothetical Public Water Supply Project

Program Component	**Indicator**	**Example**
Inputs	Intermediate	Funds allocated to project villages
Outputs	Intermediate	Number of public standpipes installed
Outcomes	Final	Distance walked to nearest public standpipe
Impact	Final	Mortality rates
Other	External	Groundwater recharge

Source: Modified from Prenusshi et al. (2000)

[6] Mansuri and Rao (2004) review studies that evaluate community-driven economic development programs in a number of sectors, including WSH, labor, agriculture, and others. Their review finds 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.

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 III).

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.

II. Evaluation of Water and Sanitation Projects

A. Objective of WSS Projects

WSS programs have far-reaching impacts, many of which are outside of the water and sanitation sector and include impacts on the environment, human health, businesses, and poverty. For example, World Bank assistance in WSS has shifted from a narrow focus on physical infrastructure to fostering operationally and financially sustainable service provision over the past decade. As a result, WSS programs and policies are influenced by the Bank's Sector Strategies in Rural Development, Private Sector Development, Urban and Local Government, and Environment.[7] It can be a challenge to clearly identify program objectives given the influences of other sectors and the breadth of program impacts.

One of the primary issues is whether to consider water and sanitation to be a final product (of WSS policies) or an intermediate product in the production of human welfare. The latter - typically measured in terms of health, education, social exclusion and income (direct and indirect impacts) – is a key concern of development practitioners. However,

[7] http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTWSS/0,,contentMDK:20204225~menuPK:511970~pagePK:148956~piPK:216618~theSitePK:337302,00.html

the World Bank Water and Sanitation Sector Board (the Board) guidelines direct TTLs to view water and sanitation as a product. In 2003, the Board agreed that all new dedicated projects involving water and sanitation will track two standard indicators (1) the project's contribution to increasing access to improved water supply services and (2) the project's contribution to increasing access to improved sanitation services.[8]

By focusing on access (a program outcome) rather than on health or poverty (a program impact), the WSS sector can measure the success of its projects by focusing within the sector. While everyone recognizes that health, education, and other welfare impacts are critical, it is impractical to require every WSS manager to design and implement projects to deliver these. Thus, there is a compelling rationale to focus on the outputs and outcomes of policies, programs and projects within the purview of WSS managers such as private sector participation, municipal management, tariff reform, and regulation.

Access, which is an intermediate outcome (see Figure 1), measures the availability of WSS services to project beneficiaries. The impact evaluation of the Bolivian Social Investment Fund (BSIF) (Newman et al., 2002) used several indicators to measure access, including: proportion of households with piped water and sanitation facilities, distance from house to water source, duration of supply, and fraction of year with adequate water. It is important to note, however, that access does not reflect actual use, efficiency or sustainability; these dimensions can be measured as final outcomes of a project intervention.

The success of WSS programs also requires that access to improved services be both efficient and sustainable. Efficiency reflects both technical and allocative aspects, but essentially reflects least cost provision, where costs are supposed to include extermalities such as environmental costs. As a result of poor efficiency in the WSS sector in past decades, many sector reform policies have focused on reducing costs and improving revenues, for example, through the introduction of private sector participation in the WSS service delivery.

Cost-benefit analysis and cost-effectiveness analysis are economics tools that evaluate respectively allocative and technical efficiency. These tools can use the results of impact evaluations as inputs in the form of X% increase in water supply coverage because of a particular policy, for example. Both types of studies compare the costs of WSS programs with benefits such as the percent increase in coverage. In cost-effectiveness analyses, benefits are measured in non-monetary terms (e.g., number of liters per capita per day, time spent traveling to and waiting at water source, number of hours spent in school, diarrhea prevalence). In cost-benefit analysis, benefits are monetized using a variety of non-market valuation methods (Pattanayak et al., 2005c; Hutton and Haller, 2004).

---

[8] The Board's 2004 WSS Program report implies that achieving poverty reduction requires a focus on "efficient sustainable water supply and sanitation services for all."
http://siteresources.worldbank.org/INTWSS/Publications/20249484/WSS_report_Final_19Feb.pdf

While impact evaluations measure the magnitude of the impacts of a WSS program, it is also important to conduct studies to assess the durability of impacts, or whether impacts are sustainable over time. By taking a short-term view such as measuring if program beneficiaries received more than 40 liters per capita per day for the first three years of operation of the project (caused by the project), we run the risk of ignoring the fundamental program delivery question – is this provision sustained through the 20-30 year life of the project? Sustainability is influenced by the financial viability of the utility (can the utility continue to produce WSS services?), affordability for beneficiaries (will beneficiaries continue to use the services?), environmental quality (will the services have a positive or negative impact on local environmental quality), and technical feasibility (is the system or water source capable of providing the WSS services over time?). There are two different approaches to measuring the sustainability of impacts. The first is to measure indicators periodically over many years. This approach is impractical when those measurements would be required long after the project cycle has ended. The second approach is to identify risks that threaten the sustainability of impacts and assess whether those risks are present in the WSS program being evaluated (White, 2005).[9]

Finally, while not explicitly mentioned in the WSS objectives, poverty alleviation objectives make it necessary to consider if WSS outcomes are equitable. This concerns equal access for all members of society, regardless of age, sex and social, cultural, religious, or ethnic status. We return to the issue of sustainability and equity in Section IV.

In conclusion, there are three goals guiding most WSS programs:

1. efficient access to safe drinking water and/or basic sanitation services;
2. sustainable access to safe drinking water and/or basic sanitation services; and
3. equitable access to safe drinking water and/or basic sanitation services.

B. Intervention Mechanisms and Outputs

In this subsection, we describe the main WSS projects, programs and policies pursued by WB TTLs. The goals and expected impacts of these programs are described in Section II A. Given that these activities have expanded beyond the provision of infrastructure to focus on broader issues underlying service provision, we distinguish WB interventions, which include sector reform initiatives (improving operator performance, PSP and SSIP, and CDD-Decentralized delivery), from the construction, expansion, and rehabilitation of infrastructure that are the outputs of initiatives.

[9] This approach is consistent with IEG's definition of sustainability that focuses on assessing sustainability by using measures from the current or near-future time periods. IEG defines sustainability as "the resilience to risk of net benefits over time." An impact evaluation, then, should analyze how resilient project outcomes are to contemporaneous risks in order to assess sustainability. Toward this end, the analyst should first identify the risks and then perform a sensitivity analysis to determine the risks to which project outcomes are most sensitive. (See The World Bank's Operational Policy 10.04).

B.1 Projects, Programs and Policies - Interventions

The recently issued World Bank Group's Program for Water Supply and Sanitation directs World Bank's assistance towards thematic and regional priorities, including:

- ✓ extending services to the urban poor;
- ✓ improving operator performance;
- ✓ increasing access to rural water supply and sanitation; and
- ✓ managing water resources effectively.

The World Bank supports countries in achieving these priorities essentially through the use of loans and guarantees, and technical assistance (*e.g.*, advisory work and policy dialogue).[10] These instruments are the primary mechanisms for achieving sector reform which are the measures taken to increase financial viability and improve the institutional performance of the WSS sector.

Three types of reform measures are predominantly used to help improve the performance of WSS sector in terms of efficiency and equity:

- ✓ improving operator performance,
- ✓ service provision by the private sector or small-scale independent providers, and
- ✓ decentralized delivery, typically relying on community demand, participation and management.

Since there are regional differences in TTLs objectives as they relate to WSS program outputs, these measures are employed to achieve a wide range of service improvements. In Africa, for example, there is a greater emphasis on installing and rehabilitating infrastructure in order to provide access to safe water from a WSS system. The current level of services tends to be lower than they are in, say, Latin America, where many people have access to the water supply system. In these areas, the quality of service (in terms of regularity, pressure, and water quality) may be the focus of WSS improvements. The measures are described below.

Improving Operator Performance

Since most utilities must substantially reduce costs and increase revenues to become financially viable, improving and extending service delivery requires the turnaround of utilities into well-functioning and financially viable entities. These entities are more likely to provide WSS services efficiently and sustainably, thus improving beneficiaries' access to WSS services, among other things (including health and income). In addition to improving access, improvements in operator performance are likely to improve indicators of utility efficiency, including reductions in unaccounted for water, and higher revenues and cost recovery rates. For example, Galiani et al. (2005) present a case study of how a privatized utility increases water production, reduces spillage,

[10] See www.worldbank.org/watsan for a description of the four official instruments – lending portfolio, policy dialogue, guarantees, and partnerships.

increases water and sewage network coverage, reduces delays in repairing among other performance metrics (see Table 7).

There are four approaches to improving operator performance in the WSS sector. The first is modifying operator institutional arrangements such that management and employees become more accountable. Business plans, standard processes and streamlined procedures, cost accounting techniques, and decentralization of responsibilities are measures to enhance accountability in a utility. This must be accompanied by capacity building for affected staff.

These changes should be accompanied by changes in the sectoral institutional framework, the second approach. These measures include a clear division of responsibilities and lines of accountability among sector and central institutions. The third approach is restructuring tariffs and subsidies to meet social, technical and/or economic objectives. Improved billing and collection procedures and higher tariffs and the revenues they generate can be used to improve services through expansion and rehabilitation.

Finally, the fourth approach is operator financing. These include approaches to improve the efficiency of public spending and using it to better leverage other sources. Output-based-Aid (OBA) subsidies are examples of specific interventions in the WSS sector that are explicitly tied to performance and results.[11] Despite interesting case studies (Drees et al., 2004; Mumssen, 2004), there are no known impact evaluations of these types of OBA strategies.

Private Sector Participation (PSP) and Small-Scale Independent Providers (SSIP)

Traditional WSS utilities have often ignored the needs of poor urban households, perceiving them as customers who are costly to serve and unable to pay their bills. However, over the next two decades, the world's population balance in all regions will tip towards urban areas and meeting the needs of this population is a challenge because of the differentiated service demands of poor households. The range of options that can be offered by private sector providers and by small scale providers (whether private or not), can better address these needs.

In this context, PSP refers private operators manage WSS services and may assume a combination of operating, commercial or financial risks to construct, expand, rehabilitate WSS systems. While PSP can help improve access to WSS services , there is also concern that the poor are not served by these providers.

Small-scale providers are assumed to serve fewer than 50,000 people each or 5,000 customers in small settlements, whether urban, peri-urban or rural. In a report by

[11] Output-Based aid (OBA), a strategy for using explicit performance based subsidies to deliver basic services—such as water, sanitation, electricity, transport, telecommunications, education, and health care—where policy concerns would justify public funding to complement or replace user fees. Two key features distinguish OBA subsidies from some other forms of publicly funded subsidy: OBA subsidies are explicit, and they are performance based.

Kariuki and Schwartz (2005), SSIPs are distinguished from other small scale providers by the fact that they are established: (i) at the initiative of a private owner or operator, which may be either a for-profit or non-profit organization; (ii) with a significant share of capital financing (25 percent or more) provided or borrowed by the private entity; and (iii) on a commercial basis (although they may be unprofitable or have non-profit status). Private and small scale providers are likely to have the same impacts on beneficiaries as other WSS providers (changes in outcomes and impacts), however, the outputs and activities may be different.

Community Demand Driven (CDD)/Decentralized Delivery

With one-third of the population of Africa and Asia living in towns with populations that range between 2,000 and 200,000 people, decentralized water supply and sanitation is fundamental to economic growth and achievement of the Millennium Development Goals. As with peri-urban and urban areas, these populations demand differentiated services but lack resources to manage and operate their WSS systems.

Since government resources are limited, decentralized delivery should establish local service providers with a minimum investment, and to ensure that reforms are put in place so that these providers can meet carefully defined cost-recovery objectives. Governments need to identify appropriate approaches for involving both service providers and town administrators (regulatory oversight) that address several aspects of service provision such as design and financing requirements that suited to localities, and need for effective professional support systems (e.g., contracting local partners to secure professional capacity).

In rural areas, CDD decentralized delivery often translates into putting the community front and center of the planning, design, implementation, and operations process and replacing career bureaucrats with qualified professionals and technocrats. The CDD philosophy is typically captured in reflecting the needs of community through the use of a participatory approach, decentralized delivery, cost sharing (typically 10-30% of capital, and 100% of O&M), and a strong component of local institutional strengthening.

The resources available to a program influence the activities that will be carried out. These activities should be in line with the ultimate objective of the program. While no means exhaustive, some common activities carried out in WSS programs include community mobilization, village planning, technical assistance, provision of financial or physical materials/supplies, and community capacity building/training.

For these types of policies, outputs are expected to include greater satisfaction with the interventions due to community participation in design, and better maintenance because of community ownership, among other things. The impact evaluations of Social Investment Funds summarized by Rawlings (2004) are examples of evaluations of CDD programs.

B.2 WSS Sector Outputs

What then are the results of programs and projects that incorporate operator reform, PSP and CDD policies? Reform is not an end in itself but a means to provide a basis for expanding access and improving the quality of service. Reform measures yield various resources that support downstream activities and outputs in the sector, including the construction, expansion, and/or rehabilitation of water supply, sanitation, and sewerage infrastructure. In addition to physical infrastructure, sector reforms provide resources that support the provision of "software" investments (e.g., education or institutional change) aimed at changing personal hygiene behaviors as well as the behavior (operation and management) of organizations, including utilities, communities, and cooperatives. Demand-side management and water conservation practices are examples of other kinds of behavior change, as well as the other outputs described in Sections II B.1 - II C.

It is useful to think about output indicators that measure either provision of "hardware" (e.g., infrastructure) or indicators that measure the provision of "software". WSS outputs that are "hardware" are simply the types of products and levels of service under the direct control of program providers, whether they are public sector, private sector, or community organizations. Outputs include installation of public infrastructure or hygiene trainings (Bosch et al. 2000; Prenusshi et al. 2000). We can classify these "hardware" WSS outputs into four categories: (1) water supply (quantity); (2) water quality; (3) household level sanitation; and (4) environmental sanitation.

Water quantity improvements include the provision of new or improved water supply infrastructure at the public or household level. Possible interventions encompass everything from a hand pump to a household connection.

Water quality improvements include improvements in clarity, odor, taste, and treatment of water to control for bacteriological and chemical contaminants. Note that water quality can be both under the control of water utilities and influenced by household behavior depending on whether water quality is measured at the source or at the point of use.

Household level Sanitation improvements include options for management and disposal of human wastes

Environmental Sanitation typically includes disposal and management of community, industrial (e.g., wastewater), agricultural and household wastes through drains, garbage dumps and wastewater treatment facilities.

Clearly, any project may include any combination of the above intervention typologies.

"Software" outputs include hygiene information and education campaigns for beneficiaries, training for sector staff, improvements in utility and sector management (e.g., appropriate bookkeeping practices, improved billing, improved accountability,

number of concession contracts awarded). These outputs will vary by WSS policy chosen, but are hypothesized to be important determinants of changes in access and other final outcomes and impacts.

C. Outcomes and Impacts of WSS Interventions

Since WSS interventions effect results in several sectors, there are many potential areas of impact. This section summarizes the main types of outcomes and impacts. The intermediate outcome is access. The indicators of access depend on the outputs of the program. Some indicators, such as the average distance from beneficiaries' homes to a water source, measure the availability of services. Other indicators, such as the liters consumed per capita per day, the number of hours of service, and the quality of drinking water, measure the quality of services. Box 1 summarizes the outcomes that were the focus of the impact evaluation study of the Bolivian SIF (Newman et al., 2002) and the indicators that were selected to measure outcomes. More details on the Bolivian SIF impact evaluation are reported in Appendix I.

Box 1: Outcomes and Indicators in the Bolivian SIF

Outcomes	Indicators
Improved access to water supply and sanitation infrastructure	Proportion of households with piped water and sanitation facilities
	Distance from house to water source
Improved availability of water supply	Hours a day of water availability
	Proportion of year with adequate water
Improved health behaviors	Proportion of households boiling water before consumption
	Proportion of households using oral rehydration therapy

WSS impacts on beneficiary well-being are categorized by Bosch et al. (2000) into four groups: (1) health improvement; (2) education; (3) gender and social inclusion; and (4) income/consumption increases. To our knowledge, no study has demonstrated that any WSS program or policy has generated all four impacts. Table 2 lists impacts and potential indicators.

Indicators for health would include changes in morbidity and mortality rates and changes in anthropometric measures. WSS improvements could be related to a variety of water-borne, water-washed and water-related diseases, but the usual focus is on diarrheal diseases, respiratory illnesses and malnutrition. Indicators for education include enrollment, attendance, and achievement rates. Indicators for inclusion can involve examining the impacts on the poor, ethnic minorities, women, or rural households. Changes in income and consumption can be measured through indicators on expenditures on WSS services, coping costs (see Pattanayak et al. 2005c), or total expenditures. Box 2 summarizes the final impacts that are the focus of an ongoing impact evaluation in India. The details of the study are summarized in Appendix I and Pattanayak et al. (2005a).

Table 2: Program Impacts of WSS Programs and Possible Indicators

Impacts	Indicators
Health improvement	• Diarrhea (three or more loose stools over 24 hour period) in last two days? two weeks? among children under 5 years in the sample • Acute respiratory infection – incidence of cough and cold in last two weeks among children under 5 in the sample • Body mass index for children under 5 in the sample
Education	• Whether children in sample are enrolled in school • Whether children in sample attend school regularly • Number of days children in sample attended school in the last month
Gender and social inclusion	• Women's perceptions about level of privacy provided by access to sanitation • Women's perceptions about safety of using water and sanitation services • Access to WSH services by poor, minorities and vulnerable groups
Income/consumption	• Household per capita income and consumption • Household coping and averting costs (expenditures on household water treatment, water storage containers) • Expenditures on medical treatment

In addition to intended impacts, analysts should also be aware of and try to measure any unintended impacts – either positive or negative – when evaluating WSS programs. These could result, for example, from environmental, socio-behavioral, or economic spillovers across communities, neighborhoods and or schools. For example, there may be spillovers due the positive externalities associated with infectious disease prevention. Both treated households (e.g., those making use of improved water or sanitation) and untreated households could experience reductions in disease because the disease-causing agents in the environment are reduced (see Miguel and Kremer, [2004] for a worm treatment example). Hygiene education messages may flow from treated communities or persons to untreated communities and persons (Dearden et al., 2003). There may be unintended costs if the water supply is insufficient for the population and subsidence or saltwater intrusion occurs as a result of well use, or if water sources are contaminated.

Box 2. Final Impacts and Indicators in Jalswarajya Evaluation

Impact	Indicator
Children's health	Under-age-five diarrhea rate
Livelihoods	Averting and coping costs; income and wealth indicators
Social inclusion	Impact on girls, women, and scheduled castes and tribes
Education	School attendance, hygiene and health literacy

III. Impact Evaluation Methodology

A. Data Requirements

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 (Bosch et al. 2000). For example, Pattanayak et al (2005a; 2005b) use surveys to collect qualitative information measuring perceptions about water quality, satisfaction with WSH services, awareness of improvements. These data on knowledge, attitudes, perceptions, and practices are frequently measured using Likert scales[12] and are tracked in order to measure changes due to the intervention.

For example, a common indicator in many WSS evaluations is water quality. Water quality attributes such as clarity, odor, and taste are observable can be gathered from either household surveys or secondary data. Qualitative information, such as household perceptions and opinions about taste, etc. are often used to measure these indicators. These measures will influence satisfaction with WSS services, but will not affect health. The presence of micro-organisms, on the other hand, is not easily observed and is better suited to quantitative measures. Microbiological test data may be available from governmental water ministries; if not, water quality tests can be conducted. Financial and logistical considerations, such as proximity to a water lab, availability of transportation, and the cost of collection, will determine the feasibility of this option. Pattanayak et al. (2005a; 2005b) use water quality tests in their evaluations of government WSH programs in Maharashtra and Orissa. They test water quality at the source and the point of use in order to understand how household behavior affects water quality.

Table 3 lists some possible sources of data for each indicator as well as actual data sources used in an ongoing impact evaluation of the Jalswarajya Program in Maharashtra, India. Ideally, to minimize time and costs, some information will be available on indicators from secondary sources. For example, intermediate indicators can usually be tracked from administrative records or governmental ministries (Prenusshi et al. 2000). They could include detailed consumer surveys conducted by the utility or even survey data collected by an organization separate from the evaluation team at an earlier point in time. Most impact evaluations conducted by the World Bank have often relied on survey data available from national agencies. Pattanayak et al. (2005a) use secondary data to identify two matched control villages for each village enrolled in the Jalswarajya project – a government program facilitating CDD (see Appendix II for details on this

[12] A **Likert scale** is a type of psychometric response scale often used in questionnaires, and is the most widely used scale in survey research. When responding to a Likert questionnaire item, respondents specify their level of agreement to a statement.

impact evaluation). Village or household level data from the Indian Housing Census, the Indian Population Census, and the National Family Health Survey were combined with information on whether the villages were enrolled in the Jalswarajya Program. These data were used to implement propensity score matching (See Section III) before the interventions were administered.

Indicators of program impacts require that information be gathered on program beneficiaries through household surveys. It may be possible to use data collected from panel surveys (e.g., LSMS, DHS, national census) or to add questions to on-going survey projects as long as the sample includes adequate numbers of treated and untreated units. There are tradeoffs to using secondary data, such as loss in accuracy or reliability. For example, most of the commonly available large scale representative surveys such as LSMS and DHS collect cursory information on water and sanitation, if they collect them at all.

However, these sources are less expensive than collecting primary data. If no data exist, then the researcher should investigate the possibility of joining planned survey efforts in the area. The third option is primary data collection, which is the most costly, but the most reliable data collection method (Prenusshi et al. 2000; Bamberger et al. 2004).

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.[13] 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). 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.

Researchers must be aware of any validity or reliability problems. For example, survey questions about personal hygiene questions are vulnerable to social desirability bias in which the respondent may provide responses that are incorrect but socially appropriate. Alternatively, observational data on these behaviors may be difficult to collect because of privacy concerns. These measurement problems are not unique to impact evaluation, or to the Water and Sanitation Sector. In addition, a number of general approaches can be applied including triangulation of data, using multiple variables to create an index, and using calibration factors (if available). For example, the USAID's Hygiene Improvement Framework relies on multiple measures of personal hygiene in

[13] The significance level (typically called the size or α) represents the probability of a type I error, which is the rejection of the null hypothesis when it is true. The power of a test is represented by $1-\beta$, where β is the probability of a type II error, which is the failure to reject the null hypothesis when it is false). Here we are talking about the significance and power of a statistically estimated impact.

order to assess hygiene levels (Kleinau et al. 2003). The framework combines numerous indicators of hygiene and collects data using survey questions, direct observation, and the presence of hygiene-related materials (soap, ash, covered water storage).

If we think there are sufficient pre-existing differences between the project and comparison groups, the evaluation design will require measurement of indicators before and after the implementation of the program (See Section III B for research designs). In these cases, timing of data collection is important. Pre-intervention or baseline measures should be taken prior to any program activity. In the case of demand responsive programs, the measurement should be before the beneficiaries have information about program.

Box 3: Data Sources Used in Bolivian SIF Impact Evaluation

The Bolivian SIF impact evaluation used a combination of secondary sources and primary data collection. In addition to information on program beneficiaries, the evaluators also collected information on facilities, the community, and water quality. Specific data sources :

- ✓ Household surveys
- ✓ Community surveys
- ✓ Water quality samples

Table 3: Possible Data Sources by Program Component

Evaluation Component	**Possible Data Sources**	**Data used in Pattanayak et al. 2005a**
Intermediate Indicators *Inputs* *Outputs*	Administrative data Expenditure data Payroll data Community surveys Program staff surveys	Government of Maharashtra data on Jalswarajya program (village applications, village action plans for program, expenditures, construction) Community Surveys (community planning processes, implementation of plans, funds received, maintenance activities)
Final Indicators *Outcomes* *Impacts*	Household surveys Existing panel data (LSMS, DHS, national census, PS/CWIQ) Consumer records (e.g., billing and metering data from utilities) Qualitative surveys Water quality samples	Household survey (access to WSH services, satisfaction with WSH services, coping costs, water sources used, knowledge and awareness of WSH, income/consumption, acute and chronic health, diarrhea rates, education, demographics,)
External Factors *Mediating Factors* *Intervening Factors*	Household surveys Qualitative surveys Direct measurement	Household survey (diarrhea rates, access to WSH services, knowledge and awareness of WSH services, income/consumption, acute and chronic health, education, demographics, satisfaction with WSH services) Community Surveys (other social programs, weather conditions, water sources)

B. Design and Analysis

One way to think about the method for an impact evaluation is in terms of two related components: the evaluation design and the estimation method. While statistical methods are loosely associated with evaluation design, some methods can be used with various designs (see Appendix IV for more information on statistical methods).

Experimental (or randomized) and quasi-experimental designs vary in how they construct the control group and estimate the counterfactual depending on the nature of the program, as well as feasibility, cost, validity, and degree of selection bias. Ravallion (2001) emphasizes that impact evaluation designs must fit the conditions and characteristics of the program. Thus, design and methods must be selected on a case-by-case basis.

Experimental design requires participants to be randomly assigned to the water and sanitation intervention. Given an adequate sample size, this ensures that participants and non-participants are statistically equivalent, and in theory controls for all selection bias. Despite the robustness of this design, they are difficult to implement (Heckman and Smith 1995), and often infeasible to implement in a WSS program due to ethical or political reasons. For example, it will be impossible to such as randomize network expansion, which depend on the locations of current networks. Randomized designs may also be impractical because it is considered unfair to deny services to people at random, or to randomly assign programs that provide benefits, rather than to target them to areas that most need them. Consequently, they are rarely done in WSS (see Section III C for some exceptions).

Another problem often encountered is the scale of a program – a WSS 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. Similarly, while the sector's emphasis on bottom-up, community-driven WSS programs may limit the use of randomized design, as a random assignment would contradict the overall intention of a community-led program although it would be possible to randomize eligibility for participation across communities.

Randomization is much more straightforward in the case of information campaigns. In addition to the randomized trial described in Box 4), and Luby et al. and Jalan and Somanathan (2004) both use randomized designs to evaluate information campaigns. Since 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. Box 4 summarizes an ongoing impact evaluation of a sanitation information, education and communication program that uses a group-randomized design.

Box 4: A Randomized Trial in Orissa, India

A WSS impact evaluation in Orissa, India (Pattanayak et al. 2005b) randomly assigned 20 villages to receive a government-funded sanitation campaign. Treatment villages receive an intensive information, education, and communication (IEC) campaign geared at stimulating demand for individual household latrines, and subsequently, financial and technical support to construct household latrines. Twenty control villages were also selected that are similar in all observable characteristics, except that they will not receive the intervention. The evaluation measures whether the IEC campaign increases household latrine uptake and, and whether this uptake improves child health within treatment villages.

Randomized designs are susceptible to validity threats due to spillover effects, and discussed in Section II C. Measures to limit these threats are discussed in Section IV B.

Due to the practical limitations of experimental design, almost all the existing WSS impact evaluations have used a **quasi-experimental design**. In this setting, the researcher must construct a control group that is as similar as possible to the treatment group, usually through a **matched comparison**.[14] Given the preponderance of matched comparisons in WSS evaluations, we review the two main ways of constructing matches:

Propensity score matching (PSM): 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 allows the calculation of a score for everyone, and then participants and non-participants are matched according to this propensity score. PSM is the most common method used to control for selection bias in the WSS sector because it is quicker and cheaper to implement than other methods, and is considered scientifically robust. It has been used in evaluations of SIFs in Armenia (Chase 2002), Bolivia (Newman et al. 2002), and in evaluations of private sector participation in Argentina (Galiani et al. 2005) and other South American cities (Clarke et al. 2004). Since this method is statistically complex, it requires a team with statistical expertise. Box 5 describes the use of PSM in an impact evaluation in India.

[14] While selection bias is a threat in any quasi-experimental design, matching can reduce selection bias by controlling for observables. Matching identifies non-program participants that are comparable in selected indicators – indicators that could affect program outcomes – to participants. This matching can be done before or after project implementation and relies on existing data sources such as national data. Matching, randomization and natural experiments rely on different techniques to construct a comparison group and thus should be distinguished from reflexive comparisons. Reflexive comparison requires that program participants be compared to themselves, before and after the intervention. This method lacks the scientific rigor (i.e. establishing causality) of matching because it cannot ensure that the change in intended results is a response to the program or to a number of other factors that could have affected program participants before and after the intervention.

Box 5: Using Matched Comparison in Jalswarajya Impact Evaluation

In the India's Jalswarajya Program (implemented by the Government of Maharashtra), communities self-select interventions by communities and program administrators target provision. This combination of bottom-up demand and targeted supply precluded the use of a randomized assignment. To construct a comparison group for an impact evaluation on child health, therefore, villages that were participating in Jalaswarjya were matched to villages that shared similar or identical characteristics based on secondary data on a set of demographic, socioeconomic, environmental, health, and WSS variables. A logit model was estimated with these data to predict a participation probability (propensity score) for all project and non-project villages. Project villages were paired with comparison villages that had similar or identical propensity scores. A comparison of village characteristics shows that the paired villages have balanced characteristics. See Appendix II for more details on this impact evaluation.

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) and Honduras (Walker et al. 2002). Box 6 summarizes the use of pipeline matching in an impact evaluation in Armenia.

Box 6: A Pipeline Comparison in Armenian Social Investment Fund (SIF)

In the Armenian SIF (Chase 2002), a pipeline comparison was used – villages where Social Funds had approved a project but the project had not yet been completed were selected as the control group. PSM was then used to match individual villages on mean per capita expenditure, mean share of food in expenditure, share of female household heads, and mean of household head's education.

Another potential means to construct a counterfactual is to rely on a natural experiment, where the evaluation team has access to what is described in econometrics as an 'instrumental variable'. We are tentative in our recommendation of this approach because there is no known application of this approach to the WSS sector. Instrumental variables are exogenous factors that influence the intervention but not the WSS outcome. State or municipality borders are examples of instruments. The instrumental variable and closely related 'control function' approach are often applied to social programs, for example, estimating returns to education (Card 2001; Heckman and Navarro-Lozano 2004). However, as Ferraro and Pattanayak (2006) suggest, "in general, good instrumental variables are hard to find. Using instrumental variables typically requires a mix of clear theoretical intuition, good quality secondary data and a solid grasp of field conditions."

One way to determine which evaluation design is feasible for an evaluation is to consider what data is available. Adapted from Prenusshi et al. (2000), Table 4 shows what data is needed for a particular evaluation design.

Table 4: Data Requirements by Evaluation Design

Evaluation Design	**Data Requirement**	
	Minimal	***Ideal***
Experimental	Single cross-section data for treatment and control group	Baseline and follow-up data for treatment and control group
Quasi-experimental Matching comparison	Single and cross-section data for treatment and control group, oversampling of control group	Baseline and follow-up data for treatment and control group, oversampling for control group

Finally, when an intervention is national in scale comparison groups cannot be identified, and it is necessary to simulate the counterfactual. This counterfactual is constructed using a theoretical model and information on the situation prior to the intervention. Shirley et al. (2000) evaluate the welfare impacts of a 1988 national reform in Chile which restructured water and sewerage subsidies and tariffs. The authors project the trends of key variables based on their linear trends in prior years to construct the counterfactual. These simulations are challenging and require the analyst to make assumptions about what WSH conditions (e.g., water supply and production, prices, access) would have been in absence of the program. Without any data on the counterfactual situation, these assumptions cannot be verified.

Contingent valuation surveys – also called 'WTP experiments' – provide a complementary type of analysis to impact evaluation by drawing on the basic logic of random assignment of information treatments. These studies typically measure how tariff design (*e.g.*, price per cubic meter) impacts coverage rates for water supply (*e.g.*, number of connections). Different sub-samples of survey respondents are asked about their decision to connect to the network after being informed about a particular tariff scenario, which is different for different households based on random assignment. Because these studies use a split sample design to assure orthogonality, data on households' 'stated' decision to connect to the network in response to the tariff information can be used to estimate coverage under alternative tariff scenarios. Pattanayak et al. (2006) review this methodology and illustrate its pros and cons using an example of PSP in water supply in Sri Lanka. Collectively, these examples suggest that information interventions can be an important component of the WSS manager's toolkit.

C. Selected Results

We briefly present results from three types of interventions: CDD-Decentralized Delivery (mostly Latin America), Private Sector Participation (Argentina), and information instruments (mostly South Asia).

C.1 CDD-Decentralization

Rawlings et al. (2004) summarize the impact evaluation of social investment funds (SIF) in water and sanitation. These studies found that household access to improved water services in SIF groups increased by 6 to 21 percent in Armenia, 10 to 18 percent in Bolivia, 5 percent in Honduras, and 22 percent in Nicaragua relative to comparison groups. In all cases surveyed (Armenia, Bolivia, Honduras, and Nicaragua), water investments led to decreases in distance to water source and time spent collecting water. In some countries, health impacts of water supply improvements were detected in terms of reduced child and infant mortality, lost working time, and less stunting. There were no health impacts detected for sewerage system improvements. Appendix III summarizes the methods used in these studies. Each study included comparison communities or households (depending on the level of analysis), but the studies differed in terms of whether measures were taken before and after the investments, and in the estimation methods. These studies' findings suggest the importance of training, user financing, and proper construction and maintenance for effective water supply projects. For sewerage projects, the results indicate the need for better targeting of the poor and for increased connections to improve access.

C.2 Private Sector Participation

Galiani et al. (2005) evaluate the impacts of Argentina's private sector participation in the delivery of water services on access and health. Between 1991 and 1999, Argentina's water systems managed by private operators were servicing about one-third of the country's municipalities and covering almost 60% of the country's population. This study uses historical mortality data for municipalities with and without PSP in the delivery of water services. Using a number of DID approaches, including PSM, the authors find that privatization decreased child mortality rates by 5 to 7 percent. To demonstrate that these health improvements are attributable to improved WSS delivery, they examined if WSS services improved in the municipalities with PSP. Improvements were measured in terms of water production (9.5% increase), water supply (27% increase), sewage drainage volume (21% increase), water leakages repaired (130% increase), and percentage of clients with appropriate water pressure (218% increase), among other things. Municipalities with PSP in the delivery of water services experienced a 2% greater increase in the proportion of households connected to the water network than municipalities that were still publicly managed. These impacts were largest in the poorest municipalities because of the increased access to services that was caused by privatization. This study is a good example of an impact evaluation that looked at both outputs (water production, etc.), outcomes (connection rates), and impacts (mortality rates). Clarke et al. (2004) use household level data from Bolivia, Brazil, and Argentina to evaluate the impact of PSP on access. Control cities are matched to treatment cities (with PSP) on the basis of population size. Using connection rates as their primary indicator, the study finds that access increases in poor and non-poor households in both treatment and control cities. Their results imply that PSP does not increase access.

C.3 Information Campaigns

Household behavior change is an essential complement WSS provision to ensure its effective use. Information, education and communication (IEC) is widely viewed as an alternative to economic or moral incentives for promoting behavior change (see Section II B for a discussion of IEC strategies and Box 4 for an example). One of the attractive features of IEC type instruments is the ability to test its effectiveness through randomized assignment at relatively low costs. Jalan and Somanathan (2004) evaluate an information program in which a random sub-sample of 500 households (from a total of 1,000 households) in New Delhi are informed about the fecal contamination of their drinking water. Seven weeks after the provision of the information, informed households were 11% more likely to begin some form of water treatment compared to the control group. Luby et al. (2004) also use a cluster randomized trial in rural Pakistan to show that IEC campaigns are effective in modifying hygiene behaviors and diarrheal outcomes.

IV. Challenges in WSS Impact Evaluation

A. Multiple Objectives, Interventions, and Outputs

Several features of the WSS sector complicate the design, implementation, and interpretation of impact evaluations. First, WSS programs typically have multiple objectives (*e.g.*, efficiency, equity, sustainability) which are achieved through the implementation of multiple interventions (*e.g.*, PSP in combination with construction and rehanilitation of infrastructure) supporting multiple outputs (*e.g.*, house connections, and toilets). Further, WSS programs are frequently one component of a multi-sectoral program such that WSS service improvements accompany changes in the health, agricultural, education, or other sectors. Careful study design early in the program planning is necessary to disentangle the impacts of each intervention and/or output. As the diversity and combinations of the WSS outputs increase, more sophisticated design (multiple year, multiple measurement) and larger sample sizes are necessary in order to statistically control for all of the combinations and reliably detect changes caused by WSS programs.

For example, Pattanayak et al. (2005a) evaluate a program in the Indian state of Maharasthra that allows communities to select any combination water, sanitation, and hygiene interventions. However, observations from field visits and discussions with local experts indicate that the chosen interventions fall into no more than 4 major intervention clusters. One reason is that the current WSH situation in all project villages is at the same basic level, which is a key criterion in project targeting. Since villages must co-finance the investment costs and pay all the operations and maintenance costs, it is impossible for villages with such basic conditions to develop and pay for sophisticated combinations of WSH strategies. Second, from a practical perspective, all communities are working with a common set of NGOs, who in turn are working within a set of guidelines provided by

the Jalswarajya program. These guidelines, in turn, are being interpreted by a small set of district staff with similar background and training. Thus, all communities are receiving very similar guidance. Finally, in these water scarce districts, there is likely to be a natural sequencing of the interventions: water supply improvements precede other interventions because access to sufficient quantities of flowing water is considered to be a critical input into personal hygiene and sanitation. Collectively, these all point to a narrow and homogeneous set of interventions selected by project villages.

B. Spillover Effects

WSS interventions may be particularly susceptible to spillover effects because water, sewage, and any associated infections flow across the landscape. Furthermore, we might expect information to flow from person to person or across communities (when the intervention is a health education campaign). Or, there may be market spillovers for example from property values. This can impact the evaluation because control communities may be 'contaminated' by the intervention. Thus, the evaluation team must first assess the feasibility of such a spillover through qualitative means and desk research. If this initial research suggests a high probability, then there are two alternatives solutions. If you can control the sample, pick observation units as far away from each other as possible but are still comparable (see Pattanayak et al., 2005b). If you have no control over the sample or think that spillovers will happen in any case, collect information on the pathways by asking about it in the survey (e.g., did you hear about this from your neighbors? What kind of information did you get?). Such information can be used to estimate the extent of spillover in the analysis (see Miguel and Kremer [2004] for the estimation of spillovers in a school-based de-worming program in Kenya).

C. Urban-Rural Differences

There are also some urban-rural differences between types of WSS projects, types of impacts, population, and community characteristics that may affect impact evaluation design. In rural areas, WSS programs are frequently components of multi-sectoral projects, as in the Social Investment Funds programs, or broader initiatives (e.g., integrated watershed management). In these cases, analysts must pay close attention to the design issues described in Sections I and IV A to ensure that the study disentangles the influence of the WSS components from the initiatives in other sectors.

Second, lower population densities in rural areas have some implications for evaluation design and analysis. First, the cost structures of alternative types of WSS services tend to favor community delivery rather than through utilities. This calls for a design that recognized the group/clustered feature of the intervention as in group

randomized trials.[15] In urban areas, expansion of networked water and sewerage services are more attractive. This basically forces a spatial pattern to the evaluation because the network only expands at certain nodes. Households and communities right at the edge of the node could make appropriate control groups for the evaluation.

Third, the fact that rural populations may face thin or missing markets for labor, agricultural product, etc. should be taken into account when selecting indicators. For instance, the household costs of coping with poor water quality may be difficult to monetize because market options (e.g., storage tanks) are less prevalent than in urban areas, intrahousehold reallocations of time and effort (e.g., women and children being assigned non-wage tasks) and the general lack of explicit prices (see Pattanayak et al., 2005 for an example of calculating coping costs).

Finally, the long-term impacts of WSS program may be broader in rural areas than in urban areas. For example, health improvements may be more likely to be caused by WSS programs because access to substitute health inputs (e.g., preventive health care, nutrition, cash income) is more limited in rural areas than in urban areas.

D. Measuring Equity and Sustainability

Equitable access to safe water and adequate sanitation for all members of society, regardless of age, sex and 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 conclusion is to engage in some form of targeting

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 is 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.[16] 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 (as discussed in Section IV C in the case of rural populations).

The sustainability of capital investments for the provision of WSS services as well as the services themselves are critical dimensions of the effectiveness of WSS

[15] In practical terms members of groups share common characteristics such that the overall information is lower in group evaluations. Usually this is accounted for by recognizing that the variance of the impact estimate will be larger – i.e., larger samples are needed to detect small / reasonable size effects.

[16] Pattanayak and Yang (2002) use a similar approach to evaluate the impact of variety of tariff designs and targeting on access to water and economic welfare. The focus is on the distributional incidence of these policies.

programs and policies. There are countless anecdotes and field stories regarding the installation of wells and toilets in communities that subsequently are not maintained and completely abandoned within 2-5 years of installation. 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. One option here would be to build impact evaluation into a follow-on project, or a series of projects.

White's (2005) 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. He 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.

The evaluations of social fund investments in water supply and sanitation also 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, all WSS impact evaluations should adopt this approach in order to build a body of evidence about sustainable WSS projects.

E. A Call for Rigorous Impact Evaluations in the WSS Sector

We have presented guidelines for evaluations in the WSS sector 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 sector. Therefore, we conclude with a call for building a rich repository of empirical examples that consider the at least the three types of policies discussed in Section II for WSS delivery.

We also use this opportunity to speculate on potential reasons for the paucity of rigorous evaluations in the sector. This discussion draws heavily on a recently published essay that assesses the lack of rigorous evaluations of biodiversity conservation policies (Ferraro and Pattanayak, 2006) and adapts it to the case of WSS.

First, one usually needs a remarkable combination of political will, a strong commitment to transparency, and a strong ethic of accountability to conduct a well-designed evaluation. Second, the diversity of donors and practitioners often leads to a plethora of objectives and it might be difficult to get agreement from local stakeholders on a set of explicit objectives and indicators. Third, WSS program staff may be unaware of state-of-the-art empirical program evaluation techniques and the biases in current analyses. Fourth, many believe that rigorous evaluations of effectiveness are expensive and thus would divert scarce funds toward "non-essential" investments. In contrast, researchers and practitioners in other policy fields have demonstrated that randomized experimental methods can be implemented in the context of small pilot programs or policies that are phased in over time. Fifth, many WSS project cycles are short. The benefits of a careful evaluation, however, will largely be realized after the project ends and will accrue to the global community. Sixth, program evaluation methods require data. In other fields of policy analysis, researchers have longstanding national surveys and historical relationships with government agencies and field practitioners that generate substantial datasets for research. Quality and quantity of water, sanitation and hygiene information are typically excluded from these data sets. Finally, on a related point, credible estimates of WSS project success depend on the ability to vary (or isolate) policy interventions in simple ways across space and time. We are well aware that within the same watershed or region, heterogeneity in institutions, income opportunities, access to markets, and other socio-economic characteristics can lead to different responses to a given intervention. However, if every village or household is exposed to a different intervention, we are left with few observations for each intervention and thus cannot make any inferences about effectiveness.

None of these problems are insurmountable. We are not advocating that every WSS program and project be evaluated with an experimental or quasi-experimental design, or that every project collect data on outcomes and covariates from treatment and control units before and after the intervention. Our concern is that there are simply too few applications to WSS, thereby impeding our ability to identify, design and justify effective interventions. Each project that builds in the methods and measurements reviewed in the paper will make a small but vital contribution towards filling the large gap in our knowledge about the effectiveness of WSS investments.

References

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.

Blum, D. and R.G. Feachem. 1983. Measuring the Impact of Water Supply and Sanitation Investments on Diarrhoeal Diseases: Problems of Methodology. International Journal of Epidemiology 12 (3): 357- 365.

Bosch, C., K. Hommann, G. Rubio, C. Sadoff, and L. Travers. 2000. Water and Sanitation. Chapter 23 in A Sourcebook for Poverty Reduction Strategies Volume 2, pages 371-404. Washington, D.C.: World Bank.

Card, D. 2001. Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*. 69:1127.

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

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.

Clarke, G., K. Kosec, and S. Wallsten. 2004. Has Private Participation in Water and Sewerage Improved Coverage? Empirical Evidence from Latin America. World Bank Policy Research Working Paper 3445.

Clasen, T., I. Roberts, T. Rabie, W. Schmidt, and S. Cairncross. 2006. Interventions to improve water quality for preventing diarrhea. *The Cochrane Database of Systematic Reviews*. Issue 3, 2006.

Cooksy, L., P. Gill, and A. Kelly. 2001. The program logic model as an integrative framework for a multimethod evaluation. *Evaluation and Program Planning* 24: 119-128.

Curtis, V., and S. Cairncross. 2003. Effect of washing hands with soap on diarrhea risk in the community, a systematic review. *Lancet Infectious Disease.* 3:275-281.

Dearden, S., L. Pritchett, and J. Brown. 2003. Learning From Neighbors: Social Learning About Child Feeding During Diarrheal Episodes. Unpublished Manuscript.

Drees, F., J. Schwartz, and A. Bakalian. 2004. "Output-based Aid in Water: Lessons in Implementation from a Pilot in Paraguay." Viewpoint No. 270. Washington, D.C.: World Bank.

Duflo, E., and M. Kremer. 2003. Use of Randomization in the Evaluation of Development Effectiveness. Unpublished paper. July 2003.

Esrey, S., J. Potash, L. Roberts, and C. Shiff. 1991. Effects of Improved Water Supply and Sanitation on Ascariasis, Diarrhea, Dracunculiasis, Hookworm Infection, Schistosomiasis, and Trachoma. *WHO Bulletin* 69(5):609-621.

Esrey, S. 1996. Water, Waste and Well-being: A Multi-Country Study. *American Journal of Epidemiology.* 143(6):608-623.

Ferraro, P.J., and S. K. Pattanayak. 2006. "Money for Nothing? A Call for Empirical Evaluation of Biodiversity Conservation Investments". *PLOS Biology* 4(4): e105 (0482-0488).

Fewtrell L. and J.M. Colford. 2004. Water, Sanitation and Hygiene: Interventions and Diarrhoea. Health, Nutrition and Population Discussion Paper. <http://www1.worldbank.org/hnp/Pubs_Discussion/Fewtrell&ColfordJuly2004.pdf>.

Galiani, S., P. Gertler, and E. Schargrodsky. 2005. Water for Life: The Impact of the Privatization of Water Services on Child Mortality. *Journal of Political Economy*. 113: 83-120.

Heckman, J. and S. Navarro-Lazano. 2004. Using Matching, Instrumental Variables, and Control Functions to Estimate Economic Choice Models. *The Review of Economics and Statistics.* 86(1): 30-57.

Heckman, J. and J. Smith. 1995. "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*. 9(2): 85-110.

Hutton, G., and L. Haller. 2004. Evaluation of the Costs and Benefits of Water and Sanitation Improvements at the Global Level. World Health Organization, Geneva.

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., and E. Somanathan. 2004. The Importance of Being Informed: Experimental Evidence on the Demand Environmental Quality. Discussion Paper 04-08, Indian Statistical Institute, Delhi Planning Unit.

Kariuki, M., and J. Schwartz. 2005. Small-Scale Private Service Providers of Water Supply and Electricity: A Review of Incidence, Structure, Pricing and Operating Characteristics. World Bank Policy Research Working Paper 3727.

Kaufmann, R. 2005. Water, Sanitation and Hygiene Interventions for Health – What Works? *Environment Matters. The World Bank Group.* Annual Review. July 2004-June 2005: 24-25.

Kleinau, E., D.F. Pyle, L. Nichols, F. Rosensweig, L. Cogswell and A. Tomasek. 2003. Guidelines for Assessing Hygiene Improvement At Household and Community Levels. Environmental Health Project, Washington, DC. 45 pages.

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

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. *JAMA.* 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

Mohr, L.B. 1988. *Impact Analysis for Program Evaluation*. Chicago, The Dorsey Press.

Millar, A., R. Simeone, and J. Carnevale. 2001. Logic models: a systems tool for performance management. *Evaluation and Program Planning* 24: 73-81.

Mumssen, Y. 2004. "Output-based Aid in Cambodia: Private operators and local communities help deliver water to the poor." OBApproaches No. 01. Washington, D.C.: World Bank.

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, S.K., C. van den Berg, J.C. Yang, and G. Van Houtven. 2006. The Use of Willingness to Pay Experiments: Estimating Demand for Piped water connections in Sri Lanka. World Bank Policy Research Working Paper no. WPS3818.

Pattanayak, S.K., Yang, J.C., Patil, S., Poulos, C., Jones, K., Kleinau, E. Corey, C., and R. Kwok.2005a. Environmental Health Impacts of Water Supply, Sanitation and Hygiene Interventions in Rural Maharashtra, India. Study Protocol, Submitted to The World Bank, January 2005a.

Pattanayak, S.K., Yang, J.C., Patil, S., Poulos, C., Jones, K., Kleinau, E. Corey, C., and R. Kwok.2005b. Environmental Health Impacts of Water Supply, Sanitation and Hygiene Interventions in Rural Orissa, India. Study Protocol, Submitted to The World Bank, January 2005.

Pattanayak, S.K., J.-C. Yang, D. Whittington, and K.C. Bal Kumar. 2005c. "Coping with Unreliable Public Water Supplies: Averting Expenditures by Households in Kathmandu, Nepal." *Water Resources Research* 41, W02012, doi: 10.1029/2003WR002443.

Pattanayak, S.K., and J.-C. Yang. 2002. "Distributional Incidence of Water Tariffs and Subsidies in Kathmandu, Nepal." RTI International, Durham, NC.

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.

Ravallion, M. 2001. The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation. *The World Bank Economic Review* 15 (1): 115-140.

Rawlings, L.B. and N.R. Schady. 2002. Impact Evaluation of Social Funds: An Introduction. *The World Bank Economic Review* 16 (2): 213-217.

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.

Shadish, W., L. Leviton, and T. Cook. 1991. Foundations of Program Evaluation: Theories of Practice. *Sage Publications.*

Shirley, M. M., Colin, X. L., and A. M. Zuluaga. 2000. Reforming urban water supply: the case of Chile. World Bank Policy Research Working Paper No 2294. World Bank, Development Research Group, Regulation and Competition Policy, Washington D.C.

Walker, I., R. Cid, F. Ordonez, and F. Rodriguez. 1999. Ex-Post Evaluation of the Honduran Social Investment Fund (FHIS 2). Produced by ESA Consultants, Honduras, for the World Bank, Latin American and Caribbean Region (LCSHD).

W.K. Kellogg Foundation. 2001. Logic Model Development Guide. <http://www.wkkf.org/Pubs/Tools/Evaluation/Pub3669.pdf>.

Weiss, C.H. 1972. *Evaluation Research: Methods of assessing program effectiveness*. Englewood Cliffs, NJ: Prentice-Hall.

White, H. 2005. Challenges in evaluating development effectiveness. IDS Working Paper 242.

The World Bank Impact Evaluation website. Accessed 2005. http://web.worldbank.org/WBSITE/EXTERNAL/TOPICS/EXTPOVERTY/EXTISPMA/0,,menuPK:384336~pagePK:149018~piPK:149093~theSitePK:384329,00.html

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.

Appendix I. Case Descriptions

CASE 1. JALSWARAJYA PROJECT	
Project Name:	Environmental Health Impacts of Water Supply, Sanitation and Hygiene Interventions in Rural Maharashtra, India
Authors:	Pattanayak, S.K., Yang, J.C., Patil, S., Poulos, C., Jones, K., Kleinau, E. Corey, C., and R. Kwok
Evaluation Period:	2004-2007
Intervention Summary:	Water and sanitation service infrastructure and personal hygiene behaviors are potentially important determinants of health. In rural India, the infrastructure for providing safe water and effective management of human wastes is typically inadequate or completely missing. Consequently, people's options for managing household water and human waste are limited, and their coping strategies often include behaviors that are harmful to their health. The Government of India (GoI) established ambitious goals for providing potable drinking water and reducing infant mortality across the country. In the state of Maharashtra, the World Bank (TWB) supported Jalswarajya Project will assist in meeting these goals. The spirit of Jalswarajya lies in voluntary participation by communities and targeted provision of rural water supply and sanitation services (RWSS) by project administrators.
Evaluation Design Summary:	The evaluation measures whether WSS interventions due to the Jalswarajya Project in Maharashtra cause differences in health outcomes for young children who live in villages that participate in the project. Specifically, it evaluates the extent to which an upgrade in water supply service, improvements in source water quality, reduction in open defecation, improvements in environmental sanitation and information and education on personal hygiene contribute towards improvements in child health outcomes. The evaluation also measures the broader impacts of WSS interventions on rural livelihoods, such as savings in time, materials and money invested in coping activities; improvements in convenience and privacy; and indirect benefits to caregivers (*e.g.*, gains in work efficiency, and time and work reallocation within the household). Villages were selected into Jalswarajya Project based on three criteria: (1) poor quality and quantity of drinking water and sanitation services, (2) high proportion of disadvantaged groups (poverty, and SCST population), and (3) institutional capacity. In order to measure the impact of WSS interventions on child health the program collected baseline and follow-up data from four sources: (1) household surveys; (2) community surveys; (3) institutional surveys; and (4) water quality samples. Propensity score matching was used to pair Jalswarajya villages (treatments) with non-Jalswarajya villages (controls) based on similar observable attributes. The impacts of WSS interventions on child health and income will be calculated using a difference-in-difference estimator to measure differences between treatments and controls.
Evaluation Methods:	Difference-in-Difference Propensity score matching
Final Indicators:	Outcome: (1) Access to water and sanitation infrastructure, (2) Improved health behaviors Impact: (1) Diarrheal rates in children under five, (2) Anthropomorphic measures in children under five, (3) Expenditure levels on water, (4) Income
Main Findings:	Ongoing
Where to Find:	Pattanayak, S.K., Yang, J.C., Patil, S., Poulos, C., Jones, K., Kleinau, E. Corey, C., and R. Kwok. 2005. Environmental Health Impacts of Water Supply, Sanitation and Hygiene Interventions in Rural Maharashtra, India. Study Protocol. Submitted to The World Bank, January 2005.

CASE 2. BOLIVIAN SOCIAL INVESTMENT FUNDS	
Project Name:	An Impact Evaluation of Education, Health, and Water Supply Investments by the Bolivian Social Investment Fund
Authors:	Newman, J.; Pradhan, M.; Rawlings, L.; Ridder, G.; Coa, R.; Evia, J.L.
Evaluation Period:	1992-1998
Intervention Summary:	The Bolivian Social Investment Fund (SIF) was created in 1991 to replace the Emergency Social Fund created in 1986. Its objectives are to improve the coverage and quality of basic services in education, health, water and sanitation. Water and sanitation projects provide either new or improved water and sanitation infrastructure. In addition, trainings on operation and maintenance of infrastructure projects and education on hygiene are often provided. The SIF is a demand-driven institution because it does not initiate projects but responds to outside initiatives by providing co-financing for investments in infrastructure, equipment, and training. The co-financing provided by the SIF generally accounts for approximately 80% of project costs, and the requesting institution provides the remaining 20%. Regional SIF offices assist communities in preparing proposals. The decision on whether to fund a project is made at the SIF central offices in La Paz. The Bolivian SIF was targeted at villages where water and sanitation infrastructure was currently unavailable and where populations were concentrated enough to provide economies of scale.
Evaluation Design Summary:	The evaluation measures whether water and sanitation interventions due to the SIF cause differences in under-age-five mortality for children who live in villages that participate in the project. External to the impacts of this project, the evaluation gathered information on the efficiency, equity, and sustainability of the SIF. The evaluation targeted two regions, the Chaco region and the Resto Rural and constructed the comparison group using propensity score matching to identify similar non-beneficiaries. The data for the evaluation were collected through a baseline survey in 1993 and a follow-up data survey in 1997-1998 in 5 provinces in the Chaco region and 17 provinces in the Resto Rural. Four types of data collection were used: (1) household surveys; (2) facilities surveys; (3) community surveys; and (4) water quality samples. A difference-in-difference estimator was used to calculate changes in the under-five mortality rate in treatment and control villages.
Evaluation Methods:	Difference-in-Difference Propensity score matching
Final Indicators:	Outcome: (1) Access to water and sanitation infrastructure, (2) Availability of water, (3) Improved health behaviors, (4) Diarrhea rates in children under five Impact: (1) Under-age-five mortality rate
Main Findings:	Main changes included a reduction in the distance to water sources and in Resto Rural, a substantial improvement in sanitation facilities. The quality of water did not improve significantly and availability of water was not able to be calculated. Under-age-five mortality was significantly reduced within treatment villages.
Where to Find:	Newman, J., M. Pradhan, L. Rawlings, G. Riddder, R.Coa, and J.L. Evia. 2002. An Impact Evaluation of Education, Health and Water Supply Investments by the Bolivian Social Investment Fund. *The World Bank Economic Review* 16(2): 241-274

Appendix II. 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 intervention

Appendix III. Water and Sanitation Impact Evaluations

Selected Results from Water and Sanitation Impact Evaluations

Study /Authors	Inter-vention	Output Type	Evaluation Design	Statistical Method	Program Results Evaluated
Armenian Social Fund (Chase 2002)	CDD	Water supply system	Quasi-experimental: PSM and pipeline matching	Single difference: with and without	Sanitation, Morbidity, Lost work time due to illness
Bolivian Social Fund (Newman et al. 2002)	CDD	Water supply system Sanitation	Quasi-experimental: PSM	DID	Under 5 mortality, Water access, Quantity and quality of water
Honduran Social Fund (Walker et al. 1999)	CDD	Water supply Sanitation	Quasi-experimental: Pipeline matching	DID, single difference	Water access, Access to toilets, diarrhea
Nicaraguan Social Fund (Pradhan and Rawlings 2002)	CDD	Water supply system Sanitation	Quasi-experimental: PSM	Single difference: with and without	Water access, Access to toilets, Malnutrition, Diarrhea
Jalswarajya, India (Pattanayak 2005a)	CDD	Water supply System, Sanitation Hygiene education	Quasi-experimental: PSM	DID	Water access, Access to toilets, Diarrhea
Orissa, India (Pattanayak 2005b)	CDD	Sanitation	Experimental	DID	Access to toilets, Diarrhea
Argentina (Galiani et al. 2005)	PSP	Water supply and sewerage	Quasi-experimental: PSM	DID	Child mortality
Bolivia, Brazil, and Argentina (Clarke et al. 2004)	PSP	Water supply and sewerage	Quasi-experimental: PSM	DID	Water Access, Access to toilets
India (Jalan and Ravallion 2003)	Public provision	Piped water supply	Quasi-experimental: PSM	Single difference: with and without	Child diarrhea

Appendix IV. Statistical Methods

Irrespective of the design applies, some analysis is necessary to tease out impact estimates – e.g., improvements in quantity and hours of water supply – attributed to the program or policy. Typically, one of the following is employed:

1. Difference in means: This method calculates program impact by comparing the value of the indicator of interest for the recipients and the non-recipients. This method can be used in both experimental and quasi-experimental designs. Difference in means lacks the scientific rigor of DID (see below) because unless the design is foolproof, pre-existing differences are not fully accounted for.

2. Multivariate regression analysis: multivariate regression analysis is the workhorse of analysts attempting to control for observable characteristics that might distinguish participants and non-participants. This method is used to estimate the impact of a program only if all possible reasons why outcomes might differ between the two groups can be controlled for. In the Honduran SIF evaluation, Walker et al. (1999) use multivariate regression to control for differences found between treatment and control groups (constructed through pipeline comparisons). The analysis specified as independent variables those variables where systematic differences existed between treatment and control groups, to control for the impact of such differences on the dependent variables.

3. Difference-in-difference (DID): is a rigorous estimation methods because it uses both baseline and follow-up survey data, and can be used with experimental, quasi-experimental and non-experimental designs. This method estimates impacts by comparing the indicator values between control and treatment groups (first difference) before and after the intervention (second difference). For an experimental design, this estimation is often referred to as a comparison of means since a direct comparison is made between the treatment and control group. Limitations of DID are that it can be expensive and time consuming, since baseline and follow-up data is needed.

DOING IMPACT EVALUATION SERIES

1. Impact Evaluation and the Project Cycle

2. Conducting Quality Impact Evaluations Under Budget, Time and Data Constraints (Joint with IEG)

3. Impact Evaluations for Slum Upgrading Interventions