

Impact Evaluation and Public Sector Programs in India: What Can We Do *Right Now*?

ELIZABETH MATHEW and JOANNE SPRAGUE

Introduction

The Government of India purports to be committed to impact, having devoted billions of rupees to development programs and built poverty reduction agendas into political campaigns. In 2005, the first Outcome Budget was released in India because it was observed that "there is a need to track not just the intermediate physical 'outputs' that are more readily measurable but the 'outcomes' which are the end objectives," emphasizing the need to focus on meaningful development results.¹ But what do these results look like, and how can we know that they are being achieved?

Development results are most often examined through impact evaluations of the initiatives that are meant to produce them. The concept of results-based development evaluation has existed for a long time, and it is certainly not new to India. As early as 1952, the Program Evaluation Organization (PEO) was instituted by the Planning Commission based on a recognized need for evaluations of public social development programs. Subsequent bodies have been set up to improve and build upon the PEO, including the recent Development Evaluation Advisory Committee in 2004 and Plan Scheme on "Strengthening Evaluation Capacity in Government" in 2005–06.² Evaluations and assessments are commissioned by many levels of government ministries, Members of Parliament, central and state government officials—on a consistent basis. As one evaluator of public schemes

Elizabeth Mathew is an Analyst with the Centre for Development Finance (CDF) at the Institute for Financial Management and Research (IFMR). Her work focuses on strategy, organizational development and impact planning for clients in the development sector.

Joanne Sprague is a Senior Analyst at the CDF at IFMR. Her work focuses on strategy advising for development sector clients, with a focus on social enterprise development and sustainable business models.

notes, "After all, a large amount of public money has gone into the Mission and it befits various groups to try and assess how well it has been spent, resulting in better performance in the future."³

However, this commitment to evaluation of development impact may be more rhetoric than action. The much-lauded Outcome Budget has been criticized both from inside and outside the government for failing to deliver: Montek Ahluwalia, Deputy Chairman of the Planning Commission, recently noted, "The exercise of outcome budget of manipulating existing data did not get the desired results."4 Evaluators of public schemes consistently list limitations of low sample sizes, lack of baselines, lack of control groups, and non-representative sampling, all indicators that the impact evaluations lack adequate planning and resources. The Development Evaluation Society of India (DESI) cites shrinking staff strength and physical infrastructure of monitoring and evaluation institutions, along with a lack of investment in human capital, as further handicaps on evaluations.⁵ The Planning Commission's own Working Group on Strengthening M&E Systems for Social Sector Development Schemes (2001) even lays out glaring gaps in the implementation of evaluations in its 39 recommendations for improvement, from over-generalized and generic objectives, to diversion of monitoring staff to other divisions, to concealment of shortcomings and manipulation of data.

How can this scenario be improved? In this paper, we first draw on a wealth of development evaluation literature to identify an "ideal type" model for public sector impact evaluations that would be both effective and feasible to implement. We then discuss some of the ways in which current evaluations of centrally sponsored schemes (CSS), one of the larger segments of central government development expenditure, fall short of this model, and the extent to which current reform efforts are moving towards it.6 In the next section, we lay out a set of incremental steps that could be taken to improve evaluations in the *current* context, to move them closer to this ideal type. Many of the currently discussed solutions focus on the evaluation system, requiring broad institutional reform, political will, or improved technical capacity, all of which take time. In contrast, these steps are modifications that can be made now, on an individual program level for existing and newly emerging schemes, to improve clarity on results and drive scheme improvement without requiring significant additional funding or institutional changes. To illustrate these steps and their feasibility, the

232 India Review

third section undertakes a thought experiment on how these changes would play out for an example scheme evaluation. We also include a brief forecast of why such steps might be resisted, and provide recommendations of how to overcome resistance to enable near-term improvements to current evaluations being undertaken.

What Do We Mean by Impact and Evaluation?

One challenge in the field of development evaluation is the lack of standard terminology.⁷ Terms such as "impact," "outcomes," "outputs" or "evaluation" are used without much attention to their meaning or differentiation. In this paper we focus on *impact* and *impact evaluation*, as defined as follows.

Impact, in general, refers to the long-term effects of a development intervention.⁸ Impact is distinct from processes or outputs in an intervention—for example, an intervention aimed at achieving greater education levels would measure the number of schools built as an output, but would look for higher literacy levels to measure the program's impact. Further, the literature defines impact as the difference between what actually happened as a result of the implementation of a program, and what would have happened if the program had not been implemented.⁹

Impact evaluation is the process of identifying and measuring the impact (positive or negative) caused by such an intervention.¹⁰ Impact often takes time to become apparent and can be caused by many factors other than one specific program. As a result, much attention is paid to the challenge of assessing *causality* in impact evaluation, or the attribution of impact to interventions. Isolating the effects of an individual program is difficult and has led to innovative evaluation methodologies such as randomized controlled trials and synthetic comparison groups. We do not advocate any particular method of impact evaluation, but we do believe it is important to evaluate impact in development programs rather than sticking to processes or outputs.

Further, many discussions on impact evaluation in India focus on improving the evaluation system—the institutional structures, interactions, and incentive systems that enable evaluation to happen. While critical, these discussions are outside the scope of this paper. When we refer to impact evaluation, we are concerned with the process of undergoing individual program evaluations, rather than the design of the institutional mechanism for impact evaluation. While this paper includes a brief overview of the institutional evaluation system in India to set the context, our recommendations will be limited to individual program evaluations, from planning for evaluation at the time of program inception through measuring the program's ex-post residual impact.

An Ideal Type for Public Program Impact Evaluations in India

In order to identify the strengths and weaknesses of current evaluations and suggest improvements, there must first be an understanding of the end goal: what a strong, impact-focused evaluation design looks like. The characteristics of "good" impact evaluations have been debated and documented widely.¹¹ One of the factors that make the field of evaluation a challenging one is its dynamism—there is no consensus on the key building blocks to an ideal evaluation. Each program is different, so each evaluation must be designed differently to address the unique aspects of the program. However, it is possible to identify a "type" or set of characteristics that would be present in a robust impact evaluation. These characteristics form a basis on which evaluations can be critiqued and standards can be improved to increase the value of evaluations.

In constructing an ideal type for public sector impact evaluations in India, we draw upon an array of guidelines and step-by-step tools for monitoring and evaluation systems by Rist and Morra Imas, Rist and Kusek, and White, as well as guidelines followed by institutions such as the International Initiative for Impact Evaluation (3ie) and the Network of Networks for Impact Evaluation (NONIE).¹² Based on these sources, an "ideal type" for public scheme evaluations in India would include several components that need to be addressed upfront, while planning to implement a program; further components to be incorporated during the actual evaluation. The ideal type has been displayed accordingly:

Components to be address during initial program set-up:

1. Upfront planning for impact evaluation, including specification of a model for development change and a strategy to evaluate progress towards this change that includes indicators, baselines, targets, and control groups;

- 234 India Review
- 2. Participation of relevant stakeholders in the planning of impact evaluation;
- 3. Appropriate timing of impact evaluation—planning early enough to enable comparisons, measuring late enough to capture lag effects; and
- 4. Adequate funding and appropriately trained evaluators.

Components to be addressed while starting the evaluation itself:

- 5. Independence of evaluators from the implementing body;
- 6. Separation of impact evaluation from other types (such as process, service delivery, fund utilization or outreach) from a budget, time, effort, and expertise standpoint; and
- 7. Adequate and representative sample of targeted beneficiaries and comparison groups.

The importance of having a causal model in mind before conducting any experiment and deriving conclusions is supported by the literature.¹³ This stage also necessitates strategic planning to ensure that the key components of successful implementation of an evaluation, such as defining indicators, measuring baseline data, identifying performance targets, and selecting test and control groups, are incorporated from the beginning. Further, getting relevant stakeholders involved at the planning stage will ensure that a sense of ownership and consensus is built into the evaluation process. To avoid potential biases and ensure honest feedback on how schemes impact their beneficiaries, it is crucial that the program officers who plan and implement schemes are separate from the evaluation experts who perform the evaluations. To get focused feedback on the "impact" of programs, it is also crucial that the design, planning, budget, time allocations, and evaluator choice for such evaluations be distinct from other evaluations. such as those for processes and service delivery. This ensures that the "impact" evaluation gets adequate attention in itself and does not get lumped in with other assessments.¹⁴ Moreover, it is important to devote ample funding and skilled manpower to ensure that these impact evaluations are rigorous and also timed well to enable meaningful evaluation. Finally, to get a complete picture of the scheme's impact on the ground, it is imperative that an adequate and representative sample of beneficiaries, as well as an adequate sample of non-recipients for comparison, is interviewed for feedback.

Shortfalls of Current Scheme Evaluations

Against this backdrop, the strengths and weaknesses of various existing evaluation studies become apparent. We first examine the general practices the PEO implements in all its evaluations, and then explore some individual examples of evaluations of public schemes.

Evaluations sanctioned by the PEO aim to "assess the performance, implementation process, service delivery and impact of schemes."¹⁵ This mandate runs counter to the ideal type from the start; it is not surprising that an evaluation that attempts to tackle so many aspects of a scheme would fall short on some. Impact is a challenging enough area to evaluate on its own, and a lack of adequate resources for evaluations means that funding and manpower must be spread thin across all the areas of assessment, dampening the utility of any given area. This challenge is further exacerbated by a lack of strong field teams, lack of technical capacity in data collection, and lack of capacity building efforts observed by the Planning Commission in several evaluations conducted by the PEO.¹⁶

An evaluation report of the National Rural Health Mission in 2009 provides an example of the potential weaknesses of current program evaluations and is to be commended for its clear articulations of its own shortcomings.¹⁷ Evaluator Kaveri Gill explains, "In many ways, it is too early and too late for certain kinds of impact evaluation of the NRHM," explicitly titling her report an evaluation of service delivery only. The report documents the failure to plan for evaluation by collecting baselines or assigning control groups-"a control area where the mission has been withheld-a political problem with government sponsored schemes in any case-is missing during the initial program design," and also notes that changes in outcome indicators may start to materialize only after a certain time period, making the evaluation ill-timed on both sides. Moreover, it lists the further caveat of small sample due to time and resource restrictions. While commendable for its transparency, the evaluation fails the timing and sample size criteria of the ideal type in addition to the upfront planning criterion.

A 2009 evaluation of the Integrated Child Development Services scheme is less transparent, and mentions the term impact assessment a few times but mostly proves to be a process and implementation evaluation.¹⁸ The report primarily assesses service delivery and recommends ways to improve scheme implementation, exemplifying how such reports often ignore impact evaluation when it is clustered with other assessment types. The only attempt to look at actual impact seems to be the effort to look at learning outcomes such as reading and counting among children who benefited from the pre-school education component of the scheme. However, this study admits that it could not identify a control group, even though the idea of controls is mentioned in the research design stage. This calls into question how much of this positive impact can be attributed to the scheme itself. Thus, the ICDS evaluation lacks the separation of impact evaluation and upfront planning criteria of the ideal type.

A 2009 evaluation of the National Rural Employment Guarantee Scheme (NREGS) contains similar issues but with an interesting twist.¹⁹ The evaluation is intended to assess efficiency in implementation, operational bottlenecks and impact assessment, among other objectives. Like other reports, it fails to separate impact evaluation from other types to give it adequate attention. In its analysis of impact, the report does identify metrics such as income range, expenditure patterns, asset base, and outstanding loans before and after the intervention. However, these metrics are not identified up-front in the scheme brief or guidelines, so there is no guarantee that other evaluations of the same scheme would measure the same indicators. The sample includes no control group, and data comparing metrics before and after implementation seem to be based on one survey with recall questions to understand metrics prior to the intervention, rather than separate pre- and post-intervention surveys. These are small details, but they reveal a strategy for assessing impact that is open to criticism and data quality issues. Moreover, the evaluators note that the scheme was not implemented "fully" or "uniformly" across the districts studied, implying a non-representative sample. They also allude to the improper timing of the attempt to assess impact: "However, the impact time of the scheme is very less, in most districts of survey it is hardly couple of months and the utility of this scheme is not up to the maximum permissible limit of 100 days at the time of survey."20 Though commendable in its efforts to measure impact indicators and understand changes caused by the program, the NREGA report fails on the separation of impact evaluation, upfront planning, proper timing and representative sampling factors of the ideal type.

In general, PEO evaluations tend to fail to meet the ideal type due to a lack of upfront planning and adequate training. Further, evaluations of CSS tend to fall short most often in separating out the evaluation of impact from that of other program components, attaining specific aspects of upfront planning for evaluation, and the criteria of timing, funding and sample size in varying amounts. The criterion that has not been mentioned specifically here is the participation of stakeholders, as most evaluations seem not to mention the involvement of stakeholders one way or the other. Based on this, we can reasonably assume that stakeholders are not often involved in evaluation planning or implementation, so current evaluations likely fail to a large extent on this count as well. The advent of social audits to enable citizens to cross-check government records with realities on the ground, made possible by the Right to Information Act in 2005, has gone a long way in increasing stakeholder participation in evaluation.²¹ However, social audits still stop at outputs, assessing whether funds have been used for their intended purposes, rather than addressing whether development outcomes have been achieved. Therefore, social audits are not as relevant to the discussion of impact evaluation, despite their value in involving stakeholders and improving accountability.

Current Reform Efforts at the Institutional Level

Though addressing issues of the institutional mechanism behind impact evaluation in India is outside the scope of this paper, it is worth a mention to set the context. There are many potential systemic reasons for the aforementioned gaps in the quality of impact evaluations of CSS. One is complexity in design: the central government is responsible for the majority of scheme funding, but it is the states that implement the schemes, leaving both sides unclear on where the responsibilities for accountability, monitoring, and evaluation lie. Mexico's evaluation system solves this problem by mandating evaluation in all public programs, and its exemplary evaluation culture is often cited as a result.²² Another issue is the difficulty of measuring outcomes and impact, which may lead program implementers to shy away from commitments to development outcomes, sticking with more directly observable outputs. In India, the push for impact evaluation is often left to individual champions in the government. However, officials switch positions often and sometimes lack the mandating power to institutionalize evaluation, making it difficult to fill this gap in evaluation culture or sustain improvements that are made.

As evidenced by the stream of incremental evaluation committees, schemes and working groups set up to advise and improve on monitoring

238 India Review

and evaluation of public programs, the Planning Commission and other levels of government recognize these shortcomings. Many of the government's efforts revolve around generic "improvements' or 'reforms," but recent efforts are stepping up the attention to specific aspects of the evaluation system. For example, the President of India called for the establishment of an Independent Evaluation Office as early as June 2009, an initiative moved forward by discussions at a two day conference on evaluation at the Planning Commission in October.²³, ²⁴ Though the architecture of this office has yet to be designed, if implemented properly the office has the potential to fully address the first criterion of our ideal type.

The Planning Commission's October conference also featured calls for improvements that would incorporate enhanced funding, capacity building, and standardization of guidelines. According to the International Initiative for Impact Evaluation (3ie)'s account of the conference, Montek Ahluwalia suggested devoting up to one percent of the Planning Commission's administrative budget, about Rs 850 crore, to evaluation, while participants called for "a focus on capacity building and boost to the professionalization of evaluation."²⁵ Further, Professor Abhijit Sen, Member of the Planning Commission, called on the Commission to set national standards and guidelines to strengthen the quality of evaluations. If implemented properly, these efforts would accommodate another two criteria of the ideal type for Indian public sector evaluations.

These efforts to reform the institutional set-up of the evaluation mechanism are commendable. The extent to which they will be implemented remains to be seen. Also, these discussions do not drill down to the details of evaluation reform: though the need for impact evaluations is mentioned, the need for specific attention to impact evaluations on their own has not been addressed. The details of what standards and guidelines for improved evaluations would look like have not been publicly outlined. The additional criteria for ideal type program evaluations, including stakeholder involvement, appropriate timing, and attention to sample sizes, have not been mentioned in the current calls for reform. As we discuss in the next section, it is possible to get a great deal closer to these improvements without the mass changes being discussed by the Planning Commission, but with a few relatively minor tweaks at the individual program level.

A More Impact-Friendly Approach

The reforms required to bring about an ideal type evaluation system in the Indian government will require the devoted attention of political champions of evaluation, a broad mandate from the Centre, and significant funding increases. In the meantime, rather than continuing to implement ineffective evaluations for the sake of continuity, small modifications can be made to improve individual evaluations within the current system. We advocate a set of these incremental steps to improve current evaluations of centrally sponsored schemes. While not ideal, and not necessarily experimentally robust, these standards will enable more clarity and better communication of progress with minimal additional burden on budgets, capacity needs, or institutional systems. They will enable better insight into impact and recommendations to improve performance, moving the current standard for evaluation quality closer to our ideal type immediately rather than waiting for sweeping systemic reform to take place.

The incremental steps we suggest to improve scheme evaluations *right now* include:

- 1. Development of a theory of impact for the scheme;
- 2. Identification of a standard set of indicators for outcomes and impacts;
- 3. Specification of data sources and baselines (where possible) for measuring outcomes and impacts;
- 4. Use of a comparison or control group; and
- 5. Mixed methods in data collection.

We use the example of the Pradhan Mantri Gram Sadak Yojana (PMGSY), a 100% centrally sponsored scheme intended to provide connectivity to all remote habitations in India via all-weather surface roads, to illustrate how these recommendations might be applied. The scheme, originally intended to be completed by 2008, aims to construct .37 million kilometers of new rural roads, connecting 0.179 million habitations, and upgrade .368 million kilometers of existing roads. The estimated cost of the scheme as of 2007 is Rs.13.2 billion.²⁶ The scheme PMGSY provides a good example for a thought experiment on evaluating public schemes because it has a wide array of publicly available information and data through its websites (www.pmgsy.org,

www.pmgsyonline.nic.in) and several mid-term assessments of the scheme that have been conducted previously.

As all development schemes are different, the specific recommendations made related to PMGSY may not be replicable for all CSS. The use of PMGSY provides an unusually transparent and data-rich scheme. Rural roads are tangible and measurable; schemes aimed at improving education quality or children's health may be harder to measure. It also may be more difficult to create consensus on the terms of evaluation in schemes that are funded by multiple sources. However, many of these recommendations can be applied individually, and these are merely incremental steps meant to encourage public programs to move towards more valuable, robust evaluations in the near term.

Step 1: Theory of Impact

It is important to clearly lay out the objectives of the scheme from the very outset—what is to be accomplished and why it is important in the broader spectrum of development, in addition to the specific project plans. From these objectives, it is then possible to lay out an assumed causal pathway describing how the project activities are likely to lead to them. This sort of pathway has been referred to as a logical framework, impact value chain, theory of change, theory-based approach, and other terminology in different evaluation methodologies, each with their own nuances. The literature is not standard on the best way to describe the theory of impact.²⁷ We define four general steps from activities to goal realization:



Few public schemes lay out the purposes of their projects in a manner that can be effectively evaluated. In the case of PMGSY, the scheme has clearly laid out the *activities* it plans to engage in and the *outputs* it intends to produce—funds to be disbursed and proposals to be sanctioned, and roads to be built and upgraded. It has even alluded to the reasons this is important, but has stopped short of laying out specific *outcomes* or *impacts* that it expects to result from its activities. The explanation of development goals expected from the construction of rural roads is limited to one sentence in the introduction of the scheme:

Rural Road Connectivity is not only a key component of Rural Development by promoting access to economic and social services and thereby generating increased agricultural incomes and productive employment opportunities in India, it is also as a result, a key ingredient in ensuring sustainable poverty reduction.²⁸

An "impact assessment" commissioned by the Ministry of Rural Development (MoRD) in 2004 described outcomes in similarly vague terms:

The rationale...is thus, to redress this situation [of poor connectivity] so that certain opportunities and services (employment, educational, health, transport, marketing facilities, etc.), which are not available in the unconnected habitation, become available to the residents.²⁹

These broad, de-linked statements, rather than clearly laid out outcomes and impacts, keep the scheme from being held accountable for such outcomes. A PMGSY causal pathway constructed based on the scheme document would be incomplete, including only activities (allocate funds and construct roads) and outputs (roads built and habitations connected). This, in turn, reduces the pressure to be accountable to the scheme's goals, decreasing incentives to invest in roads that have the highest impact or collect the data necessary to evaluate their results.

A better causal pathway lays out the outcomes and impacts of the scheme as part of the initial planning and guidelines, so that the intended results are distinctly clear and can be monitored for progress as the scheme is implemented. It also ensures that proper care is taken to construct these intended results based on a logical thought process and incorporating opinions from stakeholders, experts, and academic literature.

One potential pathway is based on the broad goals laid out in the scheme guidelines combined with past literature on rural roads. The scheme guidelines focus on the availability or access provided by the existence of roads-in particular, access to social and economic services. According to a study by Fan et al (1999), rural roads create development impact through the following pathways: 1) agricultural productivity, 2) nonfarm employment opportunities, and 3) rural agricultural wages.³⁰ With the exception of employment opportunities, it is difficult to link these objectives clearly to the increased access that rural roads provide. It is logical that agricultural productivity would be enhanced by the access roads provide to fertilizers, equipment, and new technologies. Agricultural wages could increase for many reasons: higher profits due to more cost-efficient access to inputs and better access to markets or to compete with emerging employment opportunities elsewhere. In general, economic development results from rural roads seem to revolve around three themes: 1) access to employment (agricultural or non-agricultural), 2) access to inputs (new methods and technologies in particular), and 3) access to markets for outputs (trade centers). The scheme also aims to affect "access to social services," which is not addressed in Fan's study. His 2004 literature review, however, mentions causal links to social outcomes in a number of studies.³¹ These outcomes include access to healthcare and education.

These results resonate with the broadly laid out objectives of the scheme, as well as the World Bank's objectives in its rural roads project which has supplemented the PMGSY funding. In the words of the Bank's Senior Transport Specialist Piers Vickers: "First, the proposed rural road improvements will contribute to the development of economic activity, thus providing more opportunities for employment, trade, specialization, and growth within the rural economy. Second, improved rural roads will provide better physical access to basic services, in particular healthcare and education, and thereby increase the quality of life of the poor in the project- influenced areas."³²

The impacts that would then, in theory, follow from these outcomes, based on the scheme objectives and the logic followed in the literature, would include 1) improved health, 2) improved education, and 3) increased wealth / reduction of poverty. Based on this analysis, a more robust theory of impact might look something like this:



While these socioeconomic impacts of rural road development are supported by global studies and the scheme's objectives, it is equally important to gain stakeholder consensus on the anticipated consequences of the project, through a participatory planning process to develop the impact theory. Thus, the aforementioned causal pathway documented is just an example. The theory should be at least distributed, at best co-created, with all relevant participants to ensure all potential positive and negative effects are documented.³³ In the case of this scheme which is far-reaching and involves complex layers of fund disbursement, this would mean involving the local and state government bodies taking part in the program, beneficiaries in connected and unconnected habitations, and partner organizations/implementers in the revision and finalization of the impact theory. Ideally, it would require feedback from a variety of states and habitations throughout the country to account for differences in geographical contexts.³⁴ While these stakeholder consultations may involve considerable time, effort, and resources, they constitute a valuable step towards building a sense of ownership among relevant stakeholders to ensure accurate and actionable results. In the case that resources are not readily available to be allocated to the task, consultations could be held on a relatively small scale with only highly involved stakeholders in the near term, laying the foundation for more robust stakeholder involvement in the future.

244 India Review

It is also worth noting that for different schemes, the logical assumptions associated with the theory of impact will be stronger or weaker at different points, affecting the robustness of the theory. With PMGSY, the assumptions that activities will create outputs (that constructing roads will cause more habitations to be connected), and that outputs will create outcomes (more habitations connected will increase access to various resources), are relatively straightforward. The assumption that links outcomes to goals (access will lead to improvements in development goals like health, education and wealth), on the other hand, is a bit more ambiguous. However, because the largest assumption comes at the very end of the logical chain, the theory of impact is reasonably robust: an evaluation can measure the steps and assume causality between them all the way to the outcomes. In schemes with less tangible outputs or outcomes, larger or more ambiguous assumptions may come up sooner in the logical process. Take for example a CSS such as Sarva Shiksha Abhiyan (SSA), targeted at creating universal elementary education through building and strengthening schools. One can logically assume that building more schools (activity) will improve access to education (output). The outcome here would be that more primary school aged children would go to school, but other factors such as family obligations or social marginalization could pose problems. Here, the largest logical assumption comes between outputs and outcomes, rather than between outcomes and goals. Thus, the concreteness and measurability of the theory of impact is less robust than in the PMGSY example, making SSA a more challenging scheme to evaluate. This is an unavoidable reality of the challenges of development evaluation, and does not decrease the value of generating a theory of change. It is something to keep in mind when applying this tool to other public schemes.

Step 2: Indicators for Outcomes and Impacts

The second step of the process is to make this impact theory tangible by applying concrete indicators to each level of the causal chain. The idea of indicators is no new concept, and has been put in place by the most skeletal of schemes at an activity and output level. Traditional indicators in development often list the funds utilized and "=outreach" numbers of the people touched. It is necessary to reach beyond these indicators to the results of such outreach, applying metrics to each of the outcomes and impacts listed in the impact theory. At a first look, it appears that the evaluations of PMGSY to date do not assign indicators to outcomes at all. On a closer look at the impact assessment of 2004, some state-specific reports outline a set of indicators in their introductions. For instance, the state report for Assam introduced the section on "impact on agriculture" in this manner note the indicators in bold:

A well knit rural road network can have a lot of benefits on the economy of the area and its impact on the agricultural economy can be very effective in terms of providing easy and faster access during all seasons to the markets, availability of quality inputs and increase in diversity to more cash crops. The farmers in the village can produce more agricultural product by way of generated incentives for marketing their produce at low transportation cost and importing the fertilizer, manures and seeds at cheaper rate due to the reduced cost of transportation to the village... All weather road connection also reduces the time of transporting the perisbable products like milk, vegetables, poultry products, etc. and ensures a remunerative price for such commodities throughout the year to the producer and a regular supply to the consumers.³⁵

Though this is a positive effort to outline indicators, it would have been far more useful if they had been laid out upfront, in the summaries and consolidated report, rather than being hidden in state-specific sections of the report. The fact that the indicators are buried in this manner signifies how little attention is paid to them in the assessment of the scheme's impact. It is also not clear whether the indicators have been chosen based on previous research, input from stakeholders, or simply the gut feeling of the Ministry of Rural Development.

A better indicator design involves a thoughtful process of development and refinement. At a minimum, indicators should be constructed that are clear (precise), relevant to the subject, economic (cost-effective to obtain), adequate (sufficient to benchmark performance against), and monitorable (can be validated independently).³⁶ Kusek and Rist note that some development projects use predesigned indicators from well-known sources such as the Millennium Development Goals or the UNDP's Sustainable Human Development goals, or choose proxy indicators based on externally available data.³⁷ While there are advantages of cost and cross-study aggregation to using these indicators, they are created for different purposes and carry the risk of being less relevant to the specific project. For project performance improvement we advocate the construction of indicators based on the context of the scheme, to provide the most accurate reflections of development progress. As with the impact theory, a participatory stakeholder approach, to the extent possible, is the best way to ensure that appropriate indicators are constructed.

It is also generally accepted that a multitude of indicators rather than one holistic indicator leads to a more robust methodology, as data nuances can cause a limited number of indicators to mask unintended effects.³⁸ At the same time, using too many indicators can lead to information overload and keep the results from being clear and actionable. A balance of two to seven indicators per outcome tends to be the ideal.³⁹

A preliminary indicator set for the theory of impact for PMGSY, would look something like Table 1. To construct this list we compiled the indicators most commonly used to measure health, education, trade, employment, and income outcomes in studies on rural roads, other impact evaluations, and secondary data sources in India and applied logical reasoning to ensure the relevance to the scheme and comprehensiveness within categories. The sources or rationale for each indicator are documented in the chart.

This list of indicators is meant to be a preliminary suggestion, not a finalized set. As with the impact theory, the next step required is to discuss this preliminary list with the involved stakeholders as listed previously, as well as experts in the respective fields of education, health, trade, technology, and employment. As Kusek and Rist note, indicator construction is an iterative process that takes time to finalize, since the interests of several stakeholders need to be kept in mind.⁴⁰

Step 3: Data Sources and Baselines for Outcomes and Impacts

The choice of data sources for indirect outcomes and impacts of a development scheme depends on data availability, frequency, and depth. The depth of the unit of data collection determines how rigorously it can be analyzed for comparison—say, districts or blocks where roads have been constructed vs. those where they have not been constructed. Frequency determines the extent to which outcomes can be

nent lls ed thcare

TABLE 1

(Continued)

	Rationale/source of choice of indicator	Author analysis PMGSY Impact Assessment 2004 NRRDA Assessment 2009 - Madhya Pradesh (One minus Out of School rates) Fan 2004	ASER (One minus Out of School rates) DISE	ASEK DISE	PMGSY Impact Assessment 2004 Fan 2004 ASER NFHS	PMGSY Impact Assessment 2004 - West Bengal NCERT learning achievement survevs	PMGSY Impact Ässessment 2004 NRRDA Socioeconomic Impact Assessment 2009 Fan 2004 ASFR	PMGSY Impact Assessment 2004 World Bank OED Impact Evaluation Report – Morocco 1996 ADR OFD Burnl Roade Study, 2006	ADD OED Rural Roads Study 2006 ADD OED Rural Roads Study 2006 Author analysis	PMGSY Impact Assessment 2004 Fan 2004 Author analysis
TABLE 1 (Continued)	Indicator	Time to reach public school with relevant grade level School enrollment		Girl child enrollment ratio	School attendance – number of days per year	Percentage of students taking private tuition	Teacher attendance – number of days per year	Time to reach nearest trading center	Number of visits to trading center per month	Transport cost per kg of trade items
	Component	Access to education						Access to trade centers		
								Access to economic services		

World Bank OED Impact Evaluation Report – Morocco 1996 PMGSY Impact Assessment 2004 - Assam Author analysis	 PMGSY Impact Assessment 2004 - Assam NRRDA Assessment 2009 - Madhya Pradesh Author analysis 	PMGSY Impact Assessment 2004 – Assam	Author analysis NRRDA Assessment 2009 – Madhya Pradesh Author analysis	PMGSY Impact Assessment 2004	Author analysis	NRRDA Impact Assessment 2009 – Madhya Pradesh National Health Profile	NRRDA Impact Assessment 2009 – Madhya Pradesh National Health Profile	PMGSY Impact Assessment 2004 – Assam	National Health Profile NRRDA Assessment 2009 – Madhya Pradesh NFHS	National Health Profile NRRDA Assessment 2009 – Madhya Pradesh National Health Profile	Author analysis National Health Profile PMGSY Impact Assessment 2004 – Orissa MICS	(Continued)
Use of fertilizers and other agricultural inputs – kg per household per month	Crop patterns - ratio of cash crops to subsistence crops	Use of motorized agricultural equipment – number of households with motorized equipment	Value (Rs) of crop output per year	Use of mechanized nonfarm work equipment – number of households with mechanized equipment		Percentage of deliveries that are institutional	Percentage of children immunised	Incidences of major diseases / illness such as malaria and diarrhea	Infant mortality rate	Maternal mortality rate	Life expectancy at birth Child height and weight at three years	
Access to improved methods and technology						Health						
					Impact	Social						

		TABLE 1 (Continued)	
	Component	Indicator	Rationale/source of choice of indicator
	Education	Reading comprehension level – ability to read letters, words, sentences	ASER
		Arithmetic level - ability to recognize numbers, add/ subtract	INCERT learning acmevement surveys ASER
		Literacy rate	NCERT learning achievement surveys HDI Census of India
Economic	Income / Poverty Reduction	Average monthly income per household	DJAE BPL data BPL data NSSO consumption expenditure surveys*
		Percentage of people living below poverty line	CMIE* Indicus* NRRDA Impact Assessment 2009 - Madhya Pradesh Khandker et al 2006 BPL data
		Means of livelihood - agricultural to non-agricultural ratio	NSSO consumption expenditure surveys* Author analysis
*Different sources	s choose slightly different n	aetrics to calculate household income and poverty estimate	s; the debate over these calculations is beyond the scope

of this article. For monthly income, BPL uses average monthly income per household, NSSÓ uses monthly per capita consumption expenditure, CMIE and Indicus use annual household income. For poverty, NSSO constructs estimates from the monthly per capita consumption expenditure.

analyzed across time, for pre-post analysis and lag effects. Data sources can be primary or secondary, though most project evaluations involve at least a component of primary data collection, to ensure data is gathered on the direct beneficiaries (and non-beneficiaries) of the project.

Previous evaluations of the impact of PMGSY have used small sample surveys for primary data collection only, using secondary sources only for output metrics.⁴¹ However, secondary data lower costs and are available retrospectively, which makes them particularly worthwhile to consider when attempting evaluation of current schemes for which upfront planning has not taken place. The availability and quality of secondary data in India, particularly at a unit level deep enough to enable evaluation of rural roads, is notoriously limited and must be used cautiously. Rigorous national surveys and censuses exist, but they are conducted at irregular intervals and many do not have a large enough sample size to be statistically significant below the state level. Other data sources are newer and therefore cannot serve as a baseline for a ten year old scheme.

Table 2 describes the availability and usability of secondary data sources for the relevant PMGSY outcome categories in India. A closer look at the table reveals that at least one data source in each category is available at the district level. While not an ideal substitute for primary data or road level data, these data can provide a low-cost proxy. Combining the measures captured in these data sources with the indicators laid out in the previous section, we suggest using the ASER survey and DISE District Report Cards for "access to education" outcomes-school enrollment, girl child enrollment ratio, and school attendance. For heath impacts, the DLHS data on institutional deliveries can be used while the ASER and DISE data sources can be used to capture education impacts-reading comprehension, arithmetic level, and literacy rate. For income/poverty reduction impacts, the BPL and NSSO consumption expenditure survey can be used to capture data on average monthly income per household, percentage of people below the poverty line, and agricultural to non-agricultural ratio.

For the other indicators, secondary data is either unavailable or contains too small a sample to estimate indicators at a district level, so primary data collection is required. These include the indicators for access to healthcare outcomes, half of access to education outcomes, access to trade centers outcome, access to improved methods and technology outcomes, access to employment opportunities outcomes, and most of

Outcome category	Data source	Unit	Frequency	Dates
Health	DLHS (District Level Household Survey)	District	~5 years	1999, 2003
	NFHS*	State	~6 years	1993, 1998, 2006
	MICS*	State	~5 years	1995, 2000, 2006
	Health Information of India	State	Annual	1986-2005
	National Health Profile (replaced Health Information of India)	State	Annual	2005–2008
Education	ASER	District	Annual	2005-2008
	DISE (District Information Sys- tem for Education)	District	Annual	2003–2008
	NCERT learning achievement surveys	State	~5–10 years	1990, 2000, 2005
Employment	Census	District	10 years	1951-2001
* *	NSSO Employment- Unemployment	State	~5 years	1973–2005
	NSSO Employment- Unemployment - thin	State	Annual	1990–2006
	NSSO Unorganized manufacturing sector - thin	State	~5 years	1979
Poverty reduction	BPL data	District	~5 years	2002, 2007
	NSSO Consumer Expenditure - thick	State	~5 years	1991–2005
	NCAER MISH (Market Information Survey of Households)	State	"more or less annual"	1986

 TABLE 2

 SECONDARY DATA SOURCES AVAILABLE FOR OUTCOME INDICATORS

*Also contains data on education.

health impacts. The full list of data source recommendations for the proposed indicators is shown in Table 3.

In order to determine whether a scheme has improved the outcomes it hopes to affect, it is necessary to understand the state of the indicators of those outcomes prior to the scheme's implementation. The best way to do this is to conduct a baseline assessment. The PMGSY has presumably done this for the number of habitations and km of roads it intends to construct, but it is unclear whether a baseline was determined for the outcomes and impacts that PMGSY is intended to improve. Obviously, it is too late to conduct a new baseline survey for PMGSY. However, it is possible to construct a proxy baseline based on secondary data sources collected around the time the scheme started. As highlighted previously, there exist adequate secondary data sources available for a considerable time period to construct such a baseline (except for outcomes of access to healthcare, access to trade centers and access to improved methods and technology).

		DATA SOUNCES FON CONSTRUCT	TED INDICATORS		
	Component	Indicator	Data source	Availability: Frequency	Availability: Unit
Inputs	Money Road proposals Road sanctions	Funds Km proposed Km sanctioned	pmgsyonline.nic.in NPS report pmgsyonline.nic.in NPS report pmgsyonline.nic.in NPS report	continuous continuous continuous	Road Road Road
Outputs	Roads built Roads upgraded Habitations connected	Km built Km upgraded Habitations connected	pmgsyonline.nic.in NPS report pmgsyonline.nic.in NPS report pmgsyonline.nic.in NPS report	continuous continuous continuous	Road Road Road
Outcomes Access to social services	Access to healthcare	Time to reach healthcare units (PHCs/sub-centres/district hospitals) Number of visits to healthcare units	survey survey		
		per year Number of healthcare personnel per healthcare unit	survey		
		Volume of medicine stock per healthcare unit	survey		
	Access to education	Time to reach public school with relevant orade level	survey		
		School enrollment	ASER, DISE District	Annual	District
		Girl child enrollment ratio	ASER, DISE District Report Cards	Annual	District
					(

TABLE 3 DATA SOURCES FOR CONSTRUCTED INDICATORS (Continued)

	Availability: Unit	District					
	Availability: Frequency	Annual; no base- line possible					
	Data source			~ ~ ~	~		7
		ASER survey survey	survey survey survey	survey	survey	survey	survey
TABLE 3 (Continued)	Indicator	School attendance - number of days per year Percentage of students taking private tuition Teacher attendance - number of days per year	Time to reach nearest trading center Number of visits to trading center per month Transport cost per kg of trade items	Use of fertilizers and other agricultural inputs - kg per household per month Crop patterns - ratio of cash crops to subsistence crops	Use of motorized agricultural equipment - number of households with notorized equipment	v aute (Ks) of crop output per year Use of mechanized nonfarm work equipment - number of households with mechanized equipment	Value (Ks) of nontarm output per year
	Component		Access to trade centers	Access to improved methods and technology			
			Access to economic services				

	Access to employment opportunities	Monthly employment hours	survey		
	· · · · · · · · · · · · · · · · · · ·	Percentage of population employed (formal and informal)	Census of India, survey	10 years	District
		Wage rate for labor in village - Rs per day	survey		
		Number of people entering village for employment from outside per month	survey		
Impact					
Social	Health	Percentage of deliveries that are institutional	DLHS	~5 years	District
		Percentage of children immunised	survey		
		Incidences of major diseases / illness such as malaria and diarrhea	survey		
		Infant mortality rate	survey		
		Maternal mortality rate	survey		
		Life expectancy at birth	survey		
		Child height and weight at three years	survey		
	Education	Reading comprehension level - ability	ASER	Annual; no base-	District
		v ith starts, words, sentences	A SED	A	
		Arithmetic level - ability to recognize numbers, add/subtract	AJEK	Annual; no base- line possible	District
		Literacy rate	DISE District Report Cards,	Annual; 10 years	District
			Census		
Economic	Income / Poverty Reduction	Average monthly income per household	BPL data, NSSO consumption expenditure	~5 years	District/State
		Percentage of people living below	BPL data, NSSO consumption	~5 years	District/State
		poverty line	expenditure		
		Means of livelihood - agricultural to	BPL data	~5 years	Block/District
		non-agricultural ratio			

256 India Review

The PMGSY was initiated in 2001. The suggested data source for employment outcomes is the Census of India, which was last conducted over 1999–00. For health impact, the relevant data source is the District Level Household Survey, which was taken in 1999, providing a baseline not long before the start of the scheme. The data sources for income and poverty reduction are the Below Poverty Line survey, taken in 2002, and the NSSO Consumer Expenditure 55th round (thick), conducted in 1999–00.⁴² These dates are close enough to the start of the scheme, and the intended effects are long term enough, that these can be taken as baselines as well. Education is a bit trickier: ASER is a new study, initiated in 2005, and the DISE district report cards were collected in 2003. While ASER seems like a bit of a stretch, it is feasible to use the 2003 DISE data for a stretch baseline data source.

Step 4: Comparison or Control Groups

To really evaluate the impact a scheme has had, it is necessary to isolate the scheme from any other "white noise" that may be contributing to development outcomes. Evaluators are increasingly making use of methods that involve a counterfactual, which questions what would have happened if the intervention had been absent.⁴³ Counterfactual evaluations require the use of a control or comparison group, which is similar to the group that received the intervention in all respects other than the fact that it was denied the intervention. This helps to strengthen the evidence that a change occurred as a result of an intervention rather than coincidentally, thereby strengthening the argument for causality and validating recommendations for program improvement.

The most rigorous construction of a counterfactual is the designation of a randomized control group. Of the target population, certain members are randomly selected to receive the intervention (the test group) while others do not (the control group). Stratified random sampling techniques are often used to ensure the test and control groups are sufficiently comparable, particularly in settings with a large degree of variance across the target population such as, say, the nation of India. While this technique is the strongest method to generate similar test and control groups and reduce the risk of factors besides the intervention leading to differences between the groups, it requires setting up randomly selected test and control groups from the start of the program. For a scheme such as PMGSY that has been ongoing for eight years, this is obviously not possible. Other issues of logistics, cost, and ethical questions as to the fairness of randomly providing some citizens with public services while denying them to others make randomized control groups both physically and politically infeasible for most centrally sponsored schemes.

Another method of enabling comparison is to construct a "synthetic" control group using what evaluators call quasi-experimental designs: observing groups that seem to be similar in all aspects besides the treatment, but cannot be sure because they were not randomly assigned. A multitude of quasi-experimental design methods have been developed to suit a variety of contexts and research constraints. These designs are valuable in the context of centrally sponsored schemes because they enable the construction of comparison groups even after a scheme has been implemented. One circumstance in which comparison groups can be easily constructed is in the case of "natural experiments." In natural experiments, a non-recipient group emerges not as a result of the design of the intervention, but by accident. Common examples of this scenario include natural disasters that split up groups of otherwise similar people, or natural human tendencies such as that of families with two children of the same sex to try for a third child more often than those with two children of opposite sex.⁴⁴

To strengthen the degree of comparability between synthetically created treatment and control groups, it is possible to look at the degree to which these groups are similar in other ways. One increasingly popular method to decrease the chance of selection bias is propensity score matching. With this econometric technique, data on other characteristics besides the test variable are collected and analyzed to create one-to-one "matches" of treated and non-treated units that are similar to each other in as many other characteristics as possible. The treatment and control groups are gradually constructed through this one-to-one matching, and then compared to each other for differences in development outcomes before and after the intervention. This method provides a midway point between the ideal of randomized control groups and the need to construct comparison groups after the program has been started.

Most impact evaluations of PMGSY to date have not incorporated control groups of any kind. However, a 2009 assessment of the Rural Roads Project in Madhya Pradesh and Chhattisgarh,⁴⁵ a program of PMGSY undertaken by the National Rural Roads Development Authority (NRRDA) and funded by the Asian Development Bank (ADB), involved the synthetic construction of treatment and control groups. This methodology is in line with another ADB study estimating the attribution of poverty impacts of rural roads in Indonesia, Sri Lanka and the Philippines.⁴⁶ The NRRDA evaluation, with a few minor enhancements, is a good starting point.

The NRRDA impact assessment control group methodology is as follows:

The present project resembles a case of ex-post selection process where the "benefited" and "control" groups were not formed through experimental design, rather they are selected after the projects were identified or were being implemented. Thus the nonrandom method is more suitable for the present analysis where the "control" group resembles the "benefit" group on the basis of some observed characteristics. In the present analysis, "population served" by a road is considered as the characteristics for resemblance between the "control" roads and "project" roads.⁴⁷

"Benefit," or treatment, roads were identified as those where rural roads had been constructed or upgraded under the first batch of ADB funding. "Control" roads were areas where road construction or upgradation was not included in the first batch, nor likely to be taken up during the study period. Though there may be conflating factors that determined why certain roads were included in the first batch and others were not, considering the complexity in the process of proposals, approvals and fund disbursement for PMGSY roads, it is reasonable to assume some degree of similarity between first batch roads and non-first batch roads. Thus, the NRRDA framework can be thought of loosely as a natural experiment.

The NRRDA study used one metric, population served, to generate resemblance between project and control roads. The process of propensity score matching with multiple variables could strengthen this process to increase the similarities between the project and control groups to more thoroughly isolate the variable of road construction.

To extrapolate this modified NRRDA approach to the PMGSY scheme as a whole, it is possible to look at the various phases of PMGSY implementation. Time-series data on the habitations that were unconnected at the beginning of the scheme, tracked by which have and have not been connected during each subsequent phase, are available in the "Habitation Coverage" datasets on the PMGSY website.⁴⁸ Other information provided for both connected and unconnected habitations includes total population, SC/ST population, presence of primary education and health services, electrification, and telephone connection status. While these data may not be sufficient to generate ideal propensity scores for matching, they are an improvement upon the single factor used in the NRRDA study and provide a useful starting point. To further mitigate the risks presented by constructing comparison groups on sub-optimal propensity scores, it is useful to also incorporate the collection of data on intervening factors, not just direct indicators of progress, into the primary data collection phase of the evaluation.

After the treatment and control groups have been constructed, of course, the leg work lies in collecting outcome data for each of these matches. These groups provide a starting point for the populations that should be interviewed for primary data collection as we have discussed in the previous section.⁴⁹

Step 5: Mixed Methods in Data Collection

Data collection can include gathering quantitative or qualitative data, or a combination of both. Quantitative approaches, according to Rist and Morra Imas, are appropriate when evaluators are interested in statistical analysis, require precision, are clear about what to measure, and have a large data sample.⁵⁰ Qualitative approaches are best when detailed or descriptive information is needed, quantification of results is not essential, or the metrics that can be measured are unclear.

Hawkins advocates a mixed-methods approach, which uses both quantitative and qualitative data, when an evaluator wants to deeply understand the context of why an intervention did or did not work.⁵¹ He states that mixed methods are also useful to validate information coming from different sources or in the presence of resource constraints resulting in low sample sizes, such as a lack of adequate time or funding.

Impact evaluations of PMGSY to date have varied in their data collection methods. The 2009 NRRDA impact assessment in Madhya Pradesh employed a robust mixed-methods approach, using six different survey instruments to collect data. Quantitative data was collected on traffic statistics and transport use patterns, village outcomes, and household asset changes, supplemented by qualitative data on village perceptions of program impact, as summarized in the Table 4.

Sl.No.	Survey instrument	Purpose
1.	Classified Traffic Census Count Surveys	To record change in traffic volumes, composition, etc.
2.	Transport User's Survey	To know and record the patterns of transport use
3.	Villagers' Perceptions-Village Focus Group	To identify villagers perceptions of expected and actual socio-economic and poverty reduction impacts, and record significant events and changes identified by villagers
4.	Village Primary Data (Key Informant)	To collect primary data on key indicators of impact
5.	Village Primary Data (Community Self-Monitoring)	To identify and document indicators that are especially relevant to village life
6.	Change Process	To identify the process of change associated with the project road improvements and its impact on the households

TABLE 4SURVEY INSTRUMENTS & THEIR PURPOSE IN IMPACT ASSESSMENT

Source: National Rural Roads Development Authority, "Socio-Economic Impact Assessment Report: Rural Roads Project – 1, Madhya Pradesh," April 2009. Accessible via www.pmgsy.nic.in

The impact assessment by the MoRD in 2004 is less clear. Though it is consolidated, cross-state report mentions a range of techniques used to collect both quantitative and qualitative data, including survey questionnaires, focus group discussions, open-ended interviews, and case studies, the findings presented appear to be entirely qualitative. The report makes the caveat, "The findings of the study are based on a very small sample. Thus, caution should be exercised in generalization of the findings for the entire state. The impact mentioned in the report is primarily based on the perception of the respondents."⁵²

The state-level report for the same 2004 impact assessment in Madhya Pradesh is an exception. This is perhaps because each state assessment was contracted to a separate consulting firm—in the case of Madhya Pradesh, Development & Research Services Pvt. Ltd.⁵³ Like the consolidated report and other state reports, this report also included qualitative responses on development results, but it also supplemented with quantitative information, such as that shown in Table 5.

The use and reporting of such methods across all states, made explicit in the methodology and findings in the cross-state reports of the impact assessments, would be a step in the right direction towards quality impact evaluation. The mixed-methods approach would ensure

District	Earlier average income of the households (in Rs.)	Average income of the households after the construction	Difference (In Rs.)	Percentage
Bhopal	17214	18958	1744	10.1
Datia Dhar Umaria	15750 17423 16066	16966 19275 17605	1216 1852 1539	7.7 10.6 9.5

TABLE 5 Change in income

*After about a year of the construction of PMGSY road.

Source: Ministry of Rural Development, Monitoring Division, "Quick Impact Assessment Study of the PMGSY in Madhya Pradesh," February 2004. Accessible via www.pmgsy.nic.in

that evaluations give context to why programs are or are not working rather than simply focusing on impact results in a vacuum, thereby enabling the evaluations to be followed up with actionable steps to improve programs.

Conclusion

Effective impact evaluation is a challenging task. The development evaluation culture in India suffers not only from challenges in political will and institutional design, but from the inherent difficulty in measuring development results and attributing them to the many actions that may have caused them. In this regard, public schemes in India are not alone: other nations experience the same challenges in implementing impact evaluations, and private donors and nonprofit organizations struggle just as much to define causal pathways to their development goals. Nevertheless, the scale and importance of public development programs in India mean that it is critical to work towards improving evaluation practices to identify roadblocks and maximize the capacity of public expenditures to create development impact.

The approach we have outlined is not meant to be a silver bullet or to solve all the challenges of evaluation, but merely to provide a few useful incremental steps to improve individual evaluations immediately. The PMGSY was chosen as the example for our thought experiment to make the approach easy to understand; replicating the recommendations will likely be more difficult for less transparent or data-rich schemes and, in particular, for less tangible schemes. Schemes in which the core activities are based on larger or more ambiguous assumptions about development effectiveness such as in the challenging fields of primary health and education, will be particularly difficult. The program recommendations that come out of the evaluations of these schemes must be contextualized more as a result, with the understanding that the evaluation results are based on more strategic assumptions than more straightforward schemes are. However, these improvements are still relevant to such schemes. They still go a long way in pinning down and documenting the strengths and weaknesses of the schemes' implementation, understanding their assumptions, and identifying areas for improvement in both implementation and evaluation. Furthermore, these small steps can serve as an advocacy tool to help evaluators push for more investment in and attention to impact evaluation, ideally from a scheme's inception.

The incremental steps we have outlined to improve impact evaluations in the near term may be resisted by evaluators and government officials, due to the challenges incurred even in these small steps and the political issues that the articulation of expected outcomes and description of ongoing progress may bring to the surface. It is important to note that the same challenges could become political opportunities. If outcomes could be measured, schemes would be able to take more credit for the outcomes and impacts they are producing. Pushes for increases in fund allotments, staff sizes, and new implementation strategies could be justified based on the success or failure of the current programs. Clarity on the status of progress of development programs has the potential to help all stakeholders, from funders to implementers to beneficiaries.

Evaluators can aid the process of improving the quality of evaluations by pointing out weaknesses in current evaluations explicitly. More attention can be drawn to the initial caveats of impact evaluation studies and challenges to their implementation. Program officers and government officials drafting new or revised schemes can help immensely by adding basic evaluation principles to the scheme approaches and guidelines, and amendments can be made to current scheme guidelines to ensure improvements are incorporated into future evaluations. Beneficiaries and concerned citizens can contribute on an institutional level by raising their voices to policymakers, encouraging champions of evaluation in government positions to bring their ideas forward. The pathways to improved standards for impact evaluation in India are not as challenging as they may seem; they simply need to be put into place.

NOTES

The authors would like to give special thanks to Dr. Jessica Wallack for strategic guidance and review during this research project. We also thank Jay Chaudhuri, Mohit Sharma, Raja Dhanabal, and Theresa Chen for their inputs, feedback, and assistance with data collection.

- 1. Ministry of Communications and Information Technology, "Government of India Outcome Budget, 2006–07." Accessible from http://www.dot.gov.in
- 2. For more information see descriptions on the PEO website, accessible from http://planningcommission.nic.in/reports/ peobody.html
- 3. Kaveri Gill, "A Primary Evaluation of Service Delivery under the National Rural Health Mission (NRHM): Findings from a Study in Andhra Pradesh, Uttar Pradesh, Bihar and Rajasthan," *Working Paper, PEO, Planning Commission of India*, May, 2009. Accessible from http://www.planningcommission.nic.in
- 4. "Outcome budget serves no purpose, say Montek," *The Economic Times*, October 13, 2009.
- 5. Development Evaluation Society of India. "Reviving and Sustaining Development Evaluation Culture in India." Accessible from http://www.ioce.net/members
- 6. Evaluation reports can be found on the Planning Commission website, accessible from http://www.planningcommission.nic.in
- Catherine Clark, William Rosenzweig, David Long, and Sara Olsen, "Double Bottom Line Project Report: Assessing Social Impact in Double Bottom Line Ventures, Methods Catalog," The Rockefeller Foundation, 2004. Accessible from http://www.rockfound.org
- Development Assistance Committee Working Party on Aid Evaluation, Organization for Economic Co-operation and Development, "Evaluation and Aid Effectiveness: Glossary of Key Terms in Evaluation and Results Based Management," 2002. Accessible from http://www.oecd.org
- 9. Clark, Rosenzweig, Long, and Olsen, "Double Bottom Line Project Report."
- Definition of the Independent Evaluation Group, The World Bank, Accessible from http://www.worldbank.org/ieg/ie/ (accessed October 14, 2009).
- 11. For more information see Howard White, "Theory-Based Impact Evaluation," 3ie Working Paper Series, June 2009. Accessible from http://www.3ieimpact.org and Jody Kusek and Ray Rist, *Ten Steps to a Results-Based Monitoring and Evaluation System* (Washington, D.C.: The World Bank, 2004).
- 12. Linda Morra Imas and Ray Rist, *The Road to Results* (Washington, D.C.: The World Bank, 2009). Frans Leeuw and Jos Vaessen, *Impact Evaluations and Development:* NoNIE Guidance on Impact Evaluation, The Network of Networks on Impact Evaluation, 2009, accessible from http://www.worldbank.org/ieg/nonie. International Initiative for Impact Evaluation, "How to Institutionalize Evaluation?" October 2009, accessible from http://www.3ieimpact.org. Jody Kusek and Ray Rist, *Ten Steps to a Results-Based Monitoring and Evaluation System*; White, "Theory-Based Impact Evaluation."
- 13. Angus Deaton, "Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development," *Centre for Health and Wellbeing*, Princeton University, January 2009.
- 14. Please note that by "separation" we do not imply that impact evaluations should be isolated from other types of evaluation or analyzed in a vacuum. Understanding the results of an impact evaluation in the context of its implementation process is key to providing robust recommendations for improvement. The ideal type simply requires that the budget, design, evaluator choice, etc., of an impact evaluation not be lumped in with other evaluations, as this is sure to dilute the ability to devote appropriate attention to the challenging task of evaluating impact.
- Programme Evaluation Organisation, "Development Evaluation in Programme Evaluation Organisation, volume II" Report 197, (January 2006). Accessible from http:// www.planningcommission.nic.in

264 India Review

- 16. Programme Evaluation Organisation, "Development Evaluation."
- 17. Gill, "A Primary Evaluation."
- Programme Evaluation Organisation, "Evaluation Report On Integrated Child Development Scheme (ICDS) Jammu & Kashmir," February 2009. Accessible from http:// www.planningcommission.nic.in
- Programme Evaluation Organisation, "All-India Report on Evaluation of NREGA: A Survey of Twenty Districts," 2009. Accessible from http://www.planningcommission.nic.in
- 20. Programme Evaluation Organisation, "All-India Report."
- 21. Centre for Policy Research, "Transparency and Accountability in NREGA," *Engaging* Accountability: Working Paper Series, 2009.
- 22. International Initiative for Impact Evaluation, "How to Institutionalize Evaluation?"
- 23. The President's address to the Indian Parliament, June 4, 2007. Accessible from http://www.presidentofindia.nic.in
- 24. "Planning Commission organises two day conference on evaluation," Chennai Vision. October 2009. Accessible from http://www.chennaivision.com
- 25. "India Takes Steps to Become a World Leader in Evaluation," International Initiative for Impact Evaluation News, October 14, 2009. Accessible from http://www.3ieimpact.org
- 26. National Advisory Council, "Prime Minister's Advisory Council Report: Pradhan Mantri Gram Sadak Yojana," October 17, 2007. Accessible from http://www.pmindia.gov.in
- 27. For example: Logical Framework Approach (LFA) uses the same four steps as we do in its "objective analysis" step, but uses slightly different terminology: Activities → Outputs/Immediate results → Immediate objectives/Project purpose → Development objectives (Kari Örtengren, "The Logical Framework Approach," Swedish International Development Cooperation Agency, 2004. Accessible from http://www.sida.se). The "impact value chain," profiled in Catherine Clark, William Rosenzweig, David Long, and Sara Olsen, "Double Bottom Line Project Report," also uses this logic and the same terms, with a slight additional nuance: the authors define "impact" as outcomes minus "what would have happened anyway," and assert that outcomes, whether caused by the organization or not, lead up to goal alignment.
- 28. PMGSY Scheme and Guidelines, accessible from http://www.pmgsy.nic.in
- 29. Ministry of Rural Development, Monitoring Division, "Impact Assessment of PMGSY In the States of Assam, Himachal Pradesh, Madhya Pradesh, Mizoram, Orissa, Rajasthan, Uttar Pradesh, Tamil Nadu and West Bengal," 2004. Accessible from http://www. pmgsy.nic.in
- 30. International Food Policy Research Institute, "Government Spending, Agricultural Growth and Poverty: An Analysis of Interlinkages in Rural India," Research Report 110, 1999. Accessible from http://www.ifpri.org
- Development Strategy and Governance Division, International Food Policy Research Institute, "Road Development, Economic Growth and Poverty Reduction in China," 2004. Accessible from http://www.ifpri.org
- "World Bank Supports India's Rural Roads Program," The World Bank Press Release No: 2005/104/SAR, September 23, 2004. Accessible from http://www.worldbank.org
- 33. Note the dangers of avoiding consideration of potential negative impacts, such as loss of property, as described in: The World Bank, "Pradhan Mantri Gram Sadak Yojana (PMGSY) Social Assessment of World Bank Funded Second Rural Road Project," February 2008. Accessible from http://www.worldbank.org
- 34. It should be noted that implementing a "one size fits all" development program across as diverse a nation as India carries its own risks, making it all the more important that close monitoring and evaluation take place.
- 35. Ministry of Rural Development, "Rapid Assessment of Socio-Economic Impact of PMGSY in Assam," 2004. Accessible via http://www.pmgsy.nic.in
- 36. "CREAM" criteria documented in Salvatore Schiavo-Campo, "'Performance' in the public sector," *Asian Journal of Political Science* Vol. 7, No. 2 (1999), pp 75–87. Accessible from www.informaworld.in

- 37. Kusek and Rist, Ten Steps.
- Supported by United Nations Development Program, "Governance Indicators: A User's Guide." Accessible from http://www.undp.org
- 39. Morra Imas and Rist, The Road to Results.
- 40. Kusek and Rist, Ten Steps.
- 41. The Ministry of Rural Development's Impact Assessment of PMGSY consolidated report used secondary data sources for "data on lengths of the roads, their connectivity, dates of starting and completing the construction and other details were collected from the secondary sources." Ministry of Rural Development, Monitoring Division, "Impact Assessment of PMGSY"
- 42. The authors thank an anonymous reviewer for cautioning that there may be issues of accuracy and uniform availability of data across states and districts in the recommended data sources. District-level estimates for the NSSO 55th round can be obtained by employing the small area estimation technique and using Census data for 2001.For more detail on the technique, please refer to: United Nations and India Development Foundation, "Small Area Estimates of Selected Welfare Indicators: Results for Uttaranchal," November 2005. Further, the anonymous reviewer noted that there may not be consensus on comparability of metrics across NSSO rounds, as some aspects of the NSSO questionnaire may have been modified between the 55th round and subsequent rounds.
- 43. Rusek and Rist, Ten Steps.
- Joshua Angrist and William Evans, "Children and Their Parents" Labor Supply: Evidence from Exogenous Variation in Family Size," *American Economic Review* Vol. 88, No. 3 (1998), pp. 450–77.
- 45. National Rural Roads Development Authority, "Socio-Economic Impact Assessment Report: Rural Roads Project – 1," for Chhattisgarh and Madhya Pradesh, April 2009. Accessible from http://www.pmgsy.nic.in
- 46. Operations Evaluation Department, Asian Development Bank, "When Do Rural Roads Benefit the Poor and How?" 2006. Accessible from http://www.adb.org
- National Rural Roads Development Authority, "Socio-Economic Impact Assessment Report: Rural Roads Project – 1," for Madhya Pradesh, April 2009. Accessible via www.pmgsy.nic.in
- 48. For more information please see "Habitation Reports" for data on road completion from 2001 to 2010. Accessible from http://www.pmgsyonline.nic.in/aspnet/citizens
- 49. In order to perform a pre-post comparison of the primary data collected at a habitation level to the baseline data collected from secondary data sources, the primary data will need to be aggregated at least to the district level.
- 50. Morra Imas and Rist, The Road to Results, p. 298.
- 51. Morra Imas and Rist, The Road to Results, p. 403.
- 52. Ministry of Rural Development, Monitoring Division, "Impact Assessment of PMGSY."
- 53. For more information please see Ministry of Rural Development, Monitoring Division, "Quick Impact Assessment."