Of taps and toilets: quasi-experimental protocol for evaluating community-demand-driven projects

Subhrendu K. Pattanayak, Christine Poulos, Jui-Chen Yang, Sumeet R. Patil and Kelly J. Wendland

ABSTRACT

Sustainable and equitable access to safe water and adequate sanitation are widely acknowledged as vital, yet neglected, development goals. Water supply and sanitation (WSS) policies are justified because of the usual efficiency criteria, but also major equity concerns. Yet, to date there are few scientific impact evaluations showing that WSS policies are effective in delivering social welfare outcomes. This lack of an evaluation culture is partly because WSS policies are characterized by diverse mechanisms, broad goals and the increasing importance of decentralized delivery, and partly because programme administrators are unaware of appropriate methods. We describe a protocol for a quasi-experimental evaluation of a community-demand-driven programme for water and sanitation in rural India, which addresses several evaluation challenges. After briefly reviewing policy and implementation issues in the sector, we describe key features of our protocol, including control group identification, pre-post measurement, programme theory, sample sufficiency and robust indicators. At its core, our protocol proposes to combine propensity score matching and difference-in-difference estimation. We conclude by briefly summarizing how quasi-experimental impact evaluations can address key issues in WSS policy design and when such evaluations are needed.

Key words | child health, Maharashtra, programme evaluation, propensity score matching, water and sanitation

Subhrendu K. Pattanayak (corresponding author)
Duke University,
Sanford Institute of Public Policy,
Nicholas Schools of the Environment,
and Global Health Institute,
Durham, NC 27708-0312,
USA
Tel.: +1-919-613-9306
Fax: +1-919-684-9980
E-mail: *subhrendu.pattanayak@duke.edu*

Christine Poulos
Jui-Chen Yang
RTI International,
3040 Cornwallis Road,
Research Triangle Park,
NC 27709-2194,
USA

Sumeet R. Patil
NEERMAN, Sion (E),
Mumbai 400 022,
India

Kelly J. Wendland
University of Wisconsin,
427 Lorch Street,
Madison, WI 53706,
USA

INTRODUCTION

Sustainable and equitable access to safe water and adequate sanitation are widely acknowledged as vital, yet neglected, development goals (UNDP 2006; Hutton & Bartram 2008; JMP 2008). There is growing recognition that environmental health interventions such as water and sanitation (WSS) must be reinstated into mainstream development policies because of the strong linkages to malnutrition and poverty reduction (The Lancet 2008; World Bank 2008*a*,*b*). WSS policies are justified because of the usual efficiency criteria (e.g. health externalities, economies of scale in provision), but also significant equity concerns (e.g. health, time and energy costs borne disproportionately by the poor, women and children). Despite more than US$15 billion invested annually on WSS programmes and projects, to date, we have few rigorous scientific impact evaluations showing that WSS policies are effective in delivering many of the desired outcomes (Fewtrell *et al.* 2005; Poulos *et al.* 2006; Zwane & Kremer 2007). We make this claim because of two commonly used criteria. First, a rigorous scientific evaluation must utilize some mix of control groups, baselines and covariates to establish what would have happened in the absence of the policy (the *counterfactual*) and permit the estimation of impacts. Second, the evaluation should produce evidence on a range of WSS outcomes of scalable policies and programmes, not just pilots.

doi: 10.2166/wh.2009.059

Consider several potential reasons for the lack of rigorous WSS impact evaluations that are not unique to the sector. First, many WSS project implementation cycles are short, but even the immediate impacts of the project will largely be realized after the project ends. The evaluation lessons from these immediate and other long-term impacts will accrue to the global community not to the evaluated project *per se*. Many believe that rigorous evaluations are expensive and err in considering these evaluations as non-essential investments (Ravallion 2007). Thus, one usually needs a remarkable combination of political will and foresight, commitment to transparency and an accountability ethos to conduct well-designed evaluations.

Second, the mechanisms to achieve WSS goals are broad and varied in terms of the types of service (water supply, water quality, sanitation, sewerage and hygiene), the setting (urban, peri-urban, rural) and the typology of delivery (public intervention, private interventions, decentralized delivery, expansion or rehabilitation). While these complex interventions call for carefully designed evaluation studies, most previous evaluations have used inappropriate protocols for measuring impacts.

Third, the breadth of WSS policy effects, which range from greater efficiency in the sector at the national level to improved health at the individual level, raises two challenges. First, the management information system (MIS) in many projects tracks a handful of engineering and fiscal outputs only in project communities, yielding almost no relevant information on programme impacts. Second, most impact evaluations of WSS programmes focus primarily on health or engineering indicators and, therefore, do not collect enough data to evaluate intermediate outcomes (e.g. water quantity and in-house water quality) or benefits (e.g. education, rural livelihoods, gender equity). The latter are crucial for estimating progress towards achieving the key development policy goals of poverty reduction.

Fourth, decentralized and community-level projects—particularly those that are community-demand-driven (CDD)—are an important and growing class of development projects in which communities have direct control over key project decisions, including management of investment funds (Mansuri & Rao 2004). The combination of voluntary participation in self-selected interventions by communities and targeted provision by programme administrators increases the difficulty of identifying an appropriate control group. Early evaluations (Sara & Katz 1998; Isham & Kahkonen 2002) have lacked at least one or more features of rigorous evaluations: control group, large samples, pre- and post-measurement, and specific and sensitive indicators.

Fifth, our personal experience of sectoral capacity building suggests that WSS programme staff are unaware of programme evaluation techniques and of the biases in current analyses. Most rigorous evaluations of large-scale WSS programmes are done 'outside' the sector by evaluators either focusing on health outcomes (e.g. Galiani *et al.* 2005) or considering WSS outcomes as part of a broader development package such as the Social Investment Funds (SIF) (Rawlings *et al.* 2004). Typically, these *effectiveness* evaluations employ quasi-experimental designs because the causal chain is neither short nor simple and external validity of estimates of WSS programmes' treatment effects is a critical objective (Victora *et al.* 2004). Even if randomization was deemed to be ethical, politically acceptable and effective throughout the project duration, it is not a feasible design because many WSS programmes are targeted by programme administrators and/or driven by community demand. Moreover, randomized evaluations may not be appropriate because they would answer only a narrow set of policy questions, limited by institutional constraints and be subject to randomization bias and substitution bias (Heckman & Smith 1995).

This paper presents a quasi-experimental evaluation protocol for addressing the key evaluation challenges in the WSS sector. To make our protocol credible and practical, we use the case of a real life, World Bank funded, CDD water, sanitation and hygiene (WSH) programme in rural Maharashtra, India. The study protocol was approved by the Institutional Review Board at RTI International, an external technical oversight group comprising leading public health agencies (e.g. US Centers for Disease Control and Prevention, the World Bank, UNICEF and Indian Council of Medical Research) and a local steering committee from the Office of the Secretary of Water Supply and Sanitation in Maharashtra. Throughout the design, the evaluation team worked closely with the Government of

Maharashtra. Informed consent was obtained prior to starting surveys with all respondents, including community leaders and individual household members. The programme is broadly representative of decentralized WSS delivery in the developing world.

The remainder of the paper is organized as follows. The next section briefly takes stock of the policy and implementation issues in the WSS sector, and reviews existing impact evaluations. The third section describes the design features of our protocol including control group identification, pre-post measurement, programme theory, sufficient sample and robust indicators. At the core of our protocol is the combination of propensity score matching (PSM) and difference-in-difference (DID) estimation. We focus on PSM, and not the alternative quasi-experimental designs (instrumental variables, IV, regression-discontinuity, RD, and 'control function'), because it is the only feasible design in our setting. The IV method is not appropriate because we cannot identify logical instruments that are highly correlated with the selection process, but not correlated with the health outcome. The RD method is ruled out because it relies on exact knowledge of the selection process, specifically a cutoff score with a sharp discontinuity that separates programme and non-programme villages. We have an imprecise understanding of how socio-economics, performance ability, and the RWSS situation were combined to determine selection and no evidence that local administrators strictly adhered to a sharp threshold (discontinuity). The control function method represent a general form of the well-known Heckman two-stage estimation procedures that attempts to address bias by directly modelling selection into treatment. Compared with PSM, the control function relies on a parametric strategy and therefore makes stronger (and more) assumptions about functional forms. We use 'pre-intervention' PSM to identify a control group, which is a significant deviation from most evaluation studies that rely on post-intervention PSM. To our knowledge, this strategy has been suggested only in a handful of evaluations, all outside the WSS sector (Almus *et al.* 2001; Preisser *et al.* 2003; Ho *et al.* 2007; Sills *et al.* forthcoming). We conclude with a brief summary of how quasi-experimental evaluations address key issues in WSS policy design.

BACKGROUND

This section reviews the impact evaluation literature from the WSS sector in two ways. First, we describe the WSS policy landscape and review studies that provide evidence on the effectiveness by policy or type of delivery. Second, we review the evidence by type of WSS impact. Table 1 provides a summary of recent rigorous evaluations of WSS impacts and illustrates the shallowness of the evidence base.

Evaluations by WSS policies and programmes

Increasingly, donors and aid agencies have broadened their objectives from a narrow focus on physical infrastructure to sustainable service provision. Three types of programme or policy are predominantly used to achieve the broader objectives of improvements in financial viability and institutional performance of the WSS sector: (1) improving operator performance; (2) service provision by the private sector providers (PSP) or small-scale independent providers; and (3) decentralized delivery, typically relying on demand, participation and management by communities.

The first type of sector reform focuses on helping utilities reduce costs and increase revenues to become financially viable, thereby improving and extending service delivery. Despite interesting case studies (e.g. Drees *et al.* 2004), there are no known impact evaluations of these types of reform.

The second type of policy, PSP (large-scale projects in which corporate entities, with private equity, assume operating risk and/or develop under a licence or contract), is increasingly in use. We are aware of one rigorous evaluation of PSP. Galiani *et al.* (2005) evaluate the impacts of Argentina's privatization of water services on access and health using historical mortality data for municipalities with and without privatized water services. Using DID and PSM, the authors find that privatization decreased child mortality rates by 5–7%. A variation of this policy is partnerships between the public and private sector to provide public health services (Buse & Waxman 2001). There are no rigorous evaluations of such partnerships.

With one-third of the population of Africa and Asia living in towns, the third type of policy–decentralized WSS–is fundamental to achieving the Millennium

Table 1 | Selected recent rigorous impact evaluations

			Outputs	Impacts						
Study	**Programme**	**Design and sample size**	**Water quantity**	**Water quality**	**Sanitation**	**Hygiene**	**Health**	**Education**	**Gender/social inclusion**	**Income**
Chase (2002), Armenia	CDD	Pipeline matching and PSM ($N = 5,860$)	Yes	No	Yes	No	Yes	No	No	Yes
Pradhan & Rawlings (2002), Nicaragua	CDD	PSM ($N = 4,040$)	Yes	No	Yes	No	Yes	No	No	No
Newman *et al.* (2002), Bolivia	CDD	PSM and DID ($N = 1,235$)	Yes	Yes	Yes	Yes	Yes	No	No	No
Jalan & Ravallion (2003), India	Public supply	PSM ($N = 33,000$)	Yes	No	No	No	Yes	No	No	No
Lokshin & Yemtsov (2005), Georgia	Public supply	PSM and DID ($N = 2,800$)	Yes	No	No	No	Yes	No	Yes	Yes
Galiani *et al.* (2005), Argentina	PSP	PSM and DID ($N = 40,000$)	Yes	No	No	No	Yes	No	No	No
Jalan & Somanathan (2008), India	Public supply	Randomized and DID ($N = 1,000$)	No	Yes	No	No	No	No	No	No.
Pattanayak *et al.* (forthcoming), India	CDD	Randomized and DID ($N = 1,086$)	No	No	Yes	Yes	Yes	No	Yes	Yes
Kremer *et al.* (2009), Kenya	Public supply	Randomized and DID ($N = 1,354$)	No	Yes	No	Yes	Yes	Yes	No	Yes
This paper, India	CDD	PSM and DID ($N = 10,205$)	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes
Poulos *et al.* (2009), India	PSP	PSM and DID ($N = 2,752$)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Development Goals. In rural areas, CDD often translates into putting the community 'front and centre' of the planning, design, implementation and operations processes, and replacing career bureaucrats with qualified professionals and technocrats in guiding this process. The CDD philosophy is typically captured in reflecting community needs through participatory planning, decentralized delivery, cost sharing (typically 10% of capital and 100% of operations costs) and strengthening local institutions.

Social investment funds (SIFs) exemplify CDD and have been the subject of several impact evaluations, best summarized by applications to WSS in Armenia, Bolivia, Honduras, Nicaragua and Peru (Rawlings *et al.* 2004). While these studies find improved access to water (measured by decreases in distance to water and time spent collecting water), the health impacts of water supply improvements (measured by child and infant mortality, lost working time and less stunting) are measured at only a couple of sites. Furthermore, it has been difficult to establish a causal relationship between any outcome and the participatory elements of such projects.

Finally, across all WSS policies, interventions may be delivered at the community or household level. For instance, public utilities provide services at the community level by increasing the number of sources or the quality of water for all members of the community. Alternatively, some services, particularly water treatment and hygiene education, may be provided at the household level. Indeed, many PSPs provide household level services and CDD provide mainly community level services. While some studies have compared the effectiveness of community-level WSH interventions (e.g. protection or treatment of water sources) with household-level interventions (e.g. point-of-use water treatment), several unanswered questions remain.

In particular, while the most recent systematic reviews of the epidemiologic literature suggest water treatment at the household level is more effective in preventing enteric disease than source water improvements, the effectiveness varies by the setting and some studies have methodological flaws that limit their comparability (Clasen *et al.* 2007). This review finds important differences in the methodological quality of the studies, their design, duration of follow-up, participant compliance, and prevalence of specific pathogens. It concludes that household water treatment is more effective in reducing diarrhoea than community water source treatment in rural settings, although this is not true for peri-urban, urban and refugee settings. In contrast, VanDerslice & Briscoe (1993) show that improving source water is more effective than in-house treatment in decreasing diarrhoea rate.

Outputs, outcomes and impacts in WSS

WSS policy is not an end in itself. It can expand access, improve service quality and ultimately reduce poverty. WSS policies guide decisions on the allocation of financial, social and institutional *resources* that support downstream *activities* and generate *outputs* (i.e. any direct product of programme activities that are directly controlled by providers) in the sector, including the construction, expansion and/or rehabilitation of water supply, sanitation and sewerage infrastructure. Outputs also include hygiene behaviour change campaigns. Activities could include changing incentives to influence the behaviour (e.g. operation and management) of organizations, including utilities, communities and cooperatives. Inputs could also include staff training and improvements in utilities' processes (e.g. appropriate book-keeping practices, improved billing, improved accountability, number of concession contracts awarded).

Thus, policy effects can be classified as *outputs*, *outcomes* or *impacts*. WSS *outputs* are simply the types of product and levels of service under the direct control of programme providers, whether they are public sector, private sector or community organizations (Bosch *et al.* 2000). WSS outputs are typically classified into five categories: (1) water supply (quantity); (2) water quality; (3) individual sanitation; (4) environmental sanitation; and (5) hygiene information and education campaigns.

Programme *outcomes* are the changes in behaviours, knowledge and activity of participants because of the programme. The primary outcome of WSS interventions is access or use. Some indicators measure service availability: for example, average distance from beneficiaries' homes to a water source. Other indicators measure service quality: for example, litres consumed per capita per day, the number of

hours of service and the quality of drinking water. Typically, outcomes are realized at the household or individual levels.

WSS *impacts* are the fundamental gains experienced by beneficiaries as a result of the programme that can be categorized as improvements in: (1) health; (2) education; (3) gender and social inclusion; and (4) income (Bosch *et al.* 2000). While the health impacts are studied most frequently, a small number of observational studies have measured changes in the household's costs of collecting, storing and treating water, as well as the income losses due to water-borne and water-washed illnesses (Pattanayak *et al.* 2005). There have been no rigorous studies of how WSH interventions impact education, gender and social inclusion, and poverty.

Esrey *et al.* (1985), Huttly *et al.* (1997), Curtis & Cairncross (2003) and Fewtrell *et al.* (2005) are some of the best known reviews of the health impacts of WSH interventions. Typical interventions include: (a) *hardware*–infrastructure and services that increase access to improved water quantity and quality, public taps, toilets and drains; and (b) *software* that promotes behavioural changes such as hand washing practices, faeces disposal, and safe handling and storage of drinking water and food. The impact of these interventions has tended to range from approximately 15 to 50% reduction in diarrhoea morbidity. The evaluations control for several covariates including child characteristics (e.g. gender, age and weight, race and immunization), maternal characteristics (e.g. age, education, marital status, pregnancy history, ante- and postnatal medical care, and breastfeeding practices), household socio-economics (e.g. household income, assets, employment, location, health service, family size, electricity, cooking fuel, house size and presence of domestic animals/pets) and community characteristics (e.g. population density, rainfall and location).

Fewtrell *et al.*'s (2005) meta-analysis and our own literature review identified two broad weaknesses in the evaluations. First, in spite of calls for improving research quality, half of the evaluations use inadequate research designs, particularly with respect to: (a) accounting for baseline diarrhoea rates and pre-intervention behaviours; (b) the inclusion of control groups; (c) explicit examination and control for confounders; and (d) detailed reporting and presentation of results. In addition, there is a strong possibility of publication bias.

Second, there have been almost no impact studies of taking these interventions to scale, for example, through decentralized and CDD programmes at national or regional levels. Because scaling up involves including communities with heterogeneous needs and capacity, the research design in existing studies will not be adequate. For example, CDD programmes have a long and complex causal chain, and yet studies rarely gather information about the programme along the entire chain (e.g. using process evaluations or adequacy surveys). In the case of health impacts, for example, typical analyses start with outputs such as in-house water treatment, without regard to how the household acquired such a system.

WSH interventions in India

A survey of water supply and quality conditions in India by McKenzie & Ray (2005) discusses the range of water supply technologies and institutional arrangements, and concludes, as we have, that there are few peer-reviewed studies evaluating drinking water interventions in rural India. They state that previous case studies 'tend to be brief, optimistic, and not easily comparable to one another'.

Jalan & Somanathan (2008) and Pattanayak *et al.* (forthcoming) represent two recent experimental evaluations of the impact of information on behaviours in India. For example, Pattanayak *et al.* (forthcoming) describe a randomized evaluation of an information, education and communication (IEC) campaign in rural Orissa and its impacts on use of individual household latrines (IHLs). They find that the IEC campaign substantially increased IHL uptake by about 30%. Intention-to-treat estimates show that this increase in IHL, in turn, led to a decrease in diarrhoea in children.

Such experimental evaluations of information are exceptions. In general, there have been few studies of the health impacts of public water supplies in India. The completed studies have tended to rely on data from large-scale national surveys (Hughes *et al.* 2001; Wang 2003). The cross-sectional, non-experimental and non-specific nature of national survey data raises challenges for establishing the causal impact of WSH interventions. Nor can these data be used to empirically illustrate the causal chain from WSH programme investments, through activities,

outputs, outcomes and ultimately to impacts. There is some ambiguity about the findings, even when methods are used to account for potential self-selection and targeting biases (e.g. PSM in Jalan & Ravallion 2003). To some extent these results can be explained by the inability to account for water contamination during handling and home storage (Jensen *et al.* 2002), or for insufficient flowing water and therefore hand washing (Curtis & Cairncross 2003). It is also possible that generic surveys cannot reveal real health benefits of public water supply because water supply types are too broad (ignoring issues related to water quality at the point of use, maintenance and proximity of stand posts) and surveys do not collect data on all relevant co-factors (e.g. culture, diet, geography, socio-economic status, governance) that affect the linkages between WSS and health.

APPLICATION

We now turn to a quasi-experimental evaluation protocol that addresses several challenges to measuring impacts of WSS policies. The programme we study is an example of how the Government of India (GoI) is intending to deliver WSS services to reach its 'national planning objectives' related to universal access to water and reductions in child mortality and morbidity. For example, the GoI has set an ambitious target to halve the number of people without access to safe water and basic sanitation by 2015 (GoI 2002), in excess of MDG targets. The state of Maharashtra, with support from the World Bank, has embraced a cross-sectoral, community-driven approach in the Jalswarajya (JS) programme to provide WSS services. Before describing the evaluation design, we present some background on Maharashtra and JS.

Intervention

JS was launched by the Government of Maharashtra (GoM) with support from the World Bank to improve the state's current WSS conditions in rural areas. Derived from GoI's *Swajaldhara* priniciples, the project promotes community-led service provision. Maharashtra is among the largest Indian states, with a population of approximately 100 million living in 400 cities and towns and 44,000 villages. The state is also among the most developed and prosperous in India with a variety of economic activities, relatively high literacy and per capita income, and only about half the population engaged in agriculture.

The mortality rate is 51 per 1,000 in rural infants and 68 per 1,000 in children under five (IIPS & ORC Macro 2001). Eighty-five per cent of households lack sanitation, only 23% have a household water connection, and there is little or no treatment of water in the home. We hypothesize that these poor WSS conditions contribute to the high rate of water-related diseases such as diarrhoea; 23% of children under three suffer from diarrhoea.

JS's main objectives are to increase access to rural drinking water and sanitation services, institutionalize decentralized delivery of WSS services by local governments, and improve rural livelihoods. With resources from the state and district governments, *Panchayati Raj* institutions, national and local organizations, and the World Bank, village residents organize to improve their WSS systems by choosing interventions that best meet their needs and abilities. Villages apply to the state government and are selected into the project based on the three main criteria that they: (a) have poor quality drinking water and sanitation services; (b) have a high proportion of disadvantaged groups; but (c) exhibit sufficient institutional capacity to organize and conduct community activities, such as collecting fees for water supply.

JS is being implemented by the GoM from 2003 to 2009 in approximately 2,800 villages in 26 of the state's 33 districts. This very extensive effort has been designed to address the shortcomings of previous programmes. Four sustainability principles guide the programme, such that communities must:

- submit a comprehensive application package with detailed management plan to participate;
- share the cost of projects by paying 10% of capital costs and 100% of operation and maintenance costs;
- invest in local institutions (e.g. *gram panchayats*) and new ones (e.g. village water and sanitation committees (VWSC) and social audit committees) to improve participation quality;
- take over decision making from government bureaucrats or sectoral technocrats (e.g. NGO staff).

JS is being implemented in five overlapping phases. The pilot phase comprised 30 villages in three districts; Phase I comprised 225 villages in nine districts; the remaining 17 districts will be covered in subsequent phases and batches until the target number of 2,800 villages is met. Our study uses data from a subset of 95 Phase I villages from four districts (Buldana, Nashik, Osmanabad and Sangli).

The intervention begins with selection of the villages and then progresses through three additional stages. First, communities conduct pre-planning and mobilization activities that culminate in the establishment of the VWSC. Second, the VWSC plans interventions, subject to review and approval, and launches implementation, with the social audit committee tracking the procurement, construction and finance. Third, the VWSC establishes ongoing operation and maintenance procedures. It is up to each village to customize its package of activities and outputs. In practical terms, each community is expected to make improvements in all three basic components–water, sanitation and hygiene–with the specific goal of ending the practice of open defecation.

Evaluation design

JS's flow of resources, outputs, outcomes and impacts is shown in the logic model (Figure 1). The hypothesis to be tested in this evaluation is whether the *programme outputs*–the water-sanitation-hygiene packages–will bring about improvements in child health (measured by diarrhoea prevalence and anthropometrics) and overall well-being (e.g. time savings). Naturally, we expect these impacts to be realized through improvements in *programme outcomes*: improved water quantity and quality, more latrines to reduce open defecation, improved hand washing and water handling. The evaluation must rule out any confounding influence of mediating and intervening factors. This outcome is ensured through a combination of study design, sample selection, data collection and proposed analysis that are described next.

Identifying control communities: propensity score 'pre-matching'

We used a combination of restrictions, stratification and matching to reduce sampling bias in the choice of

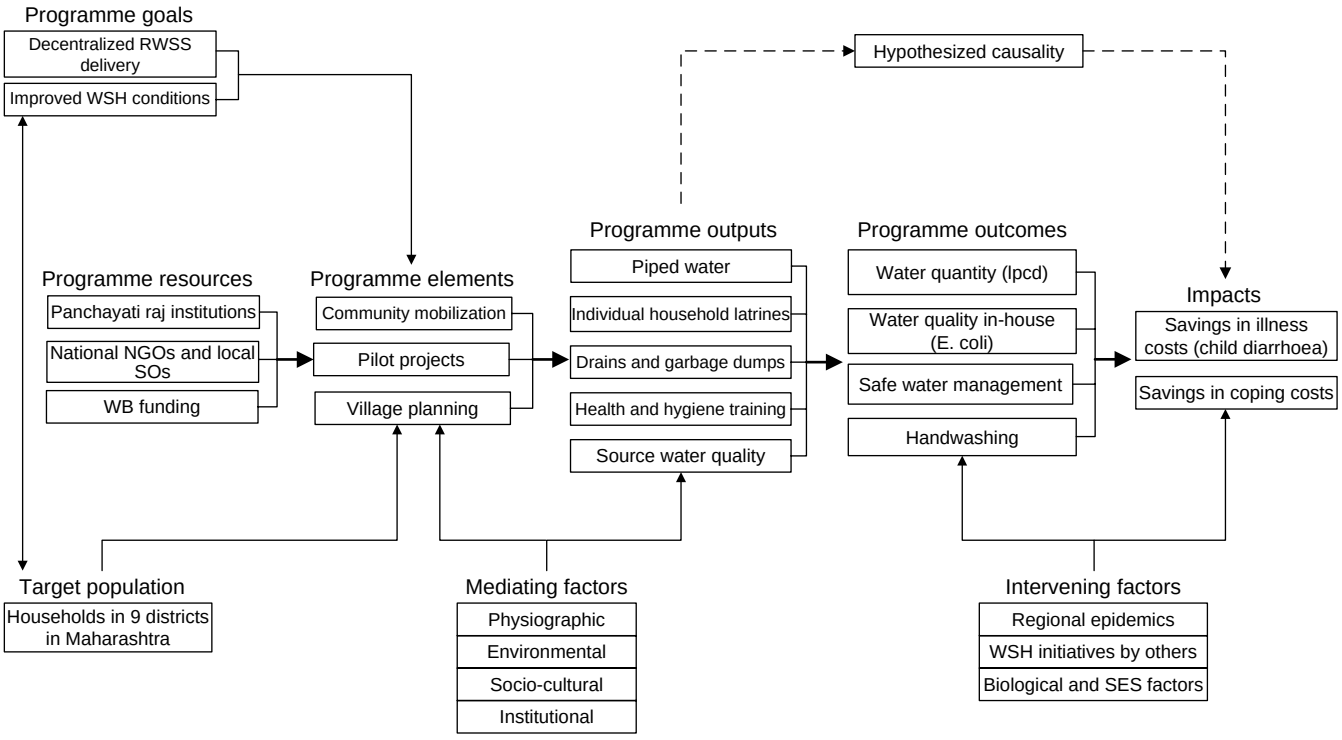

Figure 1 | Logic model for the *Jalswarajya* programme.

communities for our survey. First, we eliminated from the JS districts those which were urban or coastal. This was done to keep the focus on rural and dry or drought-prone villages. One district was chosen from each of four geographically different regions: Osmanabad (Marathwada), Nashik (Mumbai), Sangli (western Maharashtra) and Buldana (Vidarbha).

Second, PSM is used to match each project (treatment) village to two observationally similar non-project (control) villages, from this sample of districts. Unlike previous applications of PSM in data analysis, we used the method in the design stage. Post-intervention PSM is limited by reliance on secondary data for baseline measures, or even lack of baseline data altogether (Lokshin & Yemtsov 2005). PSM was carried out as follows:

1. We estimate logit models of project participation on a pooled sample of project villages and all non-project villages. Our choice of pre-determined variables for propensity score estimation relied on factors that might influence (or proxy for) the three eligibility criteria for being a JS participant (see above) and what was available. We account for district influences on programme and administrative decisions by including a district fixed effect in the estimation of propensity scores.
2. Village level data were drawn from the 2001 and 1991 censuses, and the JS project database. Table 2 lists the variables used in the PSM model, including socio-demographic characteristics and WSS conditions available for 7,200 villages (column 1). For 6,200 of these villages, the 1999 Habitat survey also provides village level water quality data. Thus, a second PSM model was estimated using this data set with fewer observations, but more variables (column 3).
3. Table 2 shows that the two estimated logit models are statistically significant, but only explained about 10% of the variation in the data (pseudo-$R^2 = 0.1$). The pseudo-R^2 improves to 0.35 in the sub-sample of matched project and control villages. The statistically significant variables had appropriate signs; that is, they are consistent with the selection criteria of the JS programme. Models 1 and 2 yielded similar parameter estimates.
4. The predicted probabilities from the logit models provide the propensity score for JS participation, and included a prediction for non-JS villages in our sample. For each JS village, we find a non-JS village with the closest propensity score. We reduced bias by restricting the matches to within the region of common support and using a 5% trimmed distribution of estimated propensity scores, thereby eliminating four JS villages.
5. Furthermore, to account for district fixed effects, we also found a second matched control within the district using a similar process as the one above, but with scores from Model 2. There is some overlap across the two sets of matches because the unrestricted match also identified controls within the district in some cases.
6. Each matching strategy was checked for balance in key covariates across JS and matched non-JS villages and the reduction in average bias (or difference) across all matching variables. Table 3 confirms that these criteria were satisfied for all villages retained for further evaluation, with matching reducing bias by between 7 and 750%. With one exception, matching eliminated statistically significant differences in the means between treatment and control villages. Owing to space limitations, we have not included covariate balance and overall bias reductions for the within-district matching (using model 2 propensity scores).
7. Next, we compared all matched pairs of programme and potential control villages for remaining bias in key variables (e.g. water supply and percentage of socially disadvantaged groups) and eliminated pairs with statistically significant differences. This eliminated 27 villages from the pool. The bias reduction is confirmed with baseline household survey data (see Table 4 and below).

The final sample of 242 villages comprised 95 treatment villages (i.e. 2 pilot villages and 93 Phase I villages) and 147 control villages. The matching process ensures that the villages are comparable in terms of several observable criteria. Household selection in each of these villages is described in the survey implementation section below.

Shadish *et al.* (2002) argue that the success of the matching methodology depends on two critical features: (1) the selection of control groups from a homogeneous pool; and (2) the use of stable and reliable variables. The restriction and stratification processes enable us to meet the first criterion. Because there is a very large sample of

Table 2 | Propensity score estimation of participation in Jalswarajya (standard error in brackets)

Description	Label	Model 1	Model 2
% males in village (2001)	MALES	−5.33 (3.97)	−8.24 (4.36)
% children in village (2001)	UNDER6	0.75 (3.22)	1.56 (3.29)
% scheduled castes in village (2001)	SC	1.56 (0.74)	1.58 (0.80)
% scheduled tribes in village (2001)	ST	1.91 (0.41)	1.99 (0.44)
% female workers in village (2001)	FWORKERS	−0.64 (1.44)	−0.19 (1.59)
% cultivators in village (2001)	CULTIVATORS	−0.84 (0.68)	−0.71 (0.74)
% agricultural labourers in village (2001)	AGLABORS	0.82 (0.71)	0.82 (0.78)
% marginal workers in village (2001)	MRGWORKERS	0.53 (0.54)	0.71 (0.58)
Households in village (2001)	NUMHH	0.00 (0.00)	−0.00 (0.00)
Average household size in village (2001)	HHSIZE	0.31 (0.15)	0.35 (0.16)
Female literacy rate in village (2001)	FLITERACY	0.00 (0.01)	0.01 (0.01)
% permanent houses in block (2001)	PUCCAHOMES	−1.01 (0.77)	−0.87 (0.82)
% households with private tap in block (2001)	TAP	−0.67 (0.99)	0.00 (1.06)
% households without toilets in block (2001)	NOTOILETS	−3.71 (1.69)	−5.57 (1.83)
% households with electricity in block (2001)	ELECTRICITY	0.13 (1.29)	−1.27 (1.46)
% households who use firewood/crop residue/cow dung as cooking fuel in block (2001)	BIOFUEL	−1.39 (1.45)	−1.41 (1.58)
Water supply level (lpcd) in village (1999)	LPCD	−0.04 (0.01)	−0.05 (0.01)
Distance to nearest water in village (1999)	DISTWATER		0.00 (0.03)
Arsenic in village water (1999)	ARSENIC		0.86 (0.73)
Fluorides in village water (1999)	FLUORIDES		0.65 (0.69)
Nitrate in village water (1999)	NITRATE		1.04 (0.73)
Salinity in village water (1999)	SALINITY		−0.43 (1.04)
Odour in village water (1999)	ODOR		0.89 (0.62)
Market facility in village (1991)	MARKET		−1.02 (0.22)
Bus stop in village (1991)	BUSSTOP		−0.11 (0.24)
Railway station in village (1991)	RAILWAY		−1.27 (1.04)
Metalled roads in village (1991)	PUCCAROADS		0.24 (0.21)
Dirt roads in village (1991)	KUTCHAROADS		−0.01 (0.01)
Village area (1991)	AREA		0.00 (0.00)
Buldana District dummy	BULDANA	0.82 (0.33)	0.72 (0.38)
Nashik District dummy	NASHIK	−0.46 (0.41)	−0.08 (0.43)
Osmanabad District dummy	OSMANABAD	0.64 (0.47)	0.87 (0.50)
Sangli District dummy	SANGLI	1.12 (0.36)	1.13 (0.41)
Constant		2.33 (3.61)	7.19 (3.96)
Number of observations		7181	6201
Pseudo R^2		0.08	0.11

Table 3 | Testing covariate balance across treatment and 'matched' control villages using secondary data

	% bias reduced*	t-statistic†
MALES	20	0.29
UNDER6	68	0.74
SC	−749	1.78
ST	79	−0.36
FWORKERS	83	−0.05
CULTIVATORS	9	−1.18
AGLABORS	37	1.55
MRGWORKERS	49	−0.37
NUMHH	98	−0.03
HHSIZE	99	0.03
FLITERACY	55	−1.05
PUCCAHOMES	56	−0.57
TAP	98	0.07
NOTOILETS	66	−0.49
ELECTRICITY	77	0.82
BIOFUEL	85	0.4
LPCD	7	2.33

*Reduction in bias when comparing mean difference between treatment and *unmatched* controls with mean difference between treatment and *matched* controls. Bias is the difference in standardized means between JS and control (non-JS) villages.
†For mean difference between treatment and matched control villages.

non-project villages (>1,500 in each district), we did not run into degrees-of-freedom problems in establishing matches. Further, our use of village level aggregates and multivariate matching improve the stability and reliability of the data used in matching.

Sample size

Required sample size depends on the expected impact, the number of interventions, outcome indicators and control variables, and the unit of analysis. The goal of sample size calculations is to identify the minimum efficient number of observations needed to ensure adequate statistical power. So long as these assumptions are consistent with conditions in the field, the results of the sample size calculations are valid. Because the ultimate impact is on health outcomes that have typically been hardest to change–in particular, child diarrhoea rates–we compute the size of the sample necessary to measure these health effects. Furthermore, we inflate our calculated sample sizes because we are studying group (village) interventions and the primary outcomes are measured on individuals nested within those villages (Blitstein *et al.* 2005).

There is an inverse relationship in the calculations between the number of villages and the number of households required from each village. This relationship parallels the field trade-offs between the number of villages and the number of respondents per village. On the one hand, it is usually more difficult logistically and expensive to include more villages and fewer households per village than it is to sample fewer villages and more respondents per village. While maximizing the number of villages reduces the overall sample size, it also increases the costs of transportation during data collection. Another advantage of sampling more households from fewer villages stems from the fact that, while the intervention takes place at the community level, the decision to use improved services (e.g. a toilet) is made by a household. The proportion of the population and the sample that would be using the intervention at the endline survey is uncertain. Sampling a larger number of households in each village increases the likelihood of interviewing users, which permits investigation of factors affecting usage.

On the other hand, it is usually advantageous to maximize the number of villages and reduce the number of households per village to increase the amount of independent data (which boosts the power of the inference) and distribute the potential bias more evenly across intervention categories. Furthermore, the study is less vulnerable to statistical problems if projects are not completed on time and entire villages drop out of the project.

Individuals' data will be correlated to an unknown degree because respondents within the same village will have shared histories and common experiences that make them more alike to each other than they are to respondents in another village. This correlation introduces a component of random variation that is attributable to the village over and above the random variation associated with the individual respondents. Correlation within a village is usually expressed as the intracluster correlation coefficient (ICC).

The ICC can be expressed as the ratio of σ_c^2 and $\sigma_c^2 + \sigma_e^2$, where σ_c^2 indicates village-level variation and σ_e^2 indicates household-level variation. The ICC is the critical factor in the design effect (DEFF), which describes the

magnitude of additional variation found in a village random sampling relative to a study that employed household random sampling. The DEFF is expressed as DEFF $= 1 + (m - 1)$ ICC, where m is the number of households per village. If the ICC is very small and the number of households per village is also small, the design effect (DEFF) would be close to 1, indicating little additional variation. Increases in either of these factors will increase the DEFF and study level variation.

Ultimately, sample size estimation involves a number of parameters and assumptions, including: (a) the type I and type II error rates; (b) the anticipated effect of the intervention, often referred to as the effect size estimate; and (c) the anticipated ICC.

Type I error or an α error is the error of rejecting a null hypothesis when it is actually true. In other words, this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when the results can be attributed to chance. For example, we think that JS caused a reduction in child diarrhoea, when in reality there is no correlation between project outputs and child health. Type II error or the β error is the error of accepting a null hypothesis when the alternative hypothesis is the true state of nature. In other words, this is the error of failing to observe a difference when in truth there is one (i.e. JS has an impact).

We set the type I error rate at 0.10, and the type II error rate at 0.20 to provide a test of the intervention effect with 80% power to identify statistically meaningful differences between intervention conditions. Further, we intend to employ a two-tailed test when we assess the effect of the intervention. This conservative strategy places a heavier onus on the evaluation, but allows us to observe intervention effects that are not in the desired direction.

The anticipated intervention effect size is determined through a review of the literature. Based on Fewtrell *et al.*'s (2005) meta-analysis of the health impacts of similar interventions, we assume an estimated effect size of 30% (the approximate mid-point of their range of effectiveness). When an effect size estimate is based on a percentage change, it is important to understand and incorporate information on the baseline diarrhoea rates in the study population. The National Family Health Survey-II data (IIPS & ORC Macro 2001) indicates a child diarrhoea rate of 22% in rural Maharashtra. We also obtain our estimate of ICC from the literature. Katz *et al.* (1993) examined the clustering of diarrhoea rates at the village-level in several developing countries to estimate the DEFF. With cluster sizes standardized to 50 households, the DEFF ranged from 1.38 to 4.73. This suggests that the ICCs are in the range of 0.008 to 0.076. Hence, we used an ICC of 0.05, a conservative estimate within this range.

Our previous work in the region suggests that we can expect 10% loss to follow-up or non-compliance. Given this set of parameters, our sample size calculations indicate that sampling approximately 50 households with children five years of age or younger in each village will generate a sample with 80% power to detect an intervention impact of 30% or greater in a population with a baseline diarrhoea prevalence of 22%. This implies we need an overall sample of 3,000 individuals per intervention or a total of 9,000 individuals to evaluate the three potential interventions (as matched pairs of treatment and control) that villages may adopt under JS.

It is very important to note that all these calculations are based on best available information and buffered by a number of reasonable assumptions to help us protect the desired goals regarding statistical power and the planned tests of intervention effectiveness. For example, we incorporate conservative assumptions regarding the reduction in study level variation associated with taking repeated measures on respondents and villages. Further, we include covariates related to the outcome to further reduce random variation. These factors can improve statistical power and their place in the final evaluation will help protect our analysis in the event that our parameters are very different from their assumed values (e.g. if the diarrhoea prevalence rate is lower than the estimated 22%, the sample size would be too small).

Survey design

The cornerstone of the study is high quality measurement of key biological, socio-economic, cultural and environmental indicators. Quality is attained by careful design and field testing of the survey instruments, rigorous training of the field enumerators and supervisors, and checking and verification efforts in the field and at the data entry stage (Wassenich 2007). Collectively, such efforts can consume

as much as 9–12 months for a study of this scale. These indicators were measured at the individual, household and community levels.

We designed the household and community survey questionnaires based on survey instruments we had developed previously, literature reviews of WSH studies and advice from local advisers. Preliminary versions of the questionnaires were reviewed in focus group discussions with selected individuals, key informants and households. The questionnaires were revised and pre-tested in the field before they were finalized.

The *household* questionnaires were designed to collect data on outputs, outcomes and impacts. O*utputs* and *outcomes* include water, sanitation or hygiene interventions. *Impact* indicators include child health as measured principally by diarrhoea among children under five. A child was classified as having diarrhoea if, during the two weeks prior to the survey, a household carer reported that the child had had three or more loose stools in a 24 hour period. Data were collected on a range of individual covariates (e.g. sex, age, class, caste, religion) and household variables (e.g. family size and composition, education, housing conditions, asset holdings, occupation and expenditures, services, sanitation practices, water storage and treatment practices). Household indicators allow us to compute economic impacts in terms of coping costs (Pattanayak *et al.* 2005) and illness costs (Poulos *et al.* 2008) associated with prevention and mitigation activities.

The *community* questionnaire was administered to key informants (e.g. village heads, governing council members, etc.) to collect information on infrastructure (e.g. roads, electricity, drains, dumps, water sources, credit availability and markets) and main programmes (governmental and non-governmental). Finally, *water samples* were collected from a subset of community sources and household storage containers and tested for microbial contamination.

All questionnaires were field-tested and the enumerators were trained through lectures, role-plays and field practice. A senior field manager was in charge of all survey teams and the development of field routing plans. The manager was supported by four field executives who directly supervised the enumerator teams (each consisting of seven enumerators, a person responsible for the water survey and water sample collection, and a supervisor).

To ensure the project was progressing sufficiently well to be evaluated, an adequacy assessment was conducted from March to August 2006. This qualitative assessment used focus group discussions, interviews and village visits to gather the views of key stakeholders. First, project management and monitoring and evaluation data were collected from district WSS committees, and qualitative insights on the implementation process were gleaned from the district staff during this exercise. Second, rapid rural appraisals were conducted in all project villages by interviewing key informants and observing field conditions. These adequacy assessments triangulated supply-side data (from project MIS) with qualitative and quantitative demand-side information (from community participants) to confirm slow but steady project progress.

Survey implementation

Baseline data were collected in two phases in 2005, before the monsoon (May–June) and after (August–September). The same sample was surveyed in both seasons. Households were selected before the start of the baseline surveys in May 2005 using two steps. In each of the selected villages, we first listed and mapped all households and then identified those with at least one child under five years of age. Because there were no pre-existing data on households with children under five, house-to-house visits had to be made to identify them. Finally, in villages with 50 or fewer eligible households, all were interviewed. In villages with more than 50, a random sub-sample of 50 was interviewed. If a household was not available for an interview, then it was visited the same day or the next day at different times for up to three follow-up visits. If a household was not found or if an interview was refused, then the household was replaced with another household from the village list.

Confirmation of balancing act

This baseline data provides an opportunity to evaluate the central premise of our evaluation design: that PSM reduced differences between treatment and matched control villages in the baseline. A comparison of means in Table 4 reveals no statistically significant differences between treatment and control villages in a number of indicators including

Table 4 | Testing balance across treatment and control villages using baseline survey data

Covariate of interest	Treatment mean	Control mean	z-value*
% under 5 children with diarrhoea	11	10	1.62
% under 5 children with acute respiratory infections	21	22	−0.71
% households using private tap	18	24	−1.55
% households using private toilet	13	10	0.96
# of critical times a caregiver washes hands	2.3	2.4	−0.51
# of critical times a child washes hands	1.1	1.2	−0.44
% households treating drinking water	64	63	0.11
% households stating roads are 'main problem'	19	21	−0.84
% households stating water supply is 'main problem'	54	42	3.27
% households stating sanitation is 'main problem'	11	14	−1.72
% households stating public well water quality is bad	19	24	−1.77
% households stating public tap water quality is bad	24	22	0.44
% households stating village water-sanitation committee (VWSC) is active	20	12	2.71
% households participating in VWSC	5	3	3.35

*For mean differences after adjusting standard errors to account for clustering at the village level.

health outcomes, WSS conditions, personal hygiene behaviours, and perceptions of local health and environmental problems. However, because the baseline survey was conducted after early sensitization in treatment villages, these villages were somewhat different in exposure to public health messages, self-reported identification of the main problem and community participation. Thus, the double difference strategy for analysis will be critical to account for any baseline differences.

Data analysis plan

The pre-post data collection plan positions us to use a DID estimator (also called double-difference or first difference) and measure the 'treatment effect' by comparing the treatment and control units before and after the intervention (Heckman *et al.* 1998). The DID estimate is the mean difference in the *change* in the outcome across the intervention and control groups. That is, we can difference the outcome values for the intervention and their matched control units at post-intervention levels and then subtract any pre-existing differences in outcome values:

$$\mathrm{DID} = \{E[Y_{1t}|p(X)] - E[Y_{1c}|p(X)]\} - \{E[Y_{0t}|p(X)] - E[Y_{0c}|p(X)]\} \quad (1)$$

where Y is the outcome with subscript 1 and 0 for post-treatment and pre-treatment levels, and subscripts t and c for intervention and control unit outcomes. E is the expectations operator suggesting that this is the expected treatment effect across all treatment units. It is conditional on the propensity score of participation, $p(X)$, which depends on all relevant covariates (X) included in the first stage estimation.

DID estimators are often implemented in a regression framework by including an interaction variable for the study condition and for the treatment period. The estimated coefficient on the interaction term measures the pre-to-post change in the outcome for the affected households relative to pre-to-post change in the outcome for the unaffected households. The overall plan accounts for observable differences (by matching) as well as time-invariant unobservable differences (by differencing) between treatment and control households (Heckman *et al.* 1998). Note that the bias due to time-variant unobservable factors is likely to be negligible because the pre-treatment and post-treatment surveys are conducted within a short time period. Further control on time-invariant unobservables will be realized because control group members were drawn from very similar villages, based on the intervention probabilities estimated in the sample selection stage.

CONCLUSIONS

This paper has used a real WSS project in Maharashtra, India, to propose a quasi-experimental research protocol that can address several challenges to rigorous impact evaluation of WSS policies.

First, the combination of self-selection by communities and targeting by programme administration increases the threat of self-selection bias. This implies that control groups may be difficult to identify and the characteristics of intervention groups are correlated with both their exposure to the intervention and the outcomes that are realized. Biases of these kinds can be minimized by quasi-experimental designs that employ PSM and DID. The example reported in the paper shows how pre-intervention PSM also improves the quality of the controls and baselines.

Because we estimate a statistical model of participation in a CDD project, the first stage of the matching strategy allows us to uncover some of the political economy factors that influence self-selection and targeting (Bardhan & Mookherjee 2000). We find that the project included communities that had more socially marginal sub-populations and poorer sanitation and water conditions. Project villages are also more likely to: (a) have larger size households; (b) be located in districts with a proactive administration; and (c) be more remote (e.g. no market facilities or unpaved roads). While these results are all suggestive of targeting, it is difficult to rule out self-selection because many of these covariates would also increase community demand. Additionally, the inclusion of poor and socially marginal villages might also suggest delivery of lower quality public goods; richer communities can simply afford better quality toilets, taps and drains and therefore not participate in JS.

Second, WSS policies are complex. They have multiple objectives, use inputs from multiple sectors, provide a variety of services using different delivery mechanisms, and generate effects in multiple sectors. Our example illustrates that randomizing such a complex mix of project components would be difficult to implement and monitor, and impossible if communities are allowed to tailor project components. In contrast, we should use quasi-experimental methods that carefully select control groups, build in sufficiently large samples, conduct pre-post measurements and use appropriately analyses.

Third, the breadth of effects of WSS policies, which range from greater efficiency in the sector at the national level to improved health at the individual level, requires careful data collection and indicator selection. The application shows how the collection of pre-intervention and post-intervention data on communities, households and individuals, as well as programme implementation using a variety of qualitative and quantitative methods can be used to triangulate our measures of impacts and controls. This should permit a thorough understanding of the programme as well as measurement of a broad range of impacts, not just health.

A deeper understanding of the programme is a critical first step towards 'opening the black box of the conditional mean impact' by recognizing heterogeneity in programme delivery, acceptance and impacts (Ravallion 2007). For example, if communities had chosen and received different WSS packages at the baseline, we could have estimated a multinomial model and matched the multiple interventions (Lechner 2002). Similarly, we can estimate heterogeneous intervention effects via analysis of sub-groups, for example, by poverty status, caste or mother's education (Jalan & Ravallion 2003).

Fourth, we are not advocating that every WSS project be evaluated with an experimental or quasi-experimental design, or that every project collects data on outcomes and covariates from treatment and control units before and after the intervention. Instead, we contend that there are simply too few impact evaluations in the WSS sector, which impedes our ability to identify, design and justify effective interventions.

Ferraro & Pattanayak (2006) suggest a number of criteria to determine whether an impact evaluation is required, including the costs and benefits of the impact evaluation. Impact evaluations are most justifiable when basic process issues have been worked out (e.g. how to deliver the treatment), good behavioural theories give ambiguous predictions and the counterfactual is not clear at all *ex ante* (i.e. lots of potential confounders). In addition, evaluations will be beneficial when the project is innovative (e.g. new technology, delivery mechanisms, or institutional restructuring); is scalable, replicable, and

likely to be expanded to other settings; involves substantial resource allocations; and has well-defined interventions. In particular, they are most needed when the intervention has become popular despite a weak evidence base.

Typically, evaluations can verify whether scarce funds are being used prudently for a fraction of the programme cost (between 1 and 5%). Considering that a portion of these costs would be incurred for monitoring and evaluation and on project MIS, impact evaluation costs can be quite low. The main incremental costs of impact evaluation include tracking an expanded set of indicators and data collection in control groups. Controls may already be built into the programme implementation as a result of over-subscription (e.g. waiting lists) and or attrition (e.g. drop-outs).

Finally, when the methodology is disseminated properly, the second and third generation evaluations can be conducted at a sizeably lower cost, as shown by our own follow-on evaluations in other parts of India and in El Salvador. Once these protocols are well established, local programme administrators and collaborators can apply the lessons learned to future programme evaluations without involving costly international experts. In sum, each project that builds in elements of the protocol advocated in this paper will make a small but vital contribution towards filling the large gap in our knowledge about the effectiveness of WSS investments.

ACKNOWLEDGEMENTS

Participants at seminars at the Institute of Economic Growth (Delhi), Center for Excellence in Health Promotion Economics, Government of Maharashtra, University of California (Berkeley), the World Bank, and a joint workshop by the Indian Council for Medical Research & the US Centers for Disease Control provided helpful comments. We are grateful to two anonymous referees, Ben Arnold, Peter Berman, Gene Brantley, Jeremy Bray, Aline Coudouel, Maureen Cropper, Jack Colford, David Evans, Markus Goldstein, Jeff Hammer, Eckhard Kleinau, Priti Kumar, Kseniya Lvovsky and Caroline Van den Berg for many helpful discussions.

REFERENCES

Almus, M., Lechner, M., Pfeiffer, F. & Spengler, H. 2001 The impact of non-profit temping agencies on individual labour market success in the West German State of Rhineland-Palatinate. In *Econometric Evaluation of Labour Market Policies* (ed. M. Lechner & F. Pfeiffer), pp. 211–242. Physica-Verlag, Heidelberg.

Bardhan, P. & Mookherjee, D. 2000 Capture and governance at local and national levels. *Am. Econ. Rev.* **90**(2), 135–139.

Blitstein, J. L., Murray, D. M., Hannan, P. J. & Shadish, W. R. 2005 Increasing the degrees of freedom in future group randomized trials: the df* approach. *Eval. Rev.* **29**(3), 268–286.

Bosch, C., Hommann, K., Rubio, G., Sadoff, C. & Travers, L. 2000 Water and sanitation. In *In A Sourcebook for Poverty Reduction Strategies*. World Bank, Washington, DC, pp. 371–404.

Buse, K. & Waxman, A. 2001 Public-private partnerships: a strategy for WHO. *Bull. World Health Organ.* **79**(8), 748–754.

Chase, R. S. 2002 Supporting communities in transition: the impact of the Armenian social investment fund. *World Bank Econ. Rev.* **16**(2), 219–240.

Clasen, T., Schmidt, W.-P., Rabie, T., Roberts, I. & Cairncross, S. 2007 Interventions to improve water quality for preventing diarrhoea: systematic review and meta-analysis. *Brit. Med. J.* **334**(7597), 782.

Curtis, V. & Cairncross, S. 2003 Effect of washing hands with soap on diarrhea risk in the community, a systematic review. *Lancet Infect. Dis.* **3**, 275–281.

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

Esrey, S. A., Feachem, R. G. & Hughes, J. M. 1985 Interventions for the control of diarrhoeal diseases among young children: improving water supplies and excreta disposal facilities. *Bull. World Health Organ.* **63**, 757–772.

Ferraro, P. J. & Pattanayak, S. K. 2006 Money for nothing? A call for empirical evaluation of biodiversity conservation investments. *PLOS Biol.* **4**(4), 482–488.

Fewtrell, L., Kaufmann, R. B., Kay, D., Enanoria, W., Haller, L. & Colford, J. M. 2005 Water, sanitation, and hygiene interventions to reduce diarrhoea in less developed countries: a systematic review and meta-analysis. *Lancet Infect. Dis.* **5**(1), 42–52.

Galiani, S., Gertler, P. & Schargrodsky, E. 2005 Water for life: the impact of the privatization of water services on child mortality. *J. Pol. Econ.* **113**, 83–120.

GoI (Government of India) 2002 *Tenth Five Year Plan (2002–2007)*, Planning Commission. Available at: http://planningcommission.nic.in/plans/planrel/fiveyr/10th/default.htm, (accessed 20 November 2008).

Heckman, J. & Smith, J. 1995 Assessing the case for social experiments. *J. Econ. Perspect.* **9**(2), 85–110.

Heckman, J., Ichimura, H. & Todd, P. 1998 Matching as an econometric evaluation estimator. *Rev. Econ. Stud.* **65**, 261–294.

Ho, D., Imai, K., King, G. & Stuart, E. 2007 Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Pol. Anal.* **15**, 199–236.

Hughes, G., Lvovsky, K. & Dunleavy, M. 2001 *Environment Health In India: Priorities in Andhra Pradesh*. South Asia Environment and Social Development Unit, World Bank, Washington, DC.

Huttly, S. R., Morris, S. & Pisani, V. 1997 Prevention of diarrhea in young children in developing countries. *Bull World Health Organ.* **75**(2), 163–174.

Hutton, G. & Bartram, J. 2008 Global costs of attaining the millennium development goal for water supply and sanitation. *Bull. World Health Organ.* **86**(1), 13–19.

Isham, J. & Kahkonen, S. 2002 Institutional determinants of the impact of community-based water services: evidence from Sri Lanka and India. *Econ. Dev. Cult. Change* **50**(3), 667–691.

IIPS (International Institute for Population Sciences) & ORC Macro 2001 *National Family Health Survey (NFHS-2), 1998–99: Maharashtra*. IIPS, Mumbai.

Jalan, J. & Somanathan, E. 2008 The importance of being informed: experimental evidence on demand for environmental quality. *J. Dev. Econ.* **87**(1), 14–28.

Jalan, J. & Ravallion, M. 2003 Does piped water reduce diarrhea for children in rural India? *J. Econom.* **112**(1), 153–173.

Jensen, P. K., Ensink, J. H. J., Jayasinghe, G., van der Hoek, W., Cairncross, S. & Dalsgaard, A. 2002 Domestic transmission routes of pathogens: the problem of in-house contamination of drinking water during storage in developing countries. *Trop. Med. Int. Health* **7**(7), 604–609.

JMP (Joint Monitoring Programme for Water Supply and Sanitation) 2008 *Progress on Drinking Water and Sanitation: Special Focus on Sanitation*. UNICEF, New York; WHO, Geneva.

Katz, J., Cary, V. J., Zeger, S. L. & Sommer, A. 1993 Estimation of design effects and diarrhea clustering within households and villages. *Am. J. Epidemiol.* **138**(11), 994–1006.

Kremer, M., Leino, J., Miguel, E. & Zwane, A. 2009 *Spring Cleaning: Rural Water Impacts, Valuation and Institutions*. Working Paper. University of California, Berkeley.

The Lancet 2008 Editorial: keeping sanitation in the international spotlight. *Lancet* **371**(9618), 1045.

Lechner, M. 2002 Program heterogeneity and propensity score matching. *Rev. Econ. Stat.* **84**(2), 205–220.

Lokshin, M. & Yemtsov, R. 2005 Has rural infrastructure rehabilitation in Georgia helped the poor? *World Bank Econ. Rev.* **19**(2), 311–333.

Mansuri, G. & Rao, V. 2004 Community-based and driven development: a critical review. *World Bank Res. Obser.* **19**(1), 1–39.

McKenzie, D. & Ray, I. 2005 *Household Water Delivery Options in Urban and Rural India*, Working Paper no. 224. Stanford Center for International Development, Stanford, California.

Newman, J., Pradhan, M., Rawlings, L. B., Ridder, G., Coa, R. & Evia, J. L. 2002 An impact evaluation of education, health, and water supply investments by the Bolivian investment fund. *World Bank Econ. Rev.* **16**(2), 241–274.

Pattanayak, S. K., Yang, J.-C., Whittington, D. & Bal Kumar, K. C. 2005 Coping with unreliable public water supplies: averting expenditures by households in Kathmandu, Nepal. *Water Resour. Res.* **41**, W02012.

Pattanayak, S. K., Yang, J.-C., Dickinson, K.L., Poulos, C., Patil, S. R. Mallick, R. Blitstein, J. & Praharaj, P. Forthcoming Shame or subsidy revisited: cluster randomized evaluation of social mobilization for sanitation in Orissa, India. *Bull. World Health Organ*.

Poulos, C., Pattanayak, S. K. & Jones, K. 2006 A *Guide to Water and Sanitation Sector Impact Evaluations. Doing Impact Evaluation* Series No. 4. World Bank, Washington, DC.

Poulos, C., Patil, S. R., Yang, J.-C. & Pattanayak, S. K. 2009 *Monitoring and evaluation of health and socio-economic impacts of water and sanitation initiatives*. Working Paper. RTI International, North Carolina.

Poulos, C., Riewpaiboon, A. Stewart, J. F., Clemens, J., Guh, S., Agtini, M., Anh, D. D., Baiqing, D., Bhutta, Z., Sur, D., Whittington, D. & DOMI Typhoid COI Study Group 2008 *Cost of illness due to typhoid fever in study sites in five Asian countries*. Working Paper. RTI International, North Carolina.

Pradhan, M. & Rawlings, L. B. 2002 The impact and targeting of social infrastructure investments: lessons from the Nicaraguan social fund. *World Bank Econ. Rev.* **16**(2), 275–295.

Preisser, J. S., Young, M. L., Zaccaro, D. J. & Wolfson, M. 2003 An integrated population-averaged approach to the design, analysis and sample size determination of cluster-unit trials. *Stat. Med.* **22**, 1235–1254.

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

Ravallion, M. 2007 Evaluating anti-poverty programs. In *Handbook of Development Economics*, **Vol. 4**. (ed. T. P. Schultz & J. Strauss). Elsevier, pp. 3787–3846.

Sara, J. & Katz, T. 1998 Making rural water supply sustainable: Report on the impact of project rules. Water and Sanitation Program, Washington DC, USA.

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

Sills, E., Arriagada, R., Pattanayak, S. K., Ferraro, P., Carrasco, L. & Cordero, S. Forthcoming Private provision of public goods: applying program evaluation to evaluate 'Payments for Ecosystem Services' in Costa Rica. In *Ecomarket: Costa Rica's Experience with Payments for Environmental*

Services (eds. G. Platais & S. Pagiola), World Bank, Washington, DC.

UNDP (United Nations Development Program) 2006 *Beyond Scarcity: Power, Poverty and the Global Water Crisis. Human Development Report, 2006*. UNDP, New York.

VanDerslice, J. & Briscoe, J. 1993 All coliforms are not created equal: a comparison of the effects of water source and in-house water contamination on infantile diarrheal disease. *Water Resour. Res.* **29**(7), 1983–1995.

Victora, C. G., Habitch, J. P. & Bryce, J. 2004 Evidence-based public health: moving beyond randomized trials. *Am. J. Public Health* **94**(3), 400–405.

World Bank 2008*a* *Poverty and the Environment: Understanding Linkages at the Household Level*. World Bank, Washington, DC.

World Bank 2008*b* *Environmental Health and Child Survival: Epidemiology, Economics, Experiences*. World Bank, Washington, DC.

Wang, L. 2003 Determinants of child mortality in LDCs: empirical findings from demographic and health surveys. *Health Policy* **65**(3), 277–299.

Wassenich, P. 2007 *Data for Impact Evaluation. Doing Impact Evaluation Series No. 6*. World Bank, Washington, DC.

Zwane, A. & Kremer, M. 2007 What works in fighting diarrheal diseases in developing countries? A critical review. *World Bank Res. Obser.* **22**(1), 1–24.

First received 27 May 2008; accepted in revised form 13 September 2008. Available online May 2009