DOING IMPACT EVALUATION No. 14

Making Smart Policy: Using Impact Evaluation for Policy Making

Case Studies on Evaluations that Influenced Policy

THE WORLD BANK

Poverty Reduction and Economic Management PREM

Thematic Group on Poverty Analysis, Monitoring and Impact Evaluation

Making Smart Policy: Using Impact Evaluation for Policy Making

Case Studies on Evaluations that Influenced Policy

June 2009

Acknowledgement

This publication reviews the experiences presented in the conference "*Making Smart Policy: Using Impact Evaluation for Policymaking",* held in January 2008. The editors, Michael Bamberger and Angeli Kirk, would like to thank the presenters for their thoughtful and honest insight into their impact evaluation experiences: Orazio Attanasio, Antonie de Kemp, Jocelyne Delarue, Pascaline Dupas, Joseph Eilor, Deon Filmer, Emanuela Galasso, John Hoddinott, Michael Kremer, Emmanuel Skoufias, Miguel Urquiola, Dominique Van De Walle, Adam Wagstaff. They also thank the chairs of each of the four parallel sessions: Judy Baker, Sustainable Development; Halsey Rogers, education; Norbert Schady, CCTs; and Charles Teller, health; as well as Elizabeth King, who chaired the plenary session that brought together the lessons from the parallel sessions. This note was task managed by Emmanuel Skoufias.

TABLE OF CONTENTS:

1. Overview

Impact evaluation has blossomed in recent years as a powerful tool for enhancing development effectiveness. The numbers of both evaluations and methodologies have multiplied very quickly. This growth, however, has been uneven both geographically and across sectors, leading to questions of how to bolster impact evaluation in regions and sectors where it is least common and perhaps most needed. Additionally, as methods mature and the collection of evidence accumulates, the conversation is expanding to include reflection on how we – development practitioners, policy makers, and researchers alike – can assure that impact evaluation reaches its potential for influencing project and policy design. A key question is, how can we strategically use scarce evaluation resources more effectively? That is, how can we ensure that impact evaluations are better utilized and more influential?

To explore these issues, the World Bank, with support from DFID and the Government of the Netherlands, held a conference *Making Smart Policy: Using Impact Evaluation for Policymaking* in January 2008.[1] One session - *Evidence and Use: Parallel Sector Sessions* - brought together 12 case studies to ground the discussion in concrete examples of impact evaluations that have been completed and to provide researchers' perspectives on the ways in which they had been influential – or not – and why. *Evidence and Use* comprised four separate thematic sessions: education, conditional cash transfers (CCTs), health and sustainable development.

This publication reviews the experiences presented in the conference session and draws lessons concerning different ways that impact evaluations are utilized and how they can contribute to improving program design and policy formulation. The overview chapter begins by describing the structure and general content of the conference presentations and proposing a framework for considering utilization and influence. It then briefly describes the evaluations and pulls together the most salient examples of how they were used and the type of influence they had. Finally, lessons are drawn on ways to enhance evaluation utilization and its contribution to program design and policy.

It should be noted that the primary purpose of the report is not the discussion of the impact evaluation methodology. Nevertheless, the overview chapter includes a chart summarizing the evaluation designs and findings, as there can be linkages among the policy questions being addressed, how an evaluation was designed, the findings and how they were communicated, and how the evaluation contributed to program design and policy formulation.

The remaining chapters are devoted to more in-depth syntheses of the case studies presented in the workshop with respect to the evaluation of education, anti-poverty programs, health, and sustainable development, with a final chapter on lessons learned.

[1] The conference website, with videos of the sessions and supplementary material, may be found at: www.worldbank.org/iepolicyconference.

A. Presentation format

Presenters in each session of *Evidence and Use: Parallel Sector Sessions* were asked to reflect on an impact evaluation experience. As well as briefly describing the project, evaluation design, and findings, the speakers discussed the dissemination process, *how the evaluation findings were utilized, and what kinds of influence they had*. Interestingly, while the focus was meant to be on utilization and impact rather than project details or evaluation technique, the distinction proved somewhat artificial, as details of the design and context of both the project and the evaluation were often central to the use and influence or lack thereof. After the presentations, the groups reflected on the general lessons that could be drawn from the case studies concerning the different kinds of contributions that evaluations can make to program management and policy formulation. Guidelines were then proposed on ways to increase the utilization and influence of evaluations for development programs and policies.

There is an important caveat: this report does not offer an "impact evaluation of impact evaluations". It is difficult to interpret associations between the conduct of an evaluation and its recommendations on the one hand, and causal relations – changes in program design or an increased use of research by policymakers – on the other. (For example, did evaluations of the education system in Uganda lead to increased appreciation and use of the management information system; or was the evaluation conducted and used because there was already an awareness of the value of research and statistics?). The evidence and recommendations concerning evaluation utilization are drawn from the impressions and observations presented by the researchers who conducted the evaluations and the subsequent discussions with workshop participants. In only one case (Uganda Education for All) was a representative of the host country partner agency present. *None of the evaluators had conducted systematic studies on the utilization of their evaluations (such as interviews with stakeholders), and no kind of attribution analysis was conducted.* It is quite possible that evaluators may not be fully aware of how the evaluations were used, and their reflections on their experiences might introduce a certain bias.

Box 1: **The programs, the evaluation designs and the main findings**

Table 1 (end of chapter) describes the programs, the key evaluation questions and the main findings of each evaluation, and Table 2 summarizes the evaluation designs. More details are given in the following chapters.

Education. The objectives of the education programs in Cambodia and Uganda were to increase school enrolment and retention for low-income students, particularly girls; and in the case of Uganda to also improve education quality. The program in Chile, which already had very high enrolment rates, was intended to improve quality for low-income students through increased access to private education. In addition, all of the programs sought to enhance the efficiency of program management. Each of the evaluations was also intended to assess the effectiveness of specific interventions such as vouchers, scholarships and management training, in enhancing enrolment and/or improving quality.

Impact evaluation designs included retrospective comparisons, regression discontinuity, using data from management information systems to measure changes over the life of the project, and using secondary data to match project and comparison groups through propensity score matching.

The findings showed that in both Cambodia and Uganda enrolment and retention increased for low-income families. However, the quality of education remained low, although pilot projects in Uganda, focusing on management training showed promising results with respect to quality improvement. In Chile, contrary to popular belief, there was no evidence that vouchers improved educational outcomes. However, the "sorting" mechanisms that resulted from the scholarship programs meant that better qualified students tended to move to private schools – an outcome that was not intended and that had negative consequences for public schools and perhaps for low-income students.

Anti-poverty programs. The programs in Mexico (PROGRESA/Oportunidades) and Colombia (Families en Accion) were conditional cash transfer (CCT) programs providing cash payments to low-income families on the condition that their children enrolled in school and went for regular health check-ups (and in the case of Mexico also received nutritional supplements). The two programs were quite similar in many ways, and in fact both the program design and the evaluation design of Familias en Accion drew on the experience of the Mexican programs. The Argentina Emergency Safety Net (the "Jefes") program provided cash payments to under-employed heads of low-income households to mitigate the impact of the 2000-2002 economic crisis. Household heads of poor families received monthly cash payments on the condition that they attended education or training programs or participated in community public works programs. While the *Jefes* program could also be considered a CCT as beneficiaries were theoretically required to attend training or participate in community improvement projects, in practice this requirement was often not enforced and the Safety Net was widely considered as an entitlement program (i.e. participants were entitled to receive the payments without any conditionality).

All three evaluations used experimental or strong quasi-experimental designs. Mexico used randomized control trials (RCT) for selection of beneficiaries at each phase. Colombia and Argentina each used propensity score matching (PSM); in Colombia, recipients were matched to households in ineligible areas and, in Argentina, participants were matched to applicants who had not yet been chosen to participate.

The Colombia and Mexico evaluations both found that CCTs increased school enrolment and access to health services. All three evaluations found that they were effective in reducing the proportion of the population below the poverty line or, in the case of Argentina, effective in preventing families from falling below the poverty line. However, the impacts varied by factors such as student age and urban/rural location.

Health. The health programs comprised insecticide treated mosquito nets in Kenya to reduce the incidence of malaria; school deworming in Kenya to reduce school absenteeism due to sickness; and a health insurance scheme in China. The goals of the China program were to reduce out-of-pocket expenses by patients, to encourage greater use of preventive care, to reduce excessive use of high-tech services and to encourage the use of health services.

The two Kenyan programs used randomized control trial evaluation designs. The China evaluation was integrated with a large government health sector evaluation and used double-difference analysis with propensity score matched samples.

The evaluations of both Kenyan programs found that ability to pay was a key factor in utilization. Efforts to introduce cost-recovery significantly reduced coverage – in the case of the insecticide net program, free distribution resulted in a 63 per cent coverage rate compared to 14 per cent compared to the highest price. Deworming participation also dropped dramatically when parents were asked to pay even small amounts. The evaluation of the China health insurance program found that utilization had increased but that out-of-pocket payments did not decrease. Facilities data found that revenue had increased more than utilization. The results showed that medical insurance is not guaranteed to decrease expenses, leading to questions about the level of care provided and whether services were selected because of medical necessity or for revenues.

Sustainable development. The programs comprised microfinance programs in Morocco and Madagascar, food security in Ethiopia and the rehabilitation of rural roads in Viet Nam. All four programs were intended to achieve sustainable reductions in poverty.

Three of the evaluation designs used retrospective comparisons with different levels of rigor in the matching of the project and comparison group samples. The fourth (Morocco) used randomized control trials.

The findings showed that the Viet Nam roads program was successful in diversifying and strengthening livelihoods but the scope was more limited than planned. The Ethiopia Food Security program also achieved its main objectives but failed to achieve integration with other complementary programs. Microfinance in Madagascar was not found to have an impact on economic trends among clients. Findings for microfinance in Morocco are still forthcoming.

B. Conceptualizing utilization and influence

When assessing the use and utility of an evaluation, it is helpful to consider two components: we term them "utilization" and "influence."

Utilization: How were the evaluation findings (and even the process) used - by whom and for what purpose? The first uses that generally come to mind are those related to impact evaluation as an *assessment* tool. For example, one may conduct an evaluation in order to:

- monitor project implementation,
- measure the benefits of an existing program and check for unanticipated side effects,
- assess the distribution of participation and benefits across different segments of the target population,
- make informed changes and improvements to an ongoing project,
- test options for the design of a project that will be implemented in the future, and
- compare the cost-effectiveness or benefit/cost ratio of alternative programs for budget planning purposes.

In practice, however, impact evaluations are also very commonly used as a *political* tool. They are frequently employed to:

- provide support for decisions that agencies have already decided upon or would like to make,

- mobilize political support for high profile or controversial programs,
- provide independent support (the international prestige and perceived independence of the evaluator is often important) for terminating a politically sensitive program, and
- provide political or managerial accountability.

In fact, in the end it is likely to be the potential political benefit or detriment that causes decision makers to embrace or avoid evaluations. As a result, those who would like to promote impact evaluation as an assessment and learning tool will have to be fully aware of the given political context and navigate strategically.

Influence: In assessing the influence of an impact evaluation, there are a number of aspects one might consider:

- *What causes or facilitates an impact evaluation's influence?* It is important to remember that it is not only the findings of an impact evaluation that can have an impact. The decision to conduct an evaluation, the choice of methodology, and how the findings are disseminated and used can all have important consequences – some anticipated, others not; some desired and others not. For example, the decision to conduct an evaluation using a randomized control trial can influence who benefits from the program, how different treatments and implementation strategies are prioritized, what is measured and the criteria used to decide if the program had achieved its objectives.[2] In other cases, if findings are presented in a manner that is too technically complex for its audience, decision makers may either misinterpret the findings, leading to misinformed choices, or ignore the findings altogether.

- *Where can the evaluation's influence be seen?* Some possibilities include administrative realms such as program design and scope, or the political realm in the form of popular support for a program or its associated politicians. One may also consider the resulting perceptions and understanding of impact evaluation, by policymakers and project administrators as well as by researchers who conduct future evaluations. For high profile programs, the influence of the evaluation may also be seen in how the debate on the program is framed in the mass media.

- *How much influence did the evaluation have on the decisions and actions of managers, planners and policymakers?* Did it have a major influence, or did it only corroborate what was already known or support decisions that had already been made? That is, to what degree have any decisions actually been made differently as a result – has the impact evaluation had any impact? Decision-makers are exposed to many different sources of information, advice and pressure, of which the evaluation is only one – and usually not the most significant.

[2] A frequently cited example from the US was the decision to assess the performance of schools under the No Child Left Behind program in terms of academic performance measured through end-of-year tests. This meant that many schools were forced to modify their curricula to allow more time to coach children in how to take the tests, often resulting in reduced time for physical education, arts, and music.

While utilization and influence are distinct as concepts, in practice they are often – though not necessarily – found to overlap. For example, if an evaluation is utilized to determine the most effective project design, then the influence may be that a future project is chosen based on strong evidence rather than on other criteria. On the other hand, there are times when the influence of an evaluation does not reflect its utilization, such as a number of cases in which an evaluation was used to gain political support (the utilization) but in the process, impact evaluation in general came to be viewed as an important and even necessary tool (an impact).

C. Reviewing the evidence: how were the evaluations utilized and what kinds of influence did they have ?

The following brings together the utilization and influence that were observed in the 12 case studies. In most cases the information is based on the perceptions and experience of the evaluators themselves, although in one case a representative of the government client agency was also present. The types of use and influence seen in the presented cases can be broadly grouped into three categories: project implementation and administration, political support, and the process and culture of evaluation itself.

Project implementation and administration:

- Evaluations were often used for the design of future programs. They provided specific operational guidance or general guidance for the strategic focus. They often helped identify logistical and administrative problems that had been overlooked.
- The Ethiopian food security evaluation identified a number of process failures, although they still found positive impacts, and authorities found it useful to have learned that there were process problems, as these were practical issues that could be addressed. In Ethiopia and in China's health insurance evaluation, "bad news" was delivered sufficiently early so that it didn't just condemn a completed project – but instead provided practical guidance for improvements.
- Several cases were cited where the extensive dissemination of evaluation findings also served to raise the profile of the *category* of programs being evaluated. Examples include deworming and conditional cash transfers.
- Evaluations help clients understand their programs in a broader context. Evaluations helped identify broader systemic implications of programs and contributed to understanding of local contextual factors affecting how projects operate in different districts or locations.
- Several evaluations have made specific contributions to choices among policy alternatives. For examples, two health evaluations in Kenya helped convince government and donors to provide free anti-malarial bednets and deworming in schools rather than to seek cost-recovery by charging.

Political support:

- Evaluations are often used to justify continued funding for a program, or to ensure political support for a new or expanded program. The evaluations of the first CCTs in Mexico and Colombia are both considered to have helped convince new administrations to continue high profile programs started by their predecessors. In several cases an evaluation was used to justify a new program, even when in fact the evaluation findings did not support this new program (for example, expansion of the Colombian Familias en Accion program from rural to urban areas).

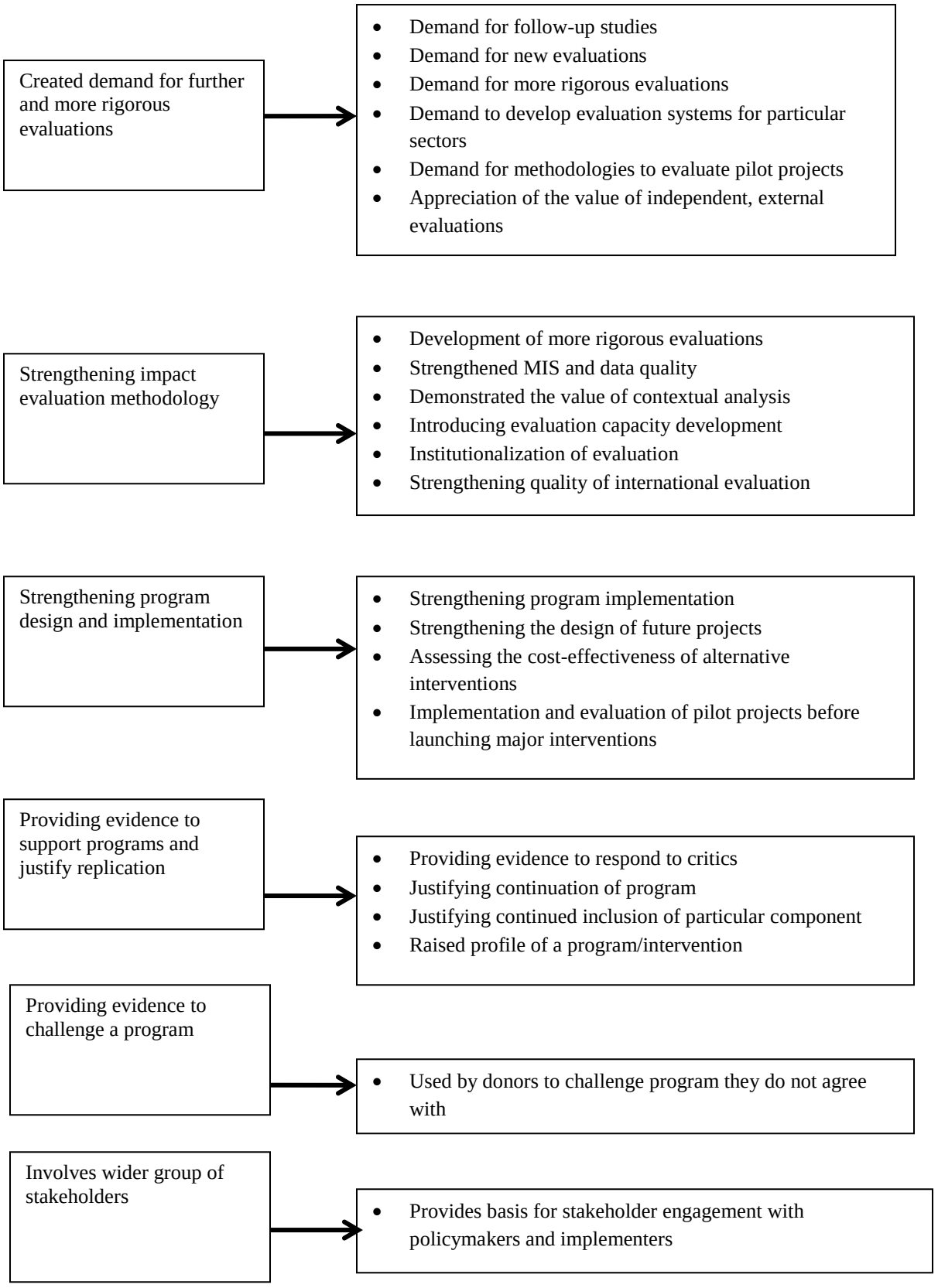
Culture of and capacity for impact evaluation:

- Evaluations that are favorably received by clients often lead to increased interest in further evaluations. Well designed and implemented evaluations have helped legitimize evaluation as a useful planning or policymaking tool. Initially many clients or local districts were either skeptical about an evaluation's utility or were afraid that the findings would be too negative or critical. In several cases attitudes became more positive and utilization increased as the evaluations progressed. Not surprisingly, it was much easier to gain acceptance for the evaluation process and findings when the findings were mainly *positive*. Well received evaluations often lead to follow-up evaluations to assess more specific issues that had been identified.
- There were, however, examples, where initial negative findings created reluctance to accept or use an evaluation, but where attitudes gradually became more favorable. The health insurance evaluation in China was very poorly received in the beginning because it showed negative results on the primary objective of reducing out-of-pocket health care expenditures (though positive results for a secondary objective of increasing use of health care services). In the end, though, authorities accepted the results and were able to use them to make some reforms (especially increased funding), and the process seemed to have increased general acceptance of impact evaluation as a tool.
- The Ethiopian food security evaluation identified a number of process failures, although they still found positive impacts, and authorities found it useful to have learned that there were process problems, as these were practical issues that could be addressed. Again, in both Ethiopia and China, however, "bad news" was delivered sufficiently early so that it didn't just condemn a completed project but instead provided practical guidance for improvements.
- Several cases were cited where the extensive dissemination of evaluation findings also served to raise the profile of the kinds of programs being evaluated. Examples include deworming and conditional cash transfers.
- Several well designed and well received evaluations have contributed to the development of a culture of evaluation and a move towards the institutionalization of evaluations rather than the ad hoc and fortuitous way in which earlier evaluations were selected and funded. Once the benefits of well designed evaluations became understood, this helped raise expectations concerning the level of rigor required in future evaluations. Methodologies, such as randomized

control trials, double-difference designs, or regression discontinuity provided models that were then replicated in other program areas.

- Where several sequential evaluations were conducted, the effect on client attitudes toward and use of evaluation is cumulative, and clients have learned to demand the kinds of information that they need and can use.
- A strengthened culture of evaluation can also stimulate evaluation capacity development, in some cases strengthening government research agencies such as the statistics bureau, in other cases training to improve the quality of monitoring data collection and use.

Figure 1: The influence and utilization of impact evaluations



D. Factors affecting evaluation utilization and influence

The following is a synthesis of the broad range of factors identified in the presentations as potentially affecting evaluation utilization.

Timing and focus on priority stakeholder issues:

- The evaluation must be timely and focus on priority issues for key stakeholders. This ensures there is a receptive audience. Timing often presents a trade-off: on the one hand, designing an evaluation to provide fast results relevant for the project at hand, in time to make changes in project design and while the project still has the attention of policymakers. On the other hand, evaluations that take longer to complete may be of higher quality and can look for longer term effects on the design of future projects and policies.
- The evaluator must be opportunistic, taking advantage of funding opportunities, or the interest of key stakeholders. Several countries that have progressed toward the institutionalization of evaluation at the national or sector level began with opportunistic selection of their first impact evaluations[3].
- The evaluator should always be on the look-out for "quick-wins" – evaluations that can be conducted quickly and economically and that provide information on an issue of immediate concern. Showing the practical utility impact evaluations can build up confidence and interest before moving on to broader and more complex evaluations.
- Also, there is value in firsts. Pioneer studies may not only be useful for showing the impact of the intervention, but in a broader context they may also change expectations about what can and should be evaluated or advance the methods that can be used. Again, even less-than-ideal evaluations that are first or early in their context may contribute by building interest in and capacity for impact evaluation.
- A series of sequential evaluations gradually builds interest, ownership and utilization.

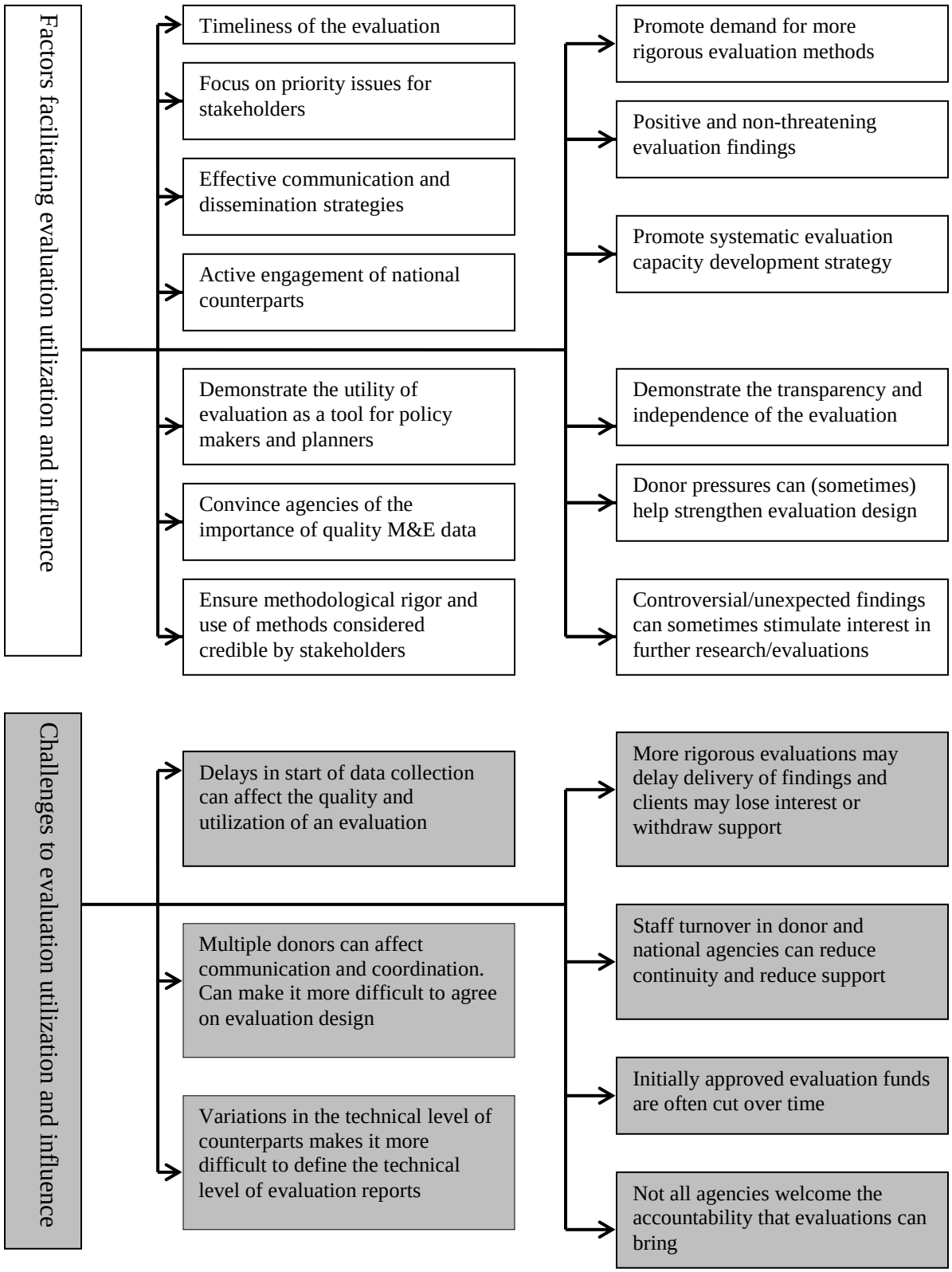
Effective dissemination

- Rapid, broad and well targeted dissemination are important determinants of utilization. One reason that many sound and potentially useful evaluations are never used is that very few people have ever seen them.
- Making data available to the academic community is also an important way of broadening interest and support for evaluations and also of legitimizing the methodologies (assuming they stand up to academic critiques as have PROGRESA and Familias en Accion).

---

[3] See IEG (2008) *Institutionalizing Impact Evaluation within the Framework of a Monitoring and Evaluation System*. The Education for All evaluations in Uganda were cited as an example of institutionalization at the sector level and the SINERGIA evaluation program under the Planning Department in Colombia is an example of institutionalization of a national impact evaluation system. The report is available at: http://lnweb90.worldbank.org/oed/oeddoclib.nsf/DocUNIDViewForJavaSearch/E629534B7C677EA78525754700715CB8/$file/inst_ie_framework_me.pdf, or at www.worldbank.org/ieg/ecd.

Figure 2: Factors affecting evaluation utilization and influence



Providing rapid feedback to government on issues such as the extent of corruption or other "hot" topics enhances utilization.

- Continuous and targeted communication builds interest and confidence and also ensures "no surprises" when the final report and recommendations are submitted. This also allows controversial or sensitive findings to be gradually introduced. Trust and open lines of communication are important confidence builders.
- Where there is existing demand for a particular evaluation, the results may partially disseminate themselves and may be more likely to be used.

Clear and well communicated messages

- Clarity and comprehensibility increase use. It helps when the evaluation results point to clear policy implications. This may also apply to the comprehension of methods. While stakeholders may be willing to "trust the experts" if an evaluation offers results that support what they want to hear, there may be a reasonable tendency to distrust results – and particularly methods – that they don't understand.

Active engagement with national counterparts

- The active involvement of national agencies in identifying the need for an evaluation, commissioning it, and deciding which international consultants to use is central to utilization.
- Close cooperation with national counterpart agencies proves critical in several ways. It gives ownership of the evaluation to stakeholders and helps ensure the evaluation focuses on important issues. It often increases quality by taking advantage of local knowledge and in several cases reduces costs (an important factor in gaining support) by combining with other ongoing studies. This cooperation can enable evaluators to modify the initial evaluation design to reflect concerns of clients – for example, changing a politically sensitive randomized design to a strong quasi-experimental design.
- Involving a wide range of stakeholders is also an important determinant of utilization. This can be achieved through consultative planning mechanisms, dissemination and ensuring that local as well as national level agencies are consulted.
- In some contexts (such as the China health insurance scheme), the involvement of the national statistical agency increases the government's trust – the results and the process have been better accepted when overseen and presented by the statistics agency.

Demonstrating the value of evaluation as a political and policymaking tool

- When evaluation is seen as a useful political tool, this greatly enhances utilization. For example, managers or policymakers often welcome specific evidence to respond to critics, support for continued funding or program expansion. Evaluation can also be seen as a way to provide more objective criticism of an unpopular program.
- Once the potential uses of planning tools such as cost-effectiveness analysis are understood, this increases the demand for, and use of, evaluations. Evaluations

can also demonstrate the practical value of good monitoring data, and increased attention to monitoring in turn generates demand for further evaluations. When evaluations show planners better ways to achieve development objectives, such as ensuring services reach the poor, this increases utilization and influence.

- Increasing concerns about corruption or poor service delivery have also been an important factor in government decisions to commission evaluations. In some cases, a new administration wishes to demonstrate its transparency and accountability or to use the evaluation to point out weaknesses in how previous administrations had managed projects.
- Evaluations that focus on local contextual issues (i.e. that are directly relevant to the work of districts and local agencies) are much more likely to be used.

The methodological quality of the evaluation and credibility of the international evaluators

- High quality of an evaluation is likely to increase its usefulness and influence. Quality improves the robustness of the findings and their policy implications and may assist in dissemination (especially in terms of publication). However, an impact evaluation of a compromised quality may still be useful if it can provide timely and relevant insight or if it ventures into new territory: new techniques, less-evaluated subject matter, or in a context where relevant stakeholders have less experience with impact evaluations.
- The credibility of international evaluators, particularly when they are seen as not tied to funding agencies, can help legitimize high profile evaluations and enhance their utilization.
- In some cases the use of what is considered "state of the art" evaluation methods, such as randomized control trials, can raise the profile of evaluation (and the agencies that use it) and increase utilization.
- New and innovative evaluations often attract more interest and support than the repetition of routine evaluations.
- On the other hand, while studies on the "frontier" may be more novel or attract more attention, subsequent related studies may be useful in confirming controversial findings and building a body of knowledge that is more accepted than a single study, especially a single study with unpopular findings.
- Evaluation methods, in addition to being methodologically sound, must also be understood and accepted by clients. Different stakeholders may have different methodological preferences.

Positive and non-threatening findings

- Positive evaluations, or those that support the views of key stakeholders, have an increased likelihood of being used. While this is not surprising, one of the reasons is that many agencies were either fearful of the negative consequences of evaluation or (to be honest) considered evaluation as a waste of time (particularly the time of busy managers) or money. Once stakeholders have appreciated that evaluations were not threatening and were actually producing useful findings, agencies have become more willing to request and use evaluations and gradually

to accept negative findings – or even to solicit evaluations to look at areas where programs were not going well.

- There is always demand for results that confirm what people want to hear. There may be some benefit in taking advantage of opportunities to present good results, especially if it helps the process of getting stakeholders to understand and appreciate the role of impact evaluation. Sometimes, though, demand can be built despite less-positive results – by special efforts to target the relevant stakeholders. Concerns over potential negative results, bad publicity, or improper handling of the results may reduce demand; sensitivity, trust-building, and creative arrangements may help overcome these fears.

Evaluation capacity development

- Evaluation capacity, especially at a local level, is an important factor in the quality of an impact evaluation that also affects the ability of stakeholders to demand, understand, trust, and utilize the results.
- Capacity building is an iterative process and may improve both demand and quality.

Pursuing easy wins alongside harder challenges

- The most effective strategy for developing a strong culture of evaluation may be two-pronged: opportunism where there are "easy wins" – willing partners, high capacity, good data, good results, etc., since these may require less effort and fewer resources and may generate familiarity with the process; and at the same time "chipping away" systematically at the harder problems where there is less capacity or less tradition of evaluation.

Table 1: The evaluation questions and the main findings for each of the evaluation

Program	Evaluation questions	Main findings
EDUCATION PROGRAMS		
1. **Cambodia: Japanese Fund for Poverty Reduction [JFPR]: Secondary School Scholarship Fund** **Goals:** *Increase enrolment and retention of girls from poor families in lower secondary schools*	▪ Do scholarships increase enrollment of girls from low-income families in secondary school? ▪ Do scholarships increase retention?	▪ Scholarship recipients had significantly lower socio-economic status than non-recipients (so program was reaching the target group) ▪ Recipients had approximately 30 per cent higher enrolment and retention than non-recipients ▪ Effect size much higher than similar programs in other countries (e.g. Progresa in Mexico)
2. **Cambodia: World Bank Girls Secondary School Scholarship Fund** [follow-up to JFPR program] **Goals:** *Improve targeting of low-income girls*	▪ Assess program impacts on: ▪ effectiveness of providing larger scholarships to poorer girls ▪ retention ▪ learning ▪ inter-household issues ▪ child labor	▪ Similar to the JFPR project (increased enrolment and retention but no effect on learning or the quality of education
3. **Uganda: Universal Primary Education (UPE)** **Goals:** *Test the effectiveness of improved management*	▪ Trends in attendance and learning since 2000 ▪ Determinants of trends ▪ Size and cost-effectiveness of each intervention ▪ Use of MIS for evaluation	▪ Progress in access to education ▪ Effectiveness of investments in teachers, classrooms, books and other facilities ▪ School management important ▪ Investments more effective if combined with improved management ▪ Quality of primary education remains poor and absenteeism and drop-outs high

Program	Evaluation questions	Main findings
4. **Uganda: UPE. Pilot program in Masindi District** **Goals:** *Test the effectiveness of improved management*	▪ Effects of improved management ▪ How does this enhance other interventions?	▪ Educational performance in project schools: o 50-60 per cent better than control schools outside the district o 35 per cent than Masindi schools not covered by the project
5. **Chile: Vouchers for private schools** **Goals:** *Improve quality of education by providing low-income students access to private education and stimulating public schools to perform better*	▪ Assessing the effects of vouchers on the quality of education ▪ Were changes due to improved quality or to skimming off better students from the public schools?	▪ No evidence that vouchers and increased choice improved educational outcomes ▪ Vouchers did lead to sorting as better students from public schools more likely to move to private schools
CONDITIONAL CASH TRANSFERS [CCT] AND POVERTY REDUCTION PROGRAMS		
1. **Familias en Accion: Colombia.** Conditional cash transfers promoting children's health and primary and secondary school enrolment **Goals:** *Short-term poverty reduction through cash transfers. Long-term investment in human capital development through increasing access to health and education*	▪ Cost-effectiveness of increasing access of poor children to health and education ▪ Effectiveness of targeting mechanisms in reaching the low-income target population ▪ Replicability of programs on a large scale ▪ Replicability in urban areas of programs developed in rural areas	▪ Increased primary school enrolment in rural but not urban areas ▪ Increased secondary school enrolment in both rural and urban areas ▪ Some improvement in rural nutrition but very limited impact in urban areas ▪ Influence on diarrhea in rural but not urban areas

Program	Evaluation questions	Main findings
2. **Progresa/ Oportunidades: Mexico.** Conditional cash transfers promoting children's health, nutrition and education. **Goals**: *As for Colombia*	▪ Are CCTs cost-effective in increasing access of poor children to health and education? *Effectiveness of key program components*: ▪ Direct monetary transfers versus in-kind grants ▪ Targeting the extremely poor versus all families ▪ New, standard targeting procedures versus existing program client lists ▪ Transfers to households versus to communities ▪ Non-discretionary rules for whole country versus flexibility for local authorities ▪ Directing benefits directly to women versus to household head ▪ Program impacts on fertility ▪ Criteria for defining size of transfer ▪ Merits of family co-responsibility and certification	▪ Poverty targeting worked well[4] ▪ PROGRESA reducesby 10% people living below poverty line ▪ Positive impact on school enrolment for boys and girls ▪ Children entering school earlier, less grade repetition and better grade progression ▪ Younger children have become more robust against illness ▪ Women's role in household decision-making increases ▪ Estimated cost-benefit ratio of 27%
3. **Jefes de Familia, Emergency Safety Net Program: Argentina.** Cash transfer for unemployed household heads with dependent children **Goals**: *Short term goal, using monthly cash transfers to stop families falling into poverty. Longer term goal of developing skills to facilitate re-entry into the labor market.*	▪ Effectiveness of cash transfers as an emergency measure to aid poor families ▪ Are programs cost-effective, efficiently managed and relatively free of corruption? ▪ Effectiveness of targeting procedures. Did they reach the intended groups? ▪ How did households respond to the program? Labor force participation, labor supply and household division of labor ▪ Impact on household income ▪ Impact on aggregate rates of poverty	*Findings on program performance* ▪ Eligibility criteria were poorly enforced – particularly with respect to women not in the labor force ▪ Targeting worked well in practice as eligibility criteria correlated with structural poverty *Findings on program impact* ▪ Prevented 10% of families falling into extreme poverty ▪ Net income gains equal to 50-65% of cash transfer ▪ Foregone income greater for previously employed and for household head than for spouse ▪ 2.5% drop in aggregate unemployment rate
HEALTH		
1. **Kenya: Bed net distribution experiment:** Free vs. Cost-Recovery **Goals**: *Increased distribution and use of insecticide-treated nets*	▪ Is free distribution or cost-recovery more effective for increasing distribution and use of nets? ▪ How price elastic is demand?	▪ Cost recovery did not increase distribution or use ▪ Cost recovery appears to reduce demand

[4] The PROGRESA findings were not reported in the conference but were taken from IFPRI (2002) PROGRESA: Breaking the Cycle of Poverty.

Program	Evaluation questions	Main findings
2. **Kenya: Deworming treatment and worm-prevention health Messages** **Goals**: *Reduced worm infections, increased prevention behaviors, improved schooling outcomes*	▪ Does (school-based) deworming improve worm load? ▪ Does it improve schooling outcomes? ▪ Do health messages on worm-prevention induce the preferred behaviors? ▪ How does cost-sharing affect uptake? ▪ How does social learning affect uptake?	▪ Deworming pills reduce worm loads among treated children and children nearby. ▪ School attendance increased; drop-outs decreased. ▪ There were no changes in worm-prevention behaviors. ▪ Cost-sharing reduced uptake. ▪ Social learning (knowing others who had taken the treatment previously) seemed to reduce uptake.
3. **China: Voluntary Health Insurance Scheme** **Goals**: *Reduced out of pocket healthcare expenditures, increased utilization of needed health services*	▪ Does the health insurance scheme reduce out of pocket expenditures? ▪ Does it increase use of services?	▪ Increased household utilization of health services ▪ No reduction in out-of-pocket payments
SUSTAINABILITY		
1. **Madagascar: ADeFI Microfinance Institution.** Provides credit to small businesses **Goals**: *Assist very small, small and medium business to develop their activities*	▪ Does participation in microfinance improve financial turnover, production, value added, staff, capital and labor productivity and capital productivity?	▪ No impact found
2. **Morocco: Al Amana Microfinance.** Provides credit to urban areas; expanding into rural areas **Goals**: *Provide access to credit for impoverished people*	▪ Activities and sales of enterprises	▪ Uptake rates were low ▪ Additional results still pending
3. **Ethiopia: Food Security Program.** Labor-intensive public works safety-net program, unconditional transfers for certain vulnerable groups, agricultural assistance and technologies	▪ Effectiveness of targeting and delivery of benefits ▪ Impacts on food security and asset growth ▪ Were constructed assets considered useful by stakeholders?	▪ Targeting was successful ▪ Food security was improved ▪ Assets constructed through the public works projects were considered useful ▪ Increased borrowing for productive purposes ▪ Increased use of agricultural technologies ▪ Frequent payment delivery delays

Program	Evaluation questions	Main findings
Goals: *Improved food security and the well-being of chronically food-insecure people in rural areas*		▪ Little overlap among program components, despite intentions
4. **Vietnam: Rural Roads** (1997-2001) **Goals**: *Rehabilitation of rural roads to commune centers, to link communities to markets and reduce poverty*	▪ Did the project fund achieve what it intended – did resources supplement or substitute for local resources? ▪ Impact on market and institutional development	▪ Fewer km of rehabilitated roads than were intended ▪ More new roads built ▪ Improved quality of roads ▪ Access to markets, goods, and services increased ▪ Livelihood diversification ▪ Increased primary school completion ▪ Some short-, some longer-term effects ▪ Larger impacts in poorer communes

Table 2: Summary of the evaluation designs

Sector	Evaluation designs
Education programs (see Table 1 for details)	1. Regression analysis to control for socio-economic differences between the two groups or to compare groups above and below the eligibility cut-off point for the maximum $60 scholarship 2. Propensity score matching to create ex-post control group 3. Quasi-experimental designs in which schools receiving project interventions are compared with schools outside the district; and with schools in treatment districts not receiving the interventions 4. Retrospective (post-test) comparison of scholarship recipients and non-recipients 5. Secondary data sets were used to increase the number of indicators (MIS data) and to analyze learning scores, household socio-economic characteristics, child labor and inter-household issues 6. Triangulation among indicators 7. When programs covered the whole country: natural restrictions or differences in geographical distribution (for example of private schools) used to create comparator group 8. Average school productivity in each commune (district) compared for private and public schools and average productivity estimated for all schools
Conditional cash transfers and poverty reduction programs	1. Randomized selection of beneficiary communities (RCT) for each phase of project 2. Pre-test/post-test comparison group design using propensity score matching and with measurement after one and four years 3. Comparison group divided into those who had starting receiving cash transfers before the baseline and those who had not 4. A propensity-score matching (PSM) design was used with households eligible to be selected for Phase 2 being used as the control group for Phase 1 5. Formal surveys combined with structured and semi-structured interviews, focus groups and workshops
Health	1. Randomization of treatments 2. Randomization, using phased-in project implementation 3. Double difference with matching 4. Integrated into the government's own evaluation and was done in collaboration with government staff
Sustainable development	1. Randomized control trial 2. Double difference with propensity score and/or judgmental matching techniques 3. First evaluation: ex-post matching of beneficiaries and non-beneficiaries 4. Second evaluation: double difference: theoretically robust but high attrition rates left low statistical significance in the results 5. Beneficiaries and non-beneficiaries compared using retrospective data 6. Controls for local conditions, events over time, etc 7. Pre-program baseline data compared with follow-up rounds in three different years
Note: This table summarizes the range of designs used by the evaluations in each sector. The following chapters provide more details on the specific design used for each of the evaluations.	

Table 3: Examples of the influence and use of the evaluations

Use		Examples
Created demand for further and more rigorous evaluations	Created demand for further evaluations	• Cambodia education • Follow-up micro-finance project - Morocco
	Promoted controversy in the academic field and encouraged further research	• Education - Chile
	Created demand for methodologies to evaluate pilot projects	• Follow-up urban project, CCT-Colombia
	Generated follow-up studies	• Assessing service delivery: Food security - Ethiopia • Health insurance - China
	Increased appreciation of the need for independent, external evaluation	• Ethiopia • Mexico
	Helped introduce impact evaluation to particular sectors	• Rural roads - Vietnam
Strengthened quality of impact evaluations	Strengthened MIS and data quality	• Demonstrated to local districts the importance of good data -Uganda education
	Strengthened MIS and data quality	• Demonstrated to local districts the importance of good data - Uganda education
	Encouraged more rigorous evaluation as a standard component of new programs	• Cambodia education
	Lead to evaluation capacity building	• Health insurance - China • Statistics agency and government - Ethiopia
	Raised the standards for evaluation	• Methods and questionnaires used in other road evaluations - Rural Roads, Vietnam
	Institutionalized impact evaluation systems	• Social sector evaluation systems introduced (Mexico) • Created a culture of evaluation (Ethiopia)
	Enhanced the role and rigor of impact evaluation internationally	• CCT-Mexico and Colombia

Contributed to program design and implementation	Identifies which components are /are not effective and improves program operation	• Raised interest in incorporating scholarship programs in government projects- Education, Cambodia • Showed investment in program management more cost-effective than building classrooms or hiring more teachers – Education, Uganda • CCT-Mexico
	Improved design of future projects	• Follow-up urban project: CCT-Colombia • Smaller grants for primary school: CCT-Colombia • CCT-Mexico
	Convinced agencies to design and test pilot projects before going to scale	• Follow-up urban project: CCT-Colombia • Self-employment program - Argentina
	Broadened program and policy options	• New labor market intervention options - Argentina
	Identified administrative and logistical problems that had been overlooked	• Food security - Ethiopia
Provided evidence to support programs	Provided evidence to respond to program critics	• Education -Uganda
	Provided evidence to support programs and justify continuation under new government	• CCT-Mexico and Colombia • Emergency Program-Argentina
	Evaluations used to justify new programs even when findings did not support this	• Government used findings to justify expansion to urban areas: CCT-Colombia • New self-employment program - Argentina
	Provided evidence to continue components agencies had planned to cut	• Community day care centers: CCT-Colombia
	Contributed to replication in other countries.	• CCT-Mexico and Colombia
	Helped agencies decide between alternative strategies	• Free distribution of mosquito nets - Kenya/ Somalia
	Raised the visibility of programs	• Deworming now commonly discussed among international agencies such as WHO and World Bank
Provided evidence to challenge programs		• Extension of CCT to urban areas - Colombia • Self-employment programs - Argentina
Involved wider group of stakeholders	Provided basis for engagement with policymakers and implementers	• Education - Uganda

Table 4: Factors affecting evaluation utilization and influence

A. Factors facilitating evaluation utilization and influence		**Examples**
Timeliness	a. The evaluation must be commissioned and the findings produced when there is current interest in the issues being studied b. At least preliminary findings should be available in time to make adjustments to program implementation	a. There was a demand for information on the questions being addressed (Madagascar) b. An interim evaluation report allowed for mid-program changes (Ethiopia)
Focus on the clients priority issues	a. The evaluation incorporated local contextual data b. Contributed to current policy debate c. Impact reduced when the evaluation does not focus on priority concerns of stakeholders	a. The focus on local contextual issues demonstrated the practical utility of the findings at the district level (Uganda) b. Many national and international agencies were already debating the merits of cost-recovery versus free distribution of bednets (Kenya) c. The evaluation focused on economic issues rather than the social and behavioral factors of concern to government (Madagascar)
Effective communication and dissemination strategies	a. Rapid and wide-spread dissemination of findings b. Clear and well communicated messages c. "No surprises". Ongoing communications and periodic one-on-one meetings to keep stakeholders informed of the progress and initial findings of the evaluation	a. Data was available on Internet and was widely used in academic publications (Mexico) b. The evaluation was rigorous but the findings were communicated in a very technical way that was difficult for non-specialists to understand (Madagascar) c. Due to frequent interactions between evaluators and stakeholders the latter became more comfortable with the evaluation process (Ethiopia)
Active engagement with national counterparts	a. National agencies involved in design and implementation of the evaluation b. Provided mechanism for greater stakeholder involvement c. Reducing costs through coordination with ongoing national surveys	a(i). Evaluators revised evaluation design in response to concerns about RCT (Cambodia) a(ii). The evaluation was commissioned by policymakers not donors (Mexico) a(iii). Ministry of Labor and Bureau of Statistics actively involved (Argentina)

	d. Evaluation integrated into ongoing government evaluation/research program	a(iv). Defined as "A true partnership from the beginning" (Madagascar) a(v). In-country team facilitated communication of evaluation progress and findings (Ethiopia) a(vi). Close cooperation with the Statistics Bureau in preparing the survey instrument (Ethiopia) b. Uganda c. Piggy-backing the evaluation with a Ministry of Labor survey (Argentina) d. Evaluation also conducted in collaboration with national agency staff (China). Recognized government concern about data security and analysis done on government computers.
Ability to demonstrate the value of evaluation as a political and policy tool	a. Findings were of practical utility to managers and policymakers b. Cost-effectiveness analysis proved a useful tool c. Findings demonstrated the programs were reaching the low-income target populations was important to policymakers	a. Ministry had for the first time specific evidence to respond to critics (Uganda education) b. Uganda c. Colombia
Demonstrated the value of good quality data	a. The practical utility of the evaluation findings demonstrated the value of good quality data	a. Local districts saw for the first time the practical utility of good evaluation data (Uganda)
The methodological quality of the evaluation and the credibility of the international evaluators	a. Rigorous methodology "set the bar" for other countries who felt the need to replicate these standards b. Credibility and independence of the international evaluators c. The use of innovative evaluation methodologies creates interest	a. The rigor of the Progresa evaluations and the surrounding publicity convinced Colombia of the need to use equally rigorous evaluations b. Mexico c. This was the first time a RCT had been conducted on micro-finance (Madagascar)
Demand for more rigorous evaluation methodologies	a. Policymakers and line ministries aware of the need for more rigorous evaluation methods	a. Large number of road projects had not been evaluated (Vietnam)
Positive and non-threatening findings	a. Initial evaluations produced positive findings, encouraging agencies to support further evaluation	a(i). Cambodia education a(ii). Uganda a(iii). Colombia a(iv). Mexico

Evaluation capacity development	a. Sequential evaluations permitted national agencies to develop capacity over time	a. Cambodia education
Unexpected findings stimulated interest in further research	a. Controversial findings stimulated interest in further research	a. Findings challenged conventional wisdom by showing there had been no improvement in school performance (Chile)
Demonstrated transparency of the evaluations	a. The ability to demonstrate transparency and professional rigor was important in countries where earlier programs had been criticized for corruption and politicization	a(i). Mexico a(ii). Argentina a(iii). Colombia
Donor pressures	a. Donors need rigorous data to justify continuation or expansion of program b. Donor pressure to ensure collection of baseline data to increase credibility of findings	a. Argentina b. In Colombia, donors pressured government to delay start of program in some areas to permit collection of baseline data
B. Challenges to evaluation utilization and influence		
Data collection often does not happen until the program has been operating for some time	a. This affects the quality of the evaluation	
Multiple donors	a. Affects communication and coordination b. May be difficult to reach consensus on evaluation design	
Variations in technical expertise of stakeholders	a. Difficult to present findings at the right technical level	
Tensions between donors and government	a. Can affect willingness to support evaluation b. Difficult to reach consensus on evaluation designs	
Long time before results are available	a. Outcomes and impacts cannot be measured for a long time. This reduces interest of many stakeholders b. Low technical capacity of patterns may slow process of data collection, analysis and dissemination	
Project staff turnover	a. People who are interested leave and replacements may not be as interested	
Funds for evaluation may be reduced	a. Originally approved evaluation funds may be reduced as project develops	
Not everyone wants accountability	a. Evaluation may be seen as a threat	

2. Education

A. Introduction

The education workshop discussed the evaluation of projects in Cambodia, Uganda, and Chile. The objectives of the education programs in Cambodia and Uganda were to increase school enrolment and retention for low-income students, particularly girls; and in the case of Uganda to also improve education quality. The program in Chile, which already had very high enrolment rates, was intended to improve quality for low-income students through increased access to private education. In addition, all of the programs sought to enhance the efficiency of program management. An overview of the education projects, the key evaluation questions, the main evaluations findings, the evaluation designs; how the evaluations were utilized, their influence on project implementation and policy, and the factors affecting utilization are presented in Chapter 1 (Tables 1 – 4). This chapter provides more detail on each of the education evaluations.

B. Getting girls into school: Evidence from a scholarship program in Cambodia

The program

The program being evaluated was the Japan Fund for Poverty Reduction (JFPR) scholarship program in Cambodia. The program, which began in 2004, awarded scholarships to poor girls who were completing 6th grade, and who wished to enter secondary school. The program tested the efficacy of scholarships as a way to increase secondary school enrolment among girls from low-income families and to encourage them to complete the full three years of lower-secondary school. The rationale for the program is the large literature documenting associations between female education and a variety of social outcomes (e.g., health, nutrition, fertility, and child mortality).

The program covered 15% of all secondary schools and in each a maximum of 45 girls were awarded scholarships. The $45 scholarship was quite large compared to the mean per capita GDP of $300. The "scholarship" program was in fact a conditional cash transfer provided to the family on the condition that the girl is enrolled in school, maintains a passing grade and maintains a high attendance rate.

A follow-up World Bank scholarship program was also discussed in the workshop. This had similar objectives to JFPR, but was able to use a more sophisticated targeting system as students were also scored on the probability of drop-out.

The evaluation (*see Table 5 at the end of the chapter***)**

The purpose of the evaluation was to test the effectiveness of a scholarship/conditional cash transfer program in increasing the transition of girls from 6th grade primary school to

the first year of lower-secondary school. As the evaluation was not commissioned until late in the project, a retrospective (ex-post) evaluation design was used. Two sources of data were used: application forms for the scholarship program (information on parental education, household composition, ownership of assets, housing materials, and distance to the nearest secondary school) and data on school enrolment and attendance collected during unannounced school visits. The analysis compared scholarship recipients (the "treated" group) and non-recipients (the "comparator" group) using regression models.

The evaluation of the follow-up World Bank project used a Regression Discontinuity Design. Girls just above the cut-off line for $60 scholarship eligibility were compared with girls just below the line. The evaluation had access to a richer database and was also able to look at learning, intrahousehold issues and child labor.

The evaluation findings

Scholarship recipients had significantly lower socio-economic status than non-recipients, confirming that the program had been successful in targeting poorer girls. After controlling for household characteristics, it was found that girls receiving scholarships had an almost 30 per cent higher attendance and enrolment rate than non-recipients, and that the effects of the program were greatest for the most disadvantaged girls – poorer, lower parental education and living further from school. These program effects compare favorably with similar programs in other countries. For example, the highly regarded PROGRESA program in Mexico was only estimated to have increased the transition from 6^{th} grade to 7^{th} grade (the first year of secondary school) by 11.1 percent.

The preliminary findings from the follow-up World Bank evaluation also showed that the scholarships affected attendance but did not improve learning.

Evaluation utilization and influence (*see Table 6 at the end of the chapter***)**

The retrospective evaluation of the JFPR, even though it was "messy" because of the limited access to baseline data, did "create an appetite that engendered a demand for the kind of more rigorous evaluation" that was implemented for the follow-up project. The cumulative effect of these two had two immediate effects: Government is planning to incorporate some of the evaluation design features in their own scholarship program, and to incorporate a rigorous evaluation into a large fast-track catalytic fund grant. Several factors increased utilization of the first evaluations and stimulated interest in more rigorous future evaluations. First, even though the original design of the JFPR did not include an impact evaluation, the methodologically "messy" retrospective evaluation was able to produce useful findings in a short period of time. It identified operational issues to address in the subsequent projects and created an appetite for more rigorous evaluations. Second, *the*

> *"At the beginning there was no appetite for evaluation. There was no demand for it. There was no appreciation of it. There was no capacity for it. And while we have overcome some of these barriers, I still think there is a limited capacity to understand and use evaluation directly."*
> Deon Filmer. Development Research Group. The World Bank

fact that the first evaluation showed the project had some positive results created interest among national stakeholders in the use of evaluation as management tool. If the evaluation had not found any positive results it might have been more difficult to convince stakeholders to support future evaluations.

Third, the program and evaluation teams *worked closely with government* to prepare a program design that would facilitate a strong evaluation and produce findings that could be used by policymakers. The Bank's willingness to replace the original RCT with a rigorous but politically less sensitive quasi-experimental design built confidence and increased then likelihood that the results would be utilized.

Several lessons were identified with respect to evaluation utilization. *Developing a demand for and a capacity to generate and use rigorous impact evaluations is a long process that evolves over the course of several evaluations.* The process will often be opportunistic taking advantage of interest and opportunities, even though the first evaluations may be "messy". It is also essential to work closely with national counterparts, to be responsive to political concerns, and every opportunity must be taken to strengthen national evaluation capacity. Finally, in cases where a clearly defined selection cut-off point can be defined and implemented (in this case the score on a poverty/probability of drop-out scale), the regression discontinuity design (RD) can provide a methodologically strong design while avoiding political and ethical concerns about RCTs. There are quite a few programs where RD designs could be considered.

C. Impact Evaluation of Primary Education in Uganda

The program

The purpose of the evaluation was to assess the effectiveness of a number of interventions introduced into the primary education system between 2000-2006 and contributing to the national goal of Universal Primary Education. The interventions included: management improvements, infrastructure, teaching materials and increased number and quality of teachers. These interventions form part of the national "full coverage" education services but were also tested in more depth in the Masindi District Education Development Project.

The evaluation

The central evaluation questions were: How have school attendance and learning achievement developed since 2000? What were the main determinants of these developments? Which interventions have the largest and most cost-effective impact on educational outputs? How effectively has the Management Information System been used for purposes of evaluation?

The evaluation was conducted at two levels: nation-wide and in the Masindi District. The evaluation was based on a program theory intervention model that identified four sets of interventions (school management, infrastructure, teaching materials and teachers)

that would enhance school performance through improving access and learning achievement; and in turn produce a set of welfare outcomes. The outcomes would be affected by local contextual factors that could affect results in each district.

Given the countrywide coverage of the education programs, the many different donors and agencies involved and the large number of contextual factors in each region, it was difficult to define a counterfactual. So a number of different approaches were used: combining different data bases to increase the range of variables included in the analysis, using triangulation to obtain independent estimates of key indicators, and using natural restrictions (e.g., remote rural areas where well educated parents do not have a choice of selecting schools with smaller class sizes); and propensity score matching to create ex-post comparator groups comparable with the intervention groups. In Masindi, a quasi-experimental design was used where schools receiving the project interventions were compared both with a comparator group from outside the district and with schools in the district that did not participate in the project.

The evaluation findings

The main findings of the evaluation were the following:

- Uganda has made enormous progress in improving access to primary education.
- The analysis confirmed the effectiveness of investments in teachers, classrooms, books and other school facilities. It also confirmed that high pupil-teacher ratios and high pupil-classroom ratios have a negative effect on learning achievements.
- There are also significant effects from teacher education and training.
- Head teacher qualification is also important.
- Investments in teachers, classrooms and books are more effective when combined with improvements school and district management. Privately funded schools, which are generally better managed, outperform government schools by 40%.
- The quality of primary education remains poor and absenteeism and dropout pose serious threats to the efficiency and effectiveness of primary education.
- The in-depth evaluation of the Masindi District Project found that educational performance in project schools were 50-60 per cent better than the comparator group from surrounding districts, and 35% better than other schools in Masindi.

Impact of the evaluation

The report was disseminated in a number of ways, including presentations in stakeholder workshops. A presentation at the National Stakeholder Conference in 2007 in Kampala to discuss measures to promote the quality of primary resulted in a pilot project being launched in 10 districts with a "rigorous impact evaluation strategy". At the same workshop, a follow up evaluation was discussed. The final report was also sent to the parliament and stakeholders in Uganda.

During the workshop, the Director of the Education Planning Department of the Ugandan Ministry of Education and Sports identified a number of domestic effects of this evaluation. At the local level, *the evaluation created a very positive response from*

district level officials, who said this was the first time they had received effective feedback about one of their programs.

At the national level, this was the first time the Ministry of Education could respond to Parliament providing concrete evidence of the impacts and cost-effectiveness of the education programs, and refuting criticisms that the money would have been better spent on other social programs. In particular, the evaluation showed that improved management could have a greater impact on education outcomes than simply building more classrooms and hiring more teachers. *By providing an objective basis for engagement with policy makers and implementers, the evaluation encouraged the involvement of a wider range of stakeholders in education sector activities.*

The evaluation has also improved the quality and effectiveness of the Education Management Information System (EMIS). Demonstrating how the information can be used in an evaluation has encouraged central agencies and local authorities to improve the quality of the data they collect. The evaluation also demonstrated the importance of contextual analysis to complement and go beyond the statistical data to understand the particular characteristics of each region and how these affect educational performance.

In the Netherlands, the report was published and sent to the parliament. The results of the report were used in the Netherlands in an extensive evaluation of (Dutch) Africa policy in 2008. One of the workshops of the conference confirmed the importance of management in schools. Also in the Netherlands, there has been a discussion of the low level of achievements in primary schools in Uganda. These findings coming out of the impact evaluation have grounded this broader "quality of primary education" discussion, linking demands to improve pupil and teacher attendance and the reduction of absenteeism to an improvement of the management in schools.

In both countries, the evaluation has contributed to an interest on impact evaluation as a management tool. In Uganda, the evaluation contributed to the mentioned initiative to enhance the quality of primary education with impact evaluation as one strategy for evidence based policy formulation and decision-making. Moreover, several officers have followed a course on impact evaluation, and one officer is doing a PhD on the impact of interventions in the education sector. The Ministry of Education and Sports (MoES) and IOB have started a new (impact) evaluation. This evaluation analyses the impact of primary education on the future of boys and girls through further education and employment opportunities.

D. The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's Voucher Program

The program

In 1981, Chile introduced nationwide school choice providing vouchers to any student wishing to attend private school. More than 1,000 private schools entered the market, the

private enrollment rate increased 20 percentage points, mainly in larger, urban and wealthier communities and a very competitive private schools market developed.

The evaluation

The evaluation examined the widely-held belief that providing vouchers and permitting parents to transfer their children to private schools will increase the effectiveness of the educational system. Two hypotheses are examined: First, private schools are more effective so that allowing children to move to private schools will increase efficiency and second, schools respond to incentives so the provision of vouchers will also encourage public schools to become more effective to avoid losing their students.

The evaluation collected data on most of the 300 communes, each of which has an autonomous government that manages schools and public services, has an average population of 39,000 and an average of 27 schools of which 18 were public, 7 private voucher schools and 2 tuition charging private schools. Three outcome measures were used: mathematics and language test scores; repetition rates; and years of schooling among 10-15 year olds. Students' socioeconomic status was measured using Ministry of Education data, classifying schools based on parents education, and the national household survey data (CASEN), that identifies the school attended by each child covered by the survey, permitting the creation of a detailed socio-economic status school profiles.

Two methodological challenges were addressed. First, how to separate the effects of school *productivity* from the effects of *sorting* (the "best" students leave the public schools and go to the private schools thus increasing average performance in private schools even without any increased productivity). Sorting could produce gains in private schools by depressing performance in public schools, both through skimming off the best students and by reducing peer pressure to perform well in public schools. This problem was partially resolved by computing average productivity effects for all schools in each commune, and while this cannot control for peer effects, it does net out the "direct" effect of changes in each sector's student composition.

The second challenge concerned how to define an adequate counterfactual for a nation-wide program for which all students are eligible to apply? The evaluation took advantage of the fact that private sector voucher schools expanded more rapidly in some markets, so that markets with slower voucher school growth could be used to approximate the counterfactual. This approach has limitations, including the effect of pre-existing differences in the characteristics of different markets, differential concurrent trends, and heterogenous treatment effects that might affect private entry and subsequent achievement growth. Several procedures were used to partially control for these factors[5].

[5] Procedures included: controls for pre-existing and concurrent trends, the identification of instrumental variables that affect the extent of private entry but are ideally uncorrelated with trends in academic outcomes, or with the productivity advantage of the private sector.

Findings of the evaluation

There was no evidence that choice improved average test scores, repetition rates, and years of schooling, but the voucher program did lead to increased sorting as the "best" students in public schools left for the private sector schools.

The evaluation made two contributions to the school choice debate. First, it pointed out the difficulties in determining to what extent observed improvements in school performance in voucher schools can be attributed to increased productivity and to what extent this is due to sorting (skimming off the best students from the public schools). Second, it appears that, due to these complicating factors, the positive effects of school vouchers may be less than claimed by many advocates. The authors stress their findings are exploratory and should not be interpreted as claiming that voucher programs do not work. They do, however, emphasize the need to understand the effects of interventions such as vouchers on the whole of the educational system, and that negative as well as positive consequences must be considered.

Evaluation utilization and influence

To disseminate the findings of the impact evaluation, the paper was published in the Journal of Public Economics and presented in academic and policy conferences. In Chile itself, newspapers discussed the evaluation and a couple of them interviewed the authors.

The main impact of the evaluation was to *promote controversy in the academic literature and to stimulate more evaluations*. The academic community tended to agree with the finding that vouchers increase stratification, but several authors challenged the finding that there was no improvement in school performance – as this goes against established theory. It is not clear, however, what impact, if any, the evaluation had in Chile. While electoral candidates have focused on the lack of quality and problems of stratification in the education system, it is not clear whether they were influenced by the evaluation as the *government was already publishing similar statistics* on the education sector.

Table 5: The evaluation questions and the evaluations designs for each of the education evaluations

Program	Evaluation questions	Evaluation design
1. Cambodia: Japanese Fund for Poverty Reduction: Secondary School Scholarship Fund **Goals**: *Increase enrolment and retention of girls from poor families in lower secondary schools*	▪ Do scholarships increase enrollment of girls from low-income families in secondary school? ▪ Do scholarships increase retention?	▪ Retrospective (post-test) comparison of scholarship recipients and non-recipients. Regression analysis to control for socio-economic differences between the two groups
2. Cambodia: World Bank Girls Secondary School Scholarship Fund [follow-up to JFPR program] ▪ Two levels of scholarship for poorest ($60) and next poorest ($45) girls **Goals**: *Improve targeting of low-income girls*	Assess program impacts on ▪ Effectiveness of providing larger scholarships to poorer girls ▪ Retention ▪ Learning ▪ Inter-household issues ▪ Child labor	▪ Regression discontinuity design comparing groups above and below the eligibility cut-off point for the maximum $60 scholarship. Richer data set also permitted analysis of learning, child labor and inter-household issues
3. Uganda: Universal Primary Education (UPE) **Goals:** *Improve attendance and quality of education through better management, infrastructure, teacher materials and more and better trained teachers*	▪ Trends in attendance and learning since 2000 ▪ Determinants of trends ▪ Size and cost-effectiveness of each intervention ▪ Use of MIS for evaluation	▪ Using MIS and other data bases to increase number of indicators ▪ Triangulation among indicators ▪ Using natural restrictions as form of control ▪ Propensity score matching to create ex-post control group
4. Uganda: UPE. Pilot program in Masindi District **Goals**: *Test the effectiveness of improved management*	▪ Effects of improved management ▪ How does this enhance other interventions?	▪ Quasi-experimental design in which Masindi District schools receiving project interventions were compared with schools outside the district; and with Masindi District schools not receiving the interventions
5. Chile: Vouchers for private schools **Goals**: *Improve the quality of education by providing low-income students access to private education and stimulating public schools to perform better*	▪ Assessing the effects of vouchers on the quality of education ▪ Determining whether changes are due to improved quality or to skimming off the better students from the public schools	▪ Secondary data used to measure math and language scores, repetition rates, average years of schooling and socioeconomic status ▪ Average school productivity in each commune (district) compared for private and public schools and average productivity estimated for all schools ▪ Selection of comparison group difficult as program covered whole country but design took advantage of the fact that public schools grew more rapidly in some areas than in others

Table 6: The possible effects and influence of each education evaluation and the reasons why the evaluations were influential

Influence/ effects of the evaluation	Reasons why influential
Cambodia secondary school scholarship projects: [Summary of Japan Fund for Poverty Reduction and World Bank projects]	
1. Created a demand for further evaluations 2. Evaluation findings raised government interest to incorporate scholarship component in their own project 3. Government decided to include rigorous evaluation component in their own projects	▪ The first evaluation produced positive findings which encouraged central and local government to support further evaluations ▪ The evaluation team worked closely with government on the design and changed the proposed RCT design in response to political concerns. This increased the government feeling of ownership of the evaluation ▪ The sequential evaluations enabled government experience and capacity to gradually develop over time
Uganda Universal Primary Education: [Summary of the National Program interventions and the Masindi District Pilot Project]	
1. The Ministry of Education was able to respond to Parliament with concrete evidence demonstrating the impacts and cost-effectiveness of the education programs. This helped defend the programs from the criticisms that the money would have been better spent on other social programs 2. Previously there had been strong pressure to build more classrooms and recruit more teachers, but the evaluation showed it is often more cost-effective to invest in improving management of the education programs 3. The evaluation involved a wider range of stakeholders, by providing a basis for engagement with policy makers and implementers 4. The evaluation improved the quality and effectiveness of the Education Management Information System (EMIS): showing how data can be used encouraged agencies to improve the quality of data collection 5. The Masindi District evaluation demonstrated the importance of contextual analysis going beyond statistical data to understand how the particular characteristics of each region affect educational performance	▪ The evaluation provided, for the first time, specific evidence and arguments to respond to critics. This enhanced the Ministry's awareness of the value of evaluation ▪ Cost-effectiveness analysis was seen to be a powerful tool, both for political and planning purposes ▪ The evaluation created positive response from district officials as this was the first time they have received feedback about their programs ▪ Provided a mechanism for greater stakeholder involvement ▪ Demonstrated how the EMIS could be used, encouraging agencies to pay greater attention to the quality of the information to put into the MIS ▪ The national level evaluation found positive findings and made the District feel more comfortable working with the evaluation ▪ The study addressed local contextual factors, making the evaluation approach and findings relevant and easy to understand by District officials
Chile: Vouchers for private schools	
1. The finding that there was no improvement in school performance was quite controversial and may have stimulated further academic research 2. Not clear whether the evaluation had any influence in Chile as the Government was already publishing extensive data on the program, and the issues of quality and accessibility to low-income families were already being discussed by politicians	

3. Anti-poverty and conditional cash transfer (CCT) programs

A. Introduction

The second session discussed the evaluation of three large anti-poverty and conditional cash transfer (CCT) programs in Colombia, Mexico and Argentina. The Mexico and Colombia programs both provided cash transfers to low-income families with children on the condition that children enrolled in school and had regular health check-ups and vaccinations. The Argentina Emergency Safety Net program provided cash transfers to unemployed household heads to reduce the risk of families falling below the poverty line, with the requirement of spending four hours per day in community work programs, training or education. However, there was considerable flexibility concerning how strictly the requirements were enforced by each municipality. An overview of the anti-poverty and conditional cash-transfer projects, the key evaluation questions, the main evaluations findings, the evaluation designs; how the evaluations were utilized, their influence on project implementation and policy, and the factors affecting utilization are presented in Chapter 1 (Tables 1 – 4). This chapter provides more detail on each of these evaluations.

B. Evaluating a Conditional Cash Transfer Program: The Experience of Familias en Accion in Colombia

The program

Familias en Accion (FeA) is a conditional cash transfer (CCT) program launched in Colombia in 2001 and funded by the World Bank and the Inter-American Development Bank. It promoted increased access to health and education by providing monthly grants to poor families on the condition that children were brought to the local clinic for regular health check-ups and vaccinations and that children attended school regularly. All payments were made to the mother on the assumption that the money was more likely to benefit children. The program operated in municipalities with populations of less than 100,000 and required a bank branch to which funds could be transferred. Beneficiaries were selected from the lowest stratum of the social security register (Sisben).

The evaluation (*see Table 7 at the end of the chapter*)

A pre-test/post-test comparison group design was used with the comparator groups selected from municipalities ineligible to participate in the program, in most cases because there was no bank branch to handle the funds transfer. The availability of good secondary data permitted the use of propensity score matching to reduce sample selection bias.

The baseline studies were conducted in 2002 with follow-ups in 2003 and 2006. A total of 57 project and 65 control municipalities were sampled with approximately 100 interviews per municipality. Political pressures due to the upcoming elections forced FeA to advance the program launch and families in a number of municipalities had already received payments before the baseline study was conducted. The World Bank and IDB were able to convince Government to delay program launch in some areas until the baseline could be conducted. Consequently the baseline was divided into two groups: those who had not received any payments prior to the baseline and those who had.

The evaluation findings

The first follow-up study (2003). Positive results could already be seen this early stage, particularly in rural areas. The evaluation attracted a lot of attention and findings were widely disseminated through a major conference in 2004 and newspaper editorials. The results of the *Second 2006 follow-up* were similar to the 2003 study: there was increased primary school enrolment (8-12 year olds) in rural but not urban areas; increased secondary school enrolment (12-17 age group) in rural and urban areas; some improvements in nutritional status in rural areas[6], but not in urban areas[7]; and an impact on diarrhea occurrence for younger rural children but not for either rural or urban areas for children over 36 months. A major concern was the lack of effects on anemia, which affects half of all poor children. Reservations were expressed in the report and in conversations with policymakers concerning the extent to which findings from the small municipalities could be extrapolated to urban areas.

Evaluation utilization and influence (*see Table 8 at the end of the chapter*)

The Government *used the results of the evaluation to justify expansion of the program to the urban areas despite the fact that the evaluation found the program much less effective* in urban areas. The government was strongly committed for political reasons to expanding the program to urban areas, and wide publicity was given to the positive results of the program in rural areas to justify the urban areas, resulting in an increase of total beneficiaries from 400,000 to 1.5 million. Although the evaluation did not justify the urban expansion, it did encourage redesigning of various program components. Most importantly, the earlier competition with Hogares de Bienestar Comunitria (HBC) was transformed into broad-based cooperation. Also small pilot interventions were introduced to refine program implementation – replacing the earlier approach of starting the program on a massive scale without time for adequate testing.

[6] There improvements in rural areas on the height per age, chronic malnutrition and weight per age but not for global malnutrition and weight per height.

[7] When the urban population was disaggregated into two age groups, improvement was found on one indicator for the under 36 month population (probability of global malnutrition) but for none of the over 36 months age group.

Lessons learned

Evaluators must adapt evaluation designs to political realities when deciding what evaluation strategies will be both technically sound and politically feasible. Evaluations of large, politically sensitive programs should be designed at an early stage, before the programs have developed a large constituency and become resistant to questioning of their goals and methods. Evaluations should begin early in the program with greater use being made of small pilot projects to assess operational procedures and viability for expansion. Finally, the Colombian experience showed that multilateral agencies can have an important role to play in promoting evaluation and ensuring it is technically sound.

C. The Role of Impact Evaluation in the PROGRESA/ Oportunidades Program of Mexico

The program

PROGRESA, now renamed Oportunidades, is a conditional cash transfer (CCT) program that provides cash directly to low-income families on the condition that children attend school regularly and family members visit health centers. PROGRESA was one of the first CCT programs in Latin America and was influential in the design of later programs in other countries. The program had two main objectives: to produce short-run effects on poverty through cash transfers, and to contribute to long-run poverty alleviation through investment in human capital (i.e. education, health and nutrition). The focus is on children because early interventions have much higher returns over the life-cycle. Payments were made to the mother to increase the likelihood that children would benefit.

The program included a number of innovative features, several of which were considered quite controversial at the time, and all of which were assessed by the evaluations. Some of the measures included: (a) direct monetary transfers instead of providing vouchers or food in-kind, or improving supply side services; (b) the programs targeted the extremely/structurally poor rather than all families; (c) PROGRESA developed a single national roster of beneficiaries rather than working from existing lists; (d) transfers were given directly to households rather than communities; (e) uniform, non-discretionary rules were introduced for the whole country; and (f) there was a requirement of family co-responsibility and certification.

The evaluation

The program began one year before the 1999 Presidential elections and there was pressure from the ruling party (PRI) to ensure that the findings of the evaluation would be available prior to the election. When Vicente Fox was elected in 2000, the new administration continued to support a rigorous, independent evaluation to provide objective evidence that their programs were more effective and transparent than those of the PRI regime that had been in power for the previous 80 years. The rigorous and expensive evaluation systems were justified on three grounds: (a) *Economic:* to improve the design and effectiveness of the programs and to compare impacts and cost-

effectiveness of different programs; (b) *Social:* Increasing transparency and accountability, and (c) *Political:* the evaluations increased the credibility of the programs, and this, combined with increased transparency and accountability, helped break with past practices, such as political influence in beneficiary selection.

The evaluation design: The program was implemented in phases, and for each phase beneficiary communities were selected randomly, with non-selected communities providing a non-biased comparator group. Randomization was politically acceptable because communities not selected in one phase were likely to be included in the next phase. Also the government was strongly committed to the use of rigorous, state-of-the-art evaluation design to ensure credibility of the findings. 24,077 households were interviewed in 320 treatment and 186 control communities. Families were interviewed at the start of the program and at several points during implementation, avoiding problems of linear extrapolation when only one post-test measurement is made.

Main findings of the evaluation

The following are some of the findings highlighted in a 2002 IFPRI report on the first post-test evaluation[8]:

- Poverty targeting worked well.
- PROGRESA reduced by 10% people living below the poverty line.
- Positive impact on school enrolment for boys and girls.
- Children entered school earlier, with less grade repetition and better grade progression.
- Younger children have become more robust against illness.
- Women's role in household decision-making increases.
- The program had an estimated cost-benefit ratio of 27%.

Evaluation utilization and influence

According to the presentation, the evaluation had the following kinds of influence:

- *Continuation of the program under a new administration.* The independence, credibility and positive outcomes of the early stages of the evaluation significantly contributed to the program's continuation under the new administration.
- *Improved operational performance.* The early operations reports identified implementation issues, such as delivery of food supplements and intra-household conflicts, and issues with targeting rules that were addressed as the program evolved.
- *Contributed to program expansion to urban areas.* A youth job creation program (Jovenes con Oportunidades) created income generating opportunities for poor households through preferential access to microcredit, housing improvements, adult education and social/health insurance.

[8] IFPRI (2002) PROGRESA: Breaking the Cycle of Poverty.

- *Contributed to the development of a more systematic policy evaluation approach in Mexico.* This move was formalized by the creation of CONEVAL (Council for Program Evaluation) in 2006.
- *Enhanced policy evaluation internationally.* The evaluation findings were available on the internet and were used widely by academic researchers. The design and findings were able to withstand critical scrutiny, greatly enhancing the credibility and influence of the findings.
- *Contributed to the initiation of CCT programs in many other countries.* Similar programs, most of which have been influenced to some extent by PROGRESA, have been started in at least 10 other countries.

D. Assessing Social Protection to the Poor: Evidence from Argentina

The program

During 2001-2002, Argentina suffered one of its worst macroeconomic crises in recent history, and in January 2002 the government launched an Emergency Safety Net Program (the "Jefes" program), co-financed by the World Bank, which by the end of 2002 reached 2 million beneficiaries. To ensure the program was only attractive to the poor, beneficiaries had to spend 4 hours per day on community work or education programs. The program was targeted for unemployed heads of household with children under 18, who received a cash transfer of 150 pesos (approximately US$ 50) per month. The program was decentralized with the details of eligibility and work/training requirements decided at the local level, causing accusations of political manipulation, and making it more difficult to introduce standardized, implementation procedures.

The evaluation

This was a large and high priority program that was being rapidly scaled-up. In addition to the need to learn about the effectiveness of the program, a rigorous and transparent evaluation was also required to address accusations of abuses and implementation problems. The World Bank also required empirical evidence to justify its financing. The cost of the evaluation was significantly reduced by piggy-backing the evaluation on an existing labor force survey. The policy questions addressed by the evaluation included: How effective was the Jefes program as a rapid and targeted poverty alleviation program and a safety net? Did the program reach the intended groups and how did they respond? Did it mitigate income loss due to the crisis and stop families falling into poverty or extreme poverty? How did it affect aggregate poverty and unemployment rates?

Evaluation methodology: The Ministry of Labor and the Statistical Institute agreed to add questions about program participation to their panel sample. The central research question was to estimate the net impact of the 150 peso monthly transfer on beneficiary household income. Net effect was expected to be less than the gross transfer due to the opportunity cost of foregone earnings. The control group was defined as household heads who had applied for the program but had not yet been accepted. As the project was

implemented in phases it was possible to use a "pipeline" evaluation design, with the control group for phase 1 comprising households selected for phase 2. This reduced selection bias because the control group wished to participate, so their motivation is similar to that of the phase 1 participants.

The evaluation findings

Evaluation findings on program performance: The evaluation found the eligibility criteria were poorly enforced, due in part to the practical difficulties of defining employment status in a country with a large informal sector. However, the targeting procedure worked quite well in practice as the eligibility criteria were correlated with structural poverty and 70 percent of beneficiaries had household per capita income in the lowest two deciles. The survey data was compared with program administrative data to check on allegations of fraud and ghost participants, as well as the practical difficulties of defining and implementing the targeting criteria. While some of the claims of abuse were corroborated, no evidence was found to substantiate many of the accusations.

Evaluation findings with respect to program impacts: About 10% of participants would have fallen below the food poverty line in the absence of the program. Many, particularly male participants, had to forego other income to participate and net income gains were equivalent to between one half and two thirds of the cash transfer. The effect of the program on aggregate poverty rates was quite small and the impact on extreme poverty was only marginal. When the economy began to bounce back in 2003, this significantly increased the opportunity cost of continued participation in the program, and the net gains to those remaining in the program dropped from two thirds to one half of the cash transfer. Half of those exiting the program found employment while about one third (mainly women) returned to their previous economic inactivity.

Evaluation utilization and influence

The rigorous evaluation was made possible due to a combination of factors. First, it provided rapid information on the implementation effectiveness of this high priority program which increased the support of the Ministry of Labor and the Statistics Bureau. The relatively low cost of the evaluation also made the decision to commission the evaluation easier. Given the controversial start of the program with the allegations of corruption and poor administration, there was also strong pressure from the World Bank to conduct an evaluation to justify continued funding. These factors, combined with the rapid, though limited, dissemination to local counterparts meant the evaluation was able to influence government policy in a number of areas: it helped justify continued financing for the Jefes program; identified future policy options, including new supply and demand-side labor market options; pressured the Ministry of Social Development to incorporate more rigorous evaluation and encouraged government to do this for other programs.

Lessons learned

This experience showed that *a well designed evaluation can give credibility to a program and can provide useful and rapid operational feedback and policy guidance*. This can be particularly useful when emergency programs must respond rapidly to challenging circumstances or when allegations of inefficiency or corruption must be investigated. *Close cooperation with national agencies is critical for creating ownership, acceptance and utilization of the findings, for improving the technical quality of the evaluation and for reducing costs through piggy-backing on an existing evaluation.*

Table 7: The evaluation questions and the evaluation designs for each of the anti-poverty and conditional cash transfer [CCT] programs

Program	Evaluation questions	Evaluation design
1. **Colombia**: **Familias en Accion.** CCT promoting children's health and primary and secondary school enrolment **Goals**: *Short-term poverty reduction through cash transfers. Long-term investment in human capital development through increasing access to health and education*	▪ Are CCTs cost-effective in increasing access of poor children to health and education? ▪ Effectiveness of targeting mechanisms in reaching low-income populations ▪ Feasibility of large-scale replicability of programs ▪ Urban replicability of programs successfully implemented in rural areas	▪ Pre-test/post-test comparison group design using propensity score matching. ▪ Comparison group divided into those who had starting receiving cash transfers before the baseline and those who had not ▪ post-test measurements after one year and four years
2. **Mexico: PROGRESA/ Oportunidades.** CCTs promoting children's health, nutrition and education. **Goals**: *As for Colombia*	▪ Are CCTs cost-effective in increasing access of poor children to health and education? *Effectiveness of key program components*: ▪ Direct monetary transfers versus in-kind grants ▪ Targeting the extremely poor versus all families ▪ New, standard targeting procedures versus existing program client lists ▪ Transfers to households versus communities ▪ Non-discretionary rules for whole country versus administrative flexibility for local authorities ▪ Directing benefits to women versus to household head ▪ Criteria for defining size of transfer	▪ Randomized selection of communities for each phase of project ▪ Pipeline design using families not selected for a given phase as the control for that phase ▪ 24,077 households interviewed in 320 treatment and 186 control communities ▪ Formal surveys were combined with structured and semi-structured interviews, focus groups and workshops
3. **Argentina**: **Jefes de Familia, Emergency Safety Net Program**. Cash transfer for unemployed household heads with dependent children **Goals**: *Short-term goal using monthly cash transfers to stop families falling into poverty. Longer-term goal of giving skills to facilitate re-entry into the labor market*	▪ Effectiveness of cash transfers as an emergency measure to aid poor families ▪ Are programs cost-effective, efficiently managed and relatively free of corruption? ▪ Effectiveness of targeting procedures. Did they reach the intended groups? ▪ How did households respond to the program in terms of labor force participation, labor supply and household division of labor? ▪ Impact on household income ▪ Impact on aggregate rates of poverty	▪ A pipeline treatment/control group design was used with households selected for Phase 2 being used as the control group for Phase 1.

Table 8: The possible effects and influence of each CCT evaluation and the reasons why they were influential

Influence/ effects of the evaluation	Reasons why influential
Familias en Acción (FeA): Colombia	
1. Evaluation influential in convincing the new Government to continue the program that had been started by its predecessor 2. Findings used to justify expansion to the urban areas [even though the evaluation findings had shown little impact in urban areas] 3. Findings used to make adjustments to the design and implementation of the urban program 4. Convinced planners to give smaller grants for attending primary school 5. Community day care centers (HCB) integrated into the program rather than being eliminated as previously planned 6. Small scale pilot programs were incorporated to test implementation strategies before going to full scale	▪ The widespread publicity given to the findings of the PROGRESA evaluations convinced Colombian policymakers of the need to introduce an equally rigorous evaluation of FeA ▪ Findings from Phase 1 were widely disseminated and, because of credibility of the international evaluators, widely accepted ▪ Findings were largely positive, making them politically more acceptable ▪ The findings showed FeA was effective in providing service access for the low-income population; and that it was possible to develop transparent and independent systems for ensuring accountability (a Presidential priority) ▪ Pressures from donors concerned that the evaluation findings did not justify the rapid urban expansion, encouraged government to incorporate a rigorous evaluation into the expanded urban program
PROGRESA/ Oportunidades CCT: Mexico	
1. Influenced the continuation of the program under a new Administration 2. Improved operational performance 3. Improved program design 4. Contributed to the introduction of more systematic social program evaluation in Mexico 5. Enhanced the role and rigor of policy evaluation internationally 6. Contributed to the promotion of CCT programs in many other countries	▪ The evaluation was commissioned by Mexican policy-makers, not donors ▪ IFPRI's reputation and independence ensured credibility of the evaluation ▪ Data was available on Internet, and was used in academic publications, increasing international familiarity with methodology and findings ▪ Findings were rapidly and widely disseminated inside and outside Mexico ▪ The findings were strongly positive – making it easier for them to be used ▪ The evaluation showed the new administration how to ensure transparency and show a break with past politicization of major programs
Emergency Safety Net Program ("Jefes"): Argentina	
1. Helped justify continued financing for the Jefes program 2. Provided feedback on future program and policy design and broadened the range of policy options on supply and demand-side labor market interventions 3. The evaluation was used by the Ministry of Social Development to justify a new self-employment program (even though the evaluation findings did not support this) 4. Encouraged the government to build-in an evaluation component to the new self-employment program and to use of evaluation at an early stage of the new program to assess viability of scaling-up	▪ Pressure from donors for rigorous evaluation to justify continued funding ▪ Accusations about corruption and poor implementation created pressure to include an independent and rigorous evaluation component. The rapid feedback on these concerns was considered valuable by Government ▪ The active involvement the Ministry of Labor and the Statistics Institute strengthened understanding of the local context and how the program operated ▪ Piggy-backing the evaluation on a Ministry of Labor survey greatly reduced cost and time requirements and strengthened local ownership ▪ Evaluations were found useful by donors and government both to justify programs they supported and to criticize those they did not

4. Health

The third session discussed the evaluation of three health interventions: insecticide-treated nets and deworming, both in Kenya, and a health insurance scheme in China. The Kenyan projects were both randomized, with the nets experiment distributing insecticide-treated nets to pregnant women at prenatal clinics for free or at subsidized prices, and the deworming intervention offering treatment to children at schools for free or with cost-sharing. The Chinese health insurance scheme was meant to reduce out-of-pocket expenditures for health care. An overview of the health projects, the key evaluation questions, the main evaluations findings, the evaluation designs; how the evaluations were utilized, their influence on project implementation and policy, and the factors affecting utilization are presented in Chapter 1 (Tables 1 – 4). This chapter provides more detail on each of these evaluations.

A. Evaluation of insecticide-treated nets in Kenya

The program and evaluation

This study explored the relative benefits of free distribution and cost recovery practices for maximizing coverage and usage of health products – specifically anti-malarial insecticide-treated nets. In particular, there was interest in the competing effects of higher prices reducing willingness or ability to pay and higher prices increasing the perceived value of the product, potentially reducing resource wastage. The experiment randomized the price of nets offered in 20 prenatal clinics from zero to 40 shillings ($.60), subsidizing the price by 100 to 90 percent, and then compared the uptake and usage of the nets at various prices. The primary evaluation interests were the effects of the various prices on demand for acquiring the nets and on usage a few months after acquiring a net. Uptake and usage rates were multiplied together to measure "effective coverage". The evaluation did not explore any potential loss of quality or service that might occur in association with eliminating cost-sharing outside of the experimental framework.

The evaluation findings

The evaluation found that demand drops very quickly as prices increase, and the highest price offered in the experiment is still lower than the common cost-sharing price. On the other hand, usage didn't vary much across price paid. Combining uptake and usage, free distribution led to a 63 percent coverage rate compared to 14 percent for the highest price group. Women who paid for the nets were not found to have worse health at the time of the prenatal visit, suggesting that ability to pay may be more of a limiting factor than need or willingness to pay and thus that full subsidies may be most effective in maximizing the effective coverage rate. However, it was noted that in Kenya, where the experiment was carried out, extensive efforts have resulted in most of the population being familiar with the benefits of nets.

Evaluation utilization and influence

At the time of the presentation, dissemination was still in the early phases. The results were first presented to the Ministry of Health. The news was well received, as they already preferred free distribution, but they noted that they would have to work to find funding and to convince donors and NGOs in the area.

To some degree, *the evaluation has disseminated itself because of high existing demand for the evidence*. An example was DFID contacting the authors before the paper was finished, to help decide whether to give or sell nets in Somalia, so it had immediate impact on other projects, as well as on the organization's official views.

There seems to have been a mixed response in private foundations. In particular, there were rumors of methodological critiques from a major net-distributing organization. From the local branch of the same organization, however, feedback was received to say that they were really pleased with the evaluation and its results. They were changing their model to dispense nets for free, and the evaluation would help them defend their choice. Since then, the local branch has helped disseminate the study.

This presentation noted some of the broader influence of impact evaluations may be harder to track, such as when evaluations contribute to a larger body of evidence. *An individual evaluation may not be entirely conclusive, but conclusions drawn from an accumulation of evidence may be more difficult to refute*.

Lessons learned

This experience showed that when an evaluation addresses an existing demand for evidence, the results may partially disseminate themselves, as interested audiences seek out the information.

Also, it seems that people tend to trust or distrust evidence based on what they already believe, looking for results that confirm what they believe and looking for ways to discredit contrary information. Perhaps one reason is that it is difficult to distinguish between good and bad evidence. Currently, there is much ongoing work to provide training in measurement and evaluation for donors and policymakers: *when individuals have a greater understanding of impact evaluation, they may be better able to recognize differing qualities of evidence they come into contact with*, allowing individual evaluations to have greater impact.

Similarly, people may not trust evidence (especially evidence contrary to their beliefs) that comes from methods they do not understand, so training in or exposure to impact evaluation as well as the use of easy-to-understand methods may make evaluation results more convincing.

In the end, an individual evaluation may not be entirely conclusive, but conclusions drawn from an accumulation of evidence may be more difficult to refute.

B. Kenyan Deworming Experiment

The program

Because intestinal worms, which can create health problems such as anemia, are expensive to diagnose but inexpensive to treat, the WHO recommends mass treatment in schools where there is high worm prevalence. Implementation often has been difficult, however, because of overlaps between ministries of health and education, as well as some question about the prioritization of worms relative to other health interventions, especially in the absence of evidence on educational benefits. Also, since treatment should be readministered every six months because of reinfection, worm treatment has not always been appealing from a sustainability point of view.

Between 1998 and 2001, the Dutch NGO Internationaal Christelijk Steunfonds Africa (ICS) and the Busia District Ministry of Health implemented the Primary School Deworming Project in 75 rural schools in Western Kenya. The intervention included deworming treatment and some worm-prevention health messages.

The evaluation

Taking advantage of the fact that treatment was rolled out in three phases to accommodate financial and administrative constraints, assignment of school to each of the phases was randomized for evaluation purposes. The final phase introduced cost-sharing to compare uptake between cost-sharing and free distribution. The data came from student and school questionnaires fielded in early 1998, 1999, and 2001 and a parent questionnaire added in 2001. The evaluation compared the groups that had been treated to those who would be treated in later rounds, with the last round adding the cost-sharing element to part of the early treatment groups.

The evaluation findings

The results showed that treatment increased school participation by an average of 15 percent of the school year, between reduced dropouts and higher attendance rates. Benefits accrued both to treated children, as well as to children nearby, because of reduced worm loads in the area. Indicators related to the health messages such as clean hands, wearing shoes, and not swimming in the lake showed no impact. Cost-sharing decreased program uptake, from 70 to 15 percent. Overall, deworming was found to be a cost-effective way of achieving schooling attendance, though no impact was found for test scores. The evaluation did not speculate on any trade-offs between eliminating cost-sharing and maintaining service quality with reduced available funds.

Evaluation utilization and influence

The findings were first disseminated to ICS and local school officials and the Ministry of Education. As a result, ICS expanded the deworming program from 75 schools to an entire province instead of moving to other types of programs. The Ministry of Education incorporated deworming into the national education plan and dropped the cost-sharing component.

The evaluation has been *actively disseminated* outside of Kenya as well, with joint efforts among the researchers, the Poverty Action Lab, the World Bank, and others. The evaluation has appeared in academic and policy publications and has generated significant interest in academic and development circles. At the same time, it has reached mass media, with mention by the US president and in the *NY Times*, for example.

As a result, a low-profile health challenge has received more recognition as both a health intervention and an education intervention. At the same time, only a fraction of those who need deworming get it, and it is still not high profile compared to other health concerns such as malaria or HIV/AIDS, and there remain bureaucratic challenges and the simple fact that treatment must be repeated.

Lessons learned

Achieving such widespread coverage may partially be the product of a fortunate combination of factors – *high-quality evaluation, high-profile authors, and surprising and compelling findings* (deworming increases school attendance and even has spillovers to children who aren't treated) – but *considerable and cooperative advocacy efforts play a key role as well*.

Perhaps it is not surprising that, in this case, the stakeholders were willing to use the findings, given their previous willingness to randomize for evaluation purposes. That is, *working with cooperative stakeholders may increase the likelihood that the evaluation will be influential*.

Despite the combination of favorable factors and cooperative stakeholders, deworming is unlikely to become a global health priority. *It seems that certain kinds of problems or interventions may have some limit to what their maximum impact can be, depending on the nature of the issues or interventions they pertain to.*

C. China: Voluntary Health Insurance Scheme

The program

After the collapse of China's Cooperative Medical Scheme in the 1980s, health facilities were allowed to charge government-determined prices for certain high-tech services in order to cross-subsidize more basic care. There was evidence, however, that these changes led to over-application of high-tech services and high out-of-pocket payments,

reducing use of needed services. In 2003, therefore, the government of China implemented the New Cooperative Medical Scheme – a voluntary (but heavily "promoted") rural primary health insurance scheme – to reduce out-of-pocket payments and encourage use of care. While the scheme is heavily subsidized, the coverage is fairly small relative to average annual health expenditures in rural China. Initially introduced in three counties per province, the scheme is meant to reach the entire country by 2010.

The evaluation

Initially there was some resistance to having an external evaluation because the government was conducting its own evaluation but, in the end, an agreement was reached that the external impact evaluation would be conducted cooperatively with government statisticians and would serve as an input into the government's evaluation. The government statisticians had strong survey experience despite limited familiarity with impact evaluation techniques.

The key evaluation questions focused on utilization of inpatient and outpatient services, out-of-pocket expenditures, and facility revenues. To examine these, the external team preferred a double difference with matching approach, using non-participant counties for comparison. The government counterpart preferred not to survey non-participant counties, using a comparison between insured and uninsured in participant counties and regression analysis to control for differences. In the end, data were collected in participant and non-participant counties, but because of non-comparability, the final analysis used a double difference with matching between insured and uninsured in participant counties.

The evaluation findings

The findings from household data showed that utilization had increased, but out-of-pocket payments did not decrease. Facilities data confirmed the increased utilization of services and found that revenues had increased more than utilization. These results showed that medical insurance is not guaranteed to decrease expenses, leading to questions about the level of care provided and whether services were selected because of medical necessity or for revenues.

Evaluation utilization and influence

Dissemination efforts included a report for the Chinese government in Chinese that included the joint findings and analysis by the government statisticians as well as a jointly-written scientific journal article. Initially, the findings were not well-received, perhaps because of the news that the primary objective of reducing out-of-pocket payments had not been achieved; however, after *internal discussions to review and explain the findings*, the government became more comfortable with the results, and the report was incorporated into the larger government. In order to address program and wider health problems raised in the evaluation, a committee was formed to include various ministries, international organizations, and other consultants and, in January

2008, the government announced a number of reform measures that included additional funding for the health insurance scheme.

Another important impact of this evaluation was capacity building in government, especially among the statisticians. They reported *that the cooperative experience had not only taught them impact evaluation principles but had also given them hands-on practice working on a real evaluation*. It has also generated more government interest in using impact evaluation in the future, leading to additional study of impact evaluation methods and consideration in the design of future surveys.

Lessons learned

An essential lesson in this evaluation is the value of the relationship among the stakeholders and the evaluators. *The choice of partners is important, and there has to be a relationship of trust*. In some cases, the trust may already exist, especially when there is already high familiarity with impact evaluation. This is not always true, however, and even where a government or organization is comfortable with impact evaluation, there may be other concerns about the potential results. In these situations – perhaps in any situation – it is necessary to take time and effort to build trust and to handle the process and the results with sensitivity. When there are "bad" results, the proper means and context for presentation and discussion may make the difference between a rejection or suppression of the results and beneficial reforms and future use of impact evaluations and other evidence for policy making.

Cooperation with the "clients" of an evaluation cannot begin too early. In this case, involving the government in the choice of survey design helped to ensure there was comfort with the evaluation methods and eventually the results – increasing utilization.

The cooperation with the local counterparts not only builds trust but also capacity. *Skills and lessons learned during one impact evaluation can be applied to future evaluations, and clients may begin to seek out new opportunities to apply these skills.*

Table 9: Summary of the programs evaluated, the evaluation questions, design and main findings - HEALTH

Program	Evaluation questions	Evaluation design
1. **Kenya: Bed net distribution experiment: Free vs. Cost-Recovery** **Goals**: *Increased distribution and use of insecticide-treated nets*	▪ Is free distribution or cost-recovery more effective for increasing distribution and use of nets? ▪ How price elastic is demand?	▪ Randomization of insecticide-treated net prices at prenatal clinics
2. **Kenya: Deworming treatment and worm-prevention health messages** **Goals**: *Reduced worm infections, increased prevention behaviors, improved schooling outcomes*	▪ Does (school-based) deworming improve worm load? ▪ Improve schooling outcomes? ▪ Do health messages on worm-prevention induce the preferred behaviors? ▪ How does cost-sharing affect uptake? ▪ How does social learning affect uptake?	▪ Randomization, using phased-in project implementation
3. **China: Voluntary Health Insurance Scheme** **Goals**: *Reduced out of pocket healthcare expenditures, increased utilization of needed health services*	▪ Does the health insurance scheme reduce out of pocket expenditures? ▪ Does it increase use of services?	▪ Double difference with matching ▪ Completed as an input into the government's own evaluation and was done in collaboration with government staff

Table10: The possible effects and influence of each evaluation and the reasons why the evaluations why the evaluations were influential - HEALTH

Influence/ effects of the evaluation	Reasons why influential
Kenya – Insecticide-treated nets	
1. Reinforced government's and NGO's decision to distribute nets for free 2. Influenced donor's (DFID) choice between free and cost-sharing distribution of nets in Somalia 3. Results seem to have been questioned among some private foundations, possibly limiting impact	▪ There was already interest in the subject – some organizations and donors were trying to decide whether to pursue free distribution or cost-sharing, and others were looking for evidence to support the decisions they had made ▪ Results contrary to some existing preferences may have led to questions on methodological soundness
Kenya - Deworming	
1. Program has been expanded, and government has discontinued cost-sharing practices 2. Deworming has become more commonly-discussed among international organizations such as WHO, World Bank, IMF 3. Deworming is now considered an education intervention	▪ There was a combination of a high-quality evaluation, high-profile authors, and compelling findings ▪ There have been concerted advocacy efforts to promote findings
China – Health Insurance	
1. Training and capacity building in impact evaluation analysis for Chinese government team 2. A committee was formed for follow-up and reforms have been announced; more resources have been allocated to the program	▪ Completed as an input into the government's own evaluation and was done in collaboration with government staff ▪ Were flexible to address government concerns on the security of their information (analysis done using government data on government computers, for example)

5. Sustainable Development

A. Introduction

The fourth session discussed the evaluation of three sets of sustainable development interventions: microfinance programs that provide small loans to poor individuals in Madagascar and Morocco; the Food Security Program implemented as a safety net response to a drought in Ethiopia; and the rehabilitation of rural roads in Vietnam. An overview of the sustainable development projects, the key evaluation questions, the main evaluations findings, the evaluation designs; how the evaluations were utilized, their influence on project implementation and policy, and the factors affecting utilization are presented in Chapter 1 (Tables 1 – 4). This chapter provides more detail on each of these evaluations.

B. Impact Evaluations of Microfinance Institutions in Madagascar and Morocco

The programs

ADEFI and Al Amana are two microfinance institutions in Madagascar and Morocco, respectively, that receive financing from the French Development Agency (AFD). ADEFI (Action pour le Développement et le Financement des Micro-Entreprises) was created in 1995. With six regional branches and 31 commercial agencies, it is a mutualist scheme that provides loans and savings services to urban micro-businesses in Madagascar.

Al Amana is the largest microfinance institution in Morocco. Originally serving only urban microenterprises when it opened in 1998, the decision was made to explore expansion into rural areas. Starting in 2006, 60 new branches were opened in 80 rural districts. For two peripheral villages in each district, one was randomly assigned a branch, with the other being phased in a year later.

The evaluations

For ADEFI, there were actually two evaluations. A first iteration was done without any specific evaluation questions in mind and was not considered rigorous enough, so a second impact evaluation was conducted using a double difference approach against a counterfactual group of non-client micro-businesses. Key evaluation questions for the second evaluation involved the impact of microfinance on indicators such as financial turnover, production, value added, staff, capital and labor productivity.

The subsequent Al Amana impact evaluation was commissioned by the organization itself, as it wanted to know the benefit of expanding into rural areas. In particular, the evaluation considers effects on agricultural and non-agricultural activities, income and

expenditures, and household security. The expansion process was designed to allow for a nationally-representative randomized control trial evaluation approach. As such, the study was the first of its kind for microfinance. Data included a pre-program survey and follow-ups after one and two years.

The evaluation findings

The results of the second ADEFI evaluation showed no impact on the participating micro-businesses. There were concerns, however, that high attrition led to low statistical significance and thus few policy implications. Thus far, the findings of the Al Amana evaluation have shown low program uptake, but the rest of the results are still pending.

Evaluation utilization and influence

The ADEFI evaluation itself seems to have had only minimal impact. First, it was disseminated only to direct stakeholders – the ADEFI and AFD, with very little readership within AFD. The bigger problems, however, seem to have been the content of the evaluation itself, which failed to appeal to its intended audience. The organization was interested more in social and behavioral rather than economic impacts, and the staff considered the methods "too statistical". The lack of clear policy implications meant that only a few minor, less-central recommendations were implemented.

In this case, however, the evaluation process proved to be useful. Lessons learned in Madagascar were applied to the evaluation in Morocco, which was considered to be much more successful. In Morocco, ongoing dissemination has involved regular meetings with AFD's operational unit in charge of microfinance projects and with other microfinance institutions in Morocco, as well as intermediary reports published and posted on AFD and Al Amana's websites. Conferences have been held for micro-finance practitioners. Planned dissemination includes an additional conference in Morocco, policy briefs and working papers, and academic articles.

Even before the final results are delivered, the Al Amana evaluation has provided some useful operational feedback and generated interest in gathering additional evidence. From the preliminary finding that take-up has been lower than expected in rural areas, Al Amana plans to adapt the design of its loans. Interest in the study has prompted a complementary study to investigate how rural households finance activities, to better design new financial products.

Lessons learned

A number of factors contributed to the evaluation's influence. First, *the subject was relevant and timely*. The organization was looking to extend credit to rural areas, and there was a question about the need and the benefits. There were no other RCT studies on microfinance, so *the evaluation benefited from a degree of novelty and recognized rigor*.

Second, there was "*true partnership from the beginning*". The organization whose project was being evaluated wanted the impact evaluation, and there were regular meetings among stakeholders. In particular, it proved useful to have geographic proximity between research team and the organization, improving communication.

Third, an *emphasis on clarity and rigor* meant that methods and results were trustworthy and understandable. Randomization is widely accepted and the technique is not difficult to explain. From the beginning of the evaluation, precise questions were identified, so the impact evaluation could be well focused.

Finally, *dissemination and high visibility were prioritized*. Active dissemination was planned from the beginning, and the choice of institution and evaluators was strategic. Al Amana is a leading microfinance organization and the largest in Morocco, and evaluators included high-level academics, to ensure that results would be published and read internationally.

C. Ethiopia's Food Security Program

The program

In response to a drought in 2001-2002, the Ethiopian government chose to reform the delivery and quality of the safety net system for vulnerable populations. The resulting Food Security Program comprises the Productive Safety Net Program, involving labor-intensive public works to construct productive community assets and a small number of cash transfers to particular vulnerable groups, and the aptly-named Other Food Security Program, providing agricultural assistance. The combination of these two components was meant to provide a safety net for emergency need while hopefully building long-term productivity and thus reducing poverty and vulnerability.

The evaluation

Because a log frame had been completed for the program, there was ready information on what it was intended to achieve and how, clarifying the evaluation questions: "It was clear to us what we were being asked to measure, what outcomes that were of particular interest to the government". These measures included the project objectives: food security, asset growth, and perceived usefulness of the works being constructed by the program. The evaluation also offered quick feedback on the implementation process – that is, it investigated the effectiveness of the targeting mechanism and the degree to which payments had actually been delivered to the intended recipients.

Because there was some debate in the beginning about whether to conduct an impact evaluation, no baseline data were collected before the program began. As a second-best option, the evaluators collected retrospective data from beneficiaries and non-beneficiaries. The evaluation then used a double difference with matching on the retrospective data. Between the lack of true baseline data and the fact that purposive

targeting with national coverage does not lend itself to the identification of an easy counterfactual, the rigor of the evaluation may have been less than ideal.

The evaluation findings

In terms of process, the evaluation found that targeting was effective, and the assets constructed through the public works projects (roads, soil and water conservation) were considered useful. However, there were notable delivery shortcomings, including delays in payments and a lack of overlap among program components. These created some questions about how to define "participation" for evaluation purposes.

As for outputs, the evaluation found evidence that food security was being improved, and among program participants there were increases in borrowing for productive purposes and use of agricultural technologies. It did not appear, however, that household assets had grown.

Evaluation utilization and influence

The Ethiopian impact evaluation was *particularly interactive among the stakeholders throughout the process*: in commissioning the research, setting priorities, and dissemination. Having many donors meant that coordination and communication among them and with the government and the evaluators was difficult. At the same time, it was hard to reach everyone with the same presentation because of differing levels of expertise. These challenges are common ones. In this case, the team chose to deal with them by using frequent meetings, many of them one-on-one, which *built trust and understanding*. Having an in-country team member helped facilitate *consistent communication*. Also, because all parties were involved throughout the entire process – "no surprises" – results were considered fairly binding and acceptable.

The interim evaluation was able to have direct and immediate influence on the program itself, in part because there was a receptive audience of government and donors who wanted to understand the results in order to help the program succeed. It shifted attention away from political matters and toward some of the administrative and logistical practicalities that were being overlooked (such as when to graduate participants). *The results it offered were relevant*, guided by the government's own log frame. It provided the results in a *timely manner*, too – before the program had ended. As a result, there were adjustments in procedures and measures, as well as follow up studies to explore delivery challenges.

Beyond impacts on the program, the evaluation experience increased government appreciation of having external evaluations to better understand programs, contributing to more of a "culture of evaluation". Also, *the evaluation team worked closely with the Statistics Agency in implementing the questionnaire, building the capacity of the Statistics Agency and government.*

Lessons learned

Starting the evaluation data collection after the evaluation may have reduced the quality of data collected and thus the insights that could be gained from this evaluation. *However, given the set of limitations, this case may be a particularly good example of how, with careful management and particular attention to communication and dissemination, even a less-than-ideal evaluation can prove to be very useful, especially if it is timely and addresses urgent questions.*

D. Rural Roads in Vietnam

The program

Between 1997 and 2001, the Vietnam Rural Transport Project I was designed to distribute funds for the rehabilitation of rural roads to commune centers in 18 provinces, with the objectives of linking communities to markets and reducing poverty. Participant communes were selected by the provinces, within minimum population-density requirements and maximum cost limits.

The evaluation

The evaluation explored whether or not the project funded what it intended – that is, whether resources supplemented or substituted for local resources designated for roads and road rehabilitation – and the impact on market and institutional development, as well as whether road might stimulate local markets or increase access to more distant markets. Key measurements included outputs such as kilometers of rehabilitated roads, new roads (which the project was not intending to fund), and road quality, as well as outcomes such as access to markets, livelihoods, and even school completion. It also examined heterogeneity of impact, particularly whether diminishing returns for villages that started off better off.

The evaluation employed a double-difference approach with propensity score match and included controls for local conditions and events over time. Data came from the Survey of Impacts of Rural Roads in Vietnam panel of 200 communes and 3000 households and included a pre-program baseline in 1997 and follow-up rounds in 1999, 2001, and 2003.

The evaluation findings

The evaluation showed that the project had resulted in more kilometers of rehabilitated roads though fewer than were expected. More new roads had been built as well, however, suggesting that the additional funding allocated to roads had primarily "stuck" to the roads sector, though not completely to rehabilitation as intended. The project improved road quality as well. The overall effect was that the project funding resulted in additional spending on roads instead of displacing regular spending on roads. As a result, access to markets, goods, and services increased, and there has been livelihood diversification.

Primary school completion has also improved. Some impacts such as demand for unskilled labor appeared in the short term but disappeared in the longer term; other impacts took longer to appear. In general, poorer communities benefited from larger impacts.

Evaluation utilization and influence

Dissemination of this impact evaluation thus far has included a number of published academic and working papers and presentations in Washington, DC; at the Ministry of Planning and Development in Mozambique; and at the Transport Research Board 2008 Annual Meeting.

The impact of the evaluation on the project itself has been low. This type of evaluation takes a long time, especially when capacity is low. In general, however, no matter how much time may be required for data collection and analysis, some projects generate impacts that take time to manifest. Additionally, the decision was made to take time to make the evaluation thorough and improve accuracy.

In this case, the benefits of the evaluation are accruing to other projects and other evaluations because of its subject and quality. Compared to health and education, impact evaluations of rural roads – and infrastructure interventions in general – are more difficult and therefore much less common. The implementation and dissemination of this evaluation has thus helped integrate impact evaluation into the roads sector and generated interest in other infrastructure evaluation: there has been high demand for information on the methods and data needs. Practitioners may be able to relate better to evaluations of interventions more similar to their own, with methodological considerations and constraints more similar to those that they face.

The quality of the evaluation has also raised the standard for rural road evaluations, which is expected to increase the quality of information future evaluations will provide. At the same time, it has made it easier for others to follow its example. Methods and questionnaires have been used for other road evaluations.

Lessons Learned

For impact evaluations, *there is often a trade off between speed and quality*. The impact of higher quality evaluations may not be seen in the actual intervention being evaluated, but the benefits may extend to other projects and evaluations by pushing the frontier of what can be evaluated and how and by setting new expectations for evaluation quality.

Timing may create another concern, if a project or evaluation takes long enough to undergo staff changes. Staff may have little incentive to start an evaluation that they will not be around to see completed (or get credit for) and, alternatively, one person interested in impact evaluation may be replaced by someone with different priorities. Staff turnover may thus result in low team interest, and low team interest often results in low government interest in evaluation.

Also, there may be special challenges that can suppress demand in sectors and regions that are less commonly evaluated. Where there is little habit or "culture of evaluation", there may be less funding and less pressure to evaluate, and perhaps higher resistance to accountability. It may require special efforts to begin to build a culture of evaluation.

Table 11 Summary of the programs evaluated, the evaluation questions, design and main findings - SDN

Program	Evaluation questions	Evaluation design
1. **Madagascar - ADeFI Microfinance Institution:** Provides credit to small businesses **Goals**: *Assist very small, small and medium business to develop their activities*	▪ Does participation in microfinance improve financial turnover, production, value added, staff, capital and labor productivity and capital productivity?	▪ First, ex-post matching of beneficiaries and non-beneficiaries. Not considered rigorous enough ▪ Second evaluation: double difference. More robust, but results were of little use – high attrition rates left low statistical significance in the results
2. **Morocco - Al Amana Microfinance:** Provides credit to urban areas; expanding into rural areas **Goals**: *Provide access to credit for impoverished people*	▪ Activities and sales of enterprises	▪ Randomized control trial
3. **Ethiopia - Food Security Program:** Labor-intensive public works safety-net program, unconditional transfers for certain vulnerable groups, agricultural assistance and technologies **Goals**: *Improved food security and the well-being of chronically food-insecure people in rural areas*	Process: ▪ Effective targeting ▪ Delivery of benefits Project objectives: ▪ From project log frame ▪ Food security ▪ Asset growth ▪ Perception that assets being constructed were useful	▪ Double difference with matching techniques ▪ Beneficiaries and non-beneficiaries were compared using retrospective data ▪ There were challenges in defining "beneficiaries" because of program delivery gaps
4. **Vietnam: Rural Roads** (1997-2001) **Goals**: *Rehabilitation of rural roads to commune centers, to link communities to markets and reduce poverty*	▪ Did the project fund what it intended – did resources supplement or substitute for local resources? ▪ Impact on market and institutional development ▪ Heterogeneity of impact	▪ Double-difference with propensity score matching ▪ Controls for local conditions, events over time, etc ▪ Used pre-program baseline data in 1997 and follow-up rounds in 1999, 2001, and 2003

Table 12 The possible effects and influence of each evaluation and the reasons why the evaluations why the evaluations were influential - SDN

Influence/ effects of the evaluation	Reasons why influential or not
Madagascar Microfinance: ADEFI and AFD	
1. Limited policy use 2. A few project changes were made, but these were not strongly linked to recommendations made by the evaluation 3. Lessons from the evaluation process were learned for a second set of impact evaluations in Morocco	▪ No dissemination beyond direct stakeholders, no efforts to plan for dissemination from the beginning ▪ No clear message for policy implications. Methods and results were hard to understand, and "too statistical" for the institution's management to understand ▪ Evaluation asked different questions than what the stakeholders were interested in ▪ The impact evaluation went unread by donors, for the most part ▪ These lessons were learned for the next evaluation
Morocco Microfinance: Al Amana and AFD	
1. Initial results have led to adaptations in Al Amana's loans 2. Interest in the study has prompted a complementary study to investigate how rural household finance activities, to design better financial products 3. [Final results of evaluation still pending]	▪ Subject was relevant and timely, there was existing demand ▪ Was first RCT on microfinance ▪ "True partnership from the beginning" – There was much cooperation and communication among stakeholders, and there was an in-country evaluation team member ▪ Clear and rigorous evaluation methods ▪ Dissemination and visibility are prioritized
Ethiopia: Food Security Program	
1. The interim evaluation brought attention to some of the administrative and logistical practicalities that were being overlooked 2. Generated follow up studies to explore delivery challenges 3. Sparked administrative dialogue on practicalities not considered initially (such as when to graduate participants), led to adjustment of procedures and measures 4. Increased government appreciation of having an external evaluation 5. Contributed to a more of a "culture of evaluation" 6. Capacity building for the Statistics Agency and government 7. Results were considered fairly binding and acceptable	▪ Timing: an interim evaluation allowed for mid-program changes ▪ There was a receptive audience of government and donors that wanted to understand the results, because of their commitment to see the program succeed ▪ "No surprises": ongoing communication, frequent and often one-on-one meetings, with space to address concerns. As a result, stakeholders became more comfortable with each other, the evaluators, and the results ▪ Close collaboration with stakeholders in the evaluation process, allowing for heavy inputs into the design Challenges: ▪ Lack of baseline may have reduced the quality of data collected and thus the insights that could be gained from the evaluation ▪ Many donors meant that coordination and communication was more difficult. Variation in expertise made it hard to present "at the right level" ▪ Government and donor tensions

Vietnam: Rural Roads Rehabilitation Project	
1. Has helped introduce impact evaluation to the infrastructure sector 2. Raises the standards for rural road evaluations 3. Methods and questionnaires have been used for other road evaluations 4. Low impact on the project itself 5. Dissemination is still in the early phases	▪ There has been high demand for dissemination on the methods and data needs for rural road evaluations, especially because infrastructure evaluations have been rare ▪ The infrastructure sector practitioners may have a harder time relating to more commonly- and easily- conducted types of evaluations, such as health Challenges: ▪ This type of evaluation takes a long time, especially when capacity is low, but generally when impacts take time to manifest ▪ Project staff turnover means that one person may be interested in the evaluation while the next may not. Low team interest leads to low government interest ▪ There were less funds and less pressure for impact evaluation when the project started ▪ Not everyone wants accountability!

6. Lessons learned: Strengthening the Utilization and Influence of Impact Evaluation

A. How are impact evaluations used?

Impact evaluations can be used as an *assessment tool* to help strengthen project and program design by providing a more systematic, rigorous, and quantifiable assessment of how a project has performed, what it has achieved (compared to its intended objectives), who has and has not benefited, and how the costs of producing the benefits compare with alternative ways of using the resources.

Impact evaluations are also used as a *political tool* to provide support for decisions that agencies have already decided upon or would like to make, to mobilize political support for high profile or controversial programs and to provide political or managerial accountability. This latter function has been important in countries where new administrations were seeking to introduce transparency into the design and implementation of high profile, politically attractive programs. Impact evaluations can also provide independent corroboration and political cover for terminating politically sensitive programs – in which case the international prestige and independence of the evaluator was found to be important. In fact, in the end it is likely to be the potential political benefit or detriment that causes decision makers to embrace or avoid evaluations, and those who would like to promote impact evaluation as an assessment and learning tool will have to be fully aware of the given political context and navigate strategically.

B. What kinds of influence can impact evaluations have?

The twelve impact evaluations discussed in this report were utilized and had influence in three broad areas: project implementation and administration; providing political support for or against a program; and promoting a culture of evaluation and strengthening national capacity to commission, implement, and use evaluations. It is not only the findings of an impact evaluation that can have an impact. The decision to conduct an evaluation, the choice of methodology, and how the findings are disseminated and used can all have important consequences – some anticipated, others not; some desired and others not. For example, the decision to conduct an evaluation using a randomized control trial can influence who benefits from the program, how different treatments and implementation strategies are prioritized, what is measured, and the criteria used to decide if the program had achieved its objectives.[9]

[9] A frequently cited example from the US was the decision to assess the performance of schools under the No Child Left Behind program in terms of academic performance measured through end-of-year tests. This meant that many schools were forced to modify their curricula to allow more time to coach children in how to take the tests, often resulting in reduced time for physical education, arts, and music.

The influence of evaluations can be seen in administrative realms such as program design and scope or the political realm in the form of popular support for a program or its associated politicians. Understanding the role of impact evaluation is also a process that evolves as managers, policymakers and other stakeholders become more familiar with how evaluations are formulated, implemented and used. For high profile programs, the influence of the evaluation may also be seen in how the debate on the program is framed in the mass media.

C. Guidelines for strengthening evaluation utilization and influence

The following is a synthesis of the broad range of factors identified in the presentations as potentially affecting evaluation utilization.

Timing and focus on priority stakeholder issues:

- The evaluation must be timely and focus on priority issues for key stakeholders. Timing often presents a trade-off: on the one hand, designing an evaluation to provide fast results relevant for the project at hand, in time to make changes in project design and while the project still has the attention of policymakers. On the other hand, evaluations that take longer to complete may be of higher quality and can look for longer term effects on the design of future projects and policies.
- Cooperation with the "clients" of an evaluation cannot begin too early. In this case, involving the government in the choice of survey design helped to ensure there was comfort with the evaluation methods and eventually the results – increasing utilization.
- The evaluator must be opportunistic, taking advantage of funding opportunities, or the interest of key stakeholders. The evaluators must work closely with national counterparts and be responsive to political concerns. Several countries that have progressed toward the institutionalization of evaluation at the national or sector level began with opportunistic selection of their first impact evaluations[10].
- The evaluator should always be on the look-out for "quick-wins" – evaluations that can be conducted quickly and economically and that provide information on issues of immediate concern. Showing the practical utility of impact evaluations can build up confidence and interest before moving on to broader and more complex evaluations.
- There is value in firsts. Pioneer studies may not only show the impact of the intervention, but in a broader context they may also change expectations about what can and should be evaluated or advance the methods that can be used. Even less-than-ideal evaluations that are the first or early in their context can build interest and capacity for impact evaluation.
- A series of sequential evaluations gradually builds interest, ownership and utilization.

[10] See IEG (2008) *Institutionalizing Impact Evaluation within the Framework of a Monitoring and Evaluation System*. The Education for All evaluations in Uganda were cited as an example of institutionalization at the sector level and the SINERGIA evaluation program under the Planning Department in Colombia is an example of institutionalization of a national impact evaluation system . The report is available at: http://lnweb90.worldbank.org/oed/oeddoclib.nsf/DocUNIDViewForJavaSearch/E629534B7C677EA78525754700715CB8/$file/inst_ie_framework_me.pdf

- For impact evaluations, *there is often a trade-off between speed and quality*. The impact of higher quality evaluations may not be seen in the actual intervention being evaluated, but the benefits may extend to other projects and evaluations by pushing the frontier of what can be evaluated and how and by setting new expectations for evaluation quality.
- Timing may create another concern. If there are likely to be staff changes before an evaluation is completed, staff may have little incentive to start an evaluation that they will not see completed (or get credit for). Alternatively, one person interested in impact evaluation may be replaced by someone with different priorities.
- Starting the evaluation data collection late in the project cycle may reduce data quality and the insights that could be gained. However, with careful management and particular attention to communication and dissemination, even a less-than-ideal evaluation can prove to be very useful, especially if it is timely and addresses urgent questions.
- Also, there may be special challenges that can suppress demand in sectors and regions that are less commonly evaluated. Where there is little habit or "culture of evaluation", there may be less funding and less pressure to evaluate, and perhaps higher resistance to accountability. It may require special efforts to begin to build a culture of evaluation.

Clear and well communicated messages

- Clarity and comprehensibility increase use. It helps when the evaluation results point to clear policy implications. This may also apply to the comprehension of methods. While stakeholders may be willing to "trust the experts" if an evaluation offers results that support what they want to hear, there may be a reasonable tendency to distrust results – and particularly methods – that they don't understand.
- People tend to trust or distrust evidence based on what they already believe, looking for results that confirm what they believe and looking for ways to discredit contrary information. Perhaps one reason is that it is difficult to distinguish between good and bad evidence. Currently, there is much ongoing work to provide training in measurement and evaluation for donors and policymakers: when individuals have a greater understanding of impact evaluation, they may be better able to recognize differing qualities of evidence, allowing individual evaluations to have greater impact.

Effective dissemination

- Rapid, broad and well targeted dissemination strategies are important determinants of utilization. One reason that many sound and potentially useful evaluations are never used is that very few people have ever seen them.
- Providing rapid feedback to government on issues such as the extent of corruption or other "hot" topics enhances utilization.
- Continuous and targeted communication builds interest and confidence and also ensures "no surprises" when the final report and recommendations are submitted. This also allows controversial or sensitive findings to be gradually introduced. Trust and open lines of communication are important confidence builders.

- An individual evaluation will rarely be entirely conclusive, but conclusions drawn from an accumulation of evidence may be more difficult to refute.
- The choice of the institution and the evaluators can contribute to dissemination and credibility of the findings.
- Making data available to the academic community is also an important way of broadening interest and support for evaluations and also of legitimizing the methodologies (assuming they stand up to academic critiques as have PROGRESA and Familias en Accion).

Positive and non-threatening findings

- Positive evaluations, or those that support the views of key stakeholders, increase the likelihood they will be used. While this is not surprising, one of the reasons is that many agencies are either fearful of the negative consequences of evaluation or considered evaluation as a waste of time (particularly the time of busy managers) or money. Once stakeholders have appreciated that evaluations were not threatening and were actually producing useful findings, agencies have become more willing to request and use evaluations and gradually to accept negative findings – or even to solicit evaluations to look at areas where programs were not going well.
- There is always demand for results that confirm what people want to hear. Concerns over potential negative results, bad publicity, or improper handling of the results may reduce demand; sensitivity, trust-building, and creative arrangements may help overcome these fears. Consequently, there may be some benefit in taking advantage of opportunities to present good results, especially if it helps the process of getting stakeholders to understand and appreciate the role of impact evaluation.

Active engagement with national counterparts

- The active involvement of national agencies in identifying the need for an evaluation, commissioning it, and deciding which international consultants to use is central to utilization. It gives ownership of the evaluation to stakeholders and helps ensure the evaluation focuses on important issues. It often increases quality by taking advantage of local knowledge and in several cases reduces costs (an important factor in gaining support) by combining with other ongoing studies.
- This cooperation can enable evaluators to modify the initial evaluation design to reflect concerns of clients – for example, changing a politically sensitive randomized design to a strong quasi-experimental design.
- Involving a wide range of stakeholders is also an important determinant of utilization. This can be achieved through consultative planning mechanisms, dissemination and ensuring that local as well as national level agencies are consulted.
- In some contexts, the involvement of the national statistical agency increases the government's trust, and the results and the process have been better accepted when overseen and presented by the statistics agency.

Demonstrating the value of evaluation as a political and policymaking tool and adapting the design to the national and local political contexts

- When evaluation is seen as a useful political tool, this greatly enhances utilization. For example, managers or policymakers often welcome specific evidence to respond

to critics, support for continued funding or program expansion. Evaluation can also be seen as a way to provide more objective criticism of an unpopular program.

- Once the potential uses of planning tools such as cost-effectiveness analysis are understood, this increases the demand for, and use of, evaluations. Evaluations can also demonstrate the practical value of good monitoring data, and increased attention to monitoring in turn generates demand for further evaluations. When evaluations show planners better ways to achieve development objectives, such as ensuring services reach the poor, this increases utilization and influence.
- Increasing concerns about corruption or poor service delivery have also been an important factor in government decisions to commission evaluations. In some cases, a new administration wishes to demonstrate its transparency and accountability or to use the evaluation to point out weaknesses in how previous administrations had managed projects.
- Evaluations that focus on local contextual issues (i.e. that are directly relevant to the work of districts and local agencies) are much more likely to be used.
- In cases where a clearly defined selection cut-off point can be defined and implemented (e.g. the score on a poverty or probability of school drop-out scale), the regression discontinuity design (RD) can provide a methodologically strong design while avoiding political and ethical concerns about RCTs.
- Evaluators must adapt evaluation designs to political realities when deciding what evaluation strategies will be both technically sound and politically feasible. Evaluations of large, politically sensitive programs should be designed at an early stage before the programs have developed a large constituency and become resistant to questioning of their goals and methods. Evaluations should begin early in the program with greater use being made of small pilot projects to assess operational procedures and viability for expansion.

The methodological quality of the evaluation and credibility of the international evaluators

- High quality of an evaluation is likely to increase its usefulness and influence. Quality improves the robustness of the findings and their policy implications and may assist in dissemination (especially in terms of publication). However, an impact evaluation of a compromised quality may still be useful if it can provide timely and relevant insight or if it ventures into new territory: new techniques, less-evaluated subject matter, or in a context where relevant stakeholders have less experience with impact evaluations.
- The credibility of international evaluators, particularly when they are seen as not tied to funding agencies, can help legitimize high profile evaluations and enhance their utilization.
- In some cases, the use of what is considered "state of the art" evaluation methods such as randomized control trials can raise the profile of evaluation (and the agencies that use it) and increase utilization.
- New and innovative evaluations often attract more interest and support than the repetition of routine evaluations.
- On the other hand, while studies on the "frontier" may be more novel or attract more attention, subsequent related studies may be useful in confirming controversial

findings and building a body of knowledge that is more accepted than a single study, especially a single study with unpopular findings.

- Evaluation methods, in addition to being methodologically sound, must also be understood and accepted by clients. Different stakeholders may have different methodological preferences.

Evaluation capacity development

- Evaluation capacity, especially at a local level, is an important factor in the quality of an impact evaluation that also affects the ability of stakeholders to demand, understand, trust, and utilize the results.
- Capacity building is an iterative process and may improve both demand and quality.

D. Strategic considerations in promoting the utilization of impact evaluations

Many of the evaluations cited in this report were selected opportunistically, depending on the availability of donor funding and technical support and the interest of a particular agency, or even a small group of champions within the agency. While individual evaluations may have made a useful contribution, the cases illustrate that the effects and benefits are often cumulative, and utilization and government buy-in tend to increase where there is a sequence of evaluations. In several cases, the first evaluation was methodologically weak (for example, being commissioned late in the project and relying on retrospective data collection methods for reconstructing the baseline), but when the findings were found useful by the national counterparts, this generated demand for subsequent and more rigorous evaluations.

Effective utilization of impact evaluations is an incremental process, with the full benefits only being realized once a number of useful evaluations have been conducted. Policymakers, planners, managers and funding agencies gradually gain confidence in the value of impact evaluation once they have seen some of the practical benefits, and have learned that some of the initial concerns and reservations were not fully justified. A key element in the successful utilization is developing a system for the selection of evaluations that address key policy issues and for analysis, dissemination, and utilization of the results. All of these considerations require the *institutionalization of an impact evaluation system* with strong buy-in from key stakeholders and with a powerful central government champion, usually the ministries of finance or planning.

Institutionalization of Impact Evaluation within the Framework of a Monitoring and Evaluation System (IEG 2008) identifies a number of different paths towards the institutionalization of impact evaluation and points out that the utility and influence of many methodologically sound evaluations has been limited because they were looked upon as one-off evaluations and did not form part of a systematic strategy for selecting evaluations that addressed priority policy issues or that were linked into national budget and strategic planning. This report argues that methodologically sound and potentially useful impact evaluations do not automatically ensure the development of an evaluation

system, and that the creation of such a system requires a strong commitment on the part of government agencies and donors over a long period of time.

The present publication corroborates many of the findings of the IEG study. In addition to the recommendations and guidelines presented in the previous sections, the discussion of the evaluation presentations[11] raised the following issues:

- It is important identify and support impact evaluations that can provide findings and knowledge that will be useful to a broader audience than the project agency whose programs are being evaluated.
- The role of the evaluator should be clarified. Should they become advocates for the adoption of the evaluation findings (for example, the free distribution of anti-malaria or deworming treatments) or should their role be limited to the collection and analysis of data that the evaluation clients will interpret? While many clients require the evaluator to present recommendations, there is a concern in the evaluation profession that the requirement to present recommendations may lead to a bias in how the findings are presented (and particularly ignoring findings that do not support the recommendations).
- There is also a challenge when academics are asked to provide recommendations. The academic researcher is trained to present caveats rather than to come to firm conclusions. Also, the academic has a different set of incentives, and she or he is often judged on the number of publications (in journals that require the use of particular methodologies and give less value to policy recommendations based on the best available, but less rigorous, evidence).
- The previous point relates to a concern that the influential role of academic researchers in the program evaluation field means that many evaluations are method-driven rather than policy driven. This criticism has often been leveled at advocates of randomized control trials who are seen as ignoring important policy evaluations where it is not possible to use rigorous methods, in favor of evaluations that are less useful to policymakers and planners but where it is possible to use randomized designs.
- Further to this point was the recommendation that there is a need to consider rules and procedures for defining acceptable standards of evidence. Different fields, such as health and drug research, may traditionally use different standards of evidence and proof than those used in other fields such as conditional cash transfers and poverty analysis. Is it possible to define generally accepted standards of evidence that can apply in all sectors?
- The question of standards of evidence also applies to increasing use of mixed method evaluation designs that recognize and seek to reconcile the different criteria of evidence and proof conventionally used in quantitative and qualitative research.
- A final point concerned the question of whether all evaluation results should be disseminated. For example, if the success of an evaluation depends on close

[11] These considerations draw primarily from Michael Kremer's reflections during his presentation on the Kenyan deworming evaluation.

cooperation of national counterpart agencies, should there be situations in which these agencies can decide whether and when certain findings should be disseminated? There are other situations in which potentially important but controversial findings may be based on weak evidence (for example with small sample sizes and low statistical power). While researchers may understand that such findings must be interpreted with caution, the mass media or political supporters or critics of a program may ignore these caveats, perhaps jumping to conclusions that a program should be terminated or an innovative approach should receive major funding.

Annex 1 The case studies

All of the case studies are available on video presentations, and (except where indicated) the presentations are also available in Power Point on the conference website: www.worldbank.org/iepolicyconference.

Education

Deon Filmer. *Promoting Lower Secondary School Attendance: The Impact of the CESSP Scholarship Program in Cambodia.*

Miguel Urquiola. *The Effects of Generalized School Choice on Achievement and Stratification: Evidence from Chile's Voucher Program.*

Antonie de Kemp and Joseph Eilor. *Impact of Primary Education in Uganda* [Video presentation only].

Anti-Poverty Programs and Conditional Cash Transfers

Emmanuel Skoufias. *The Role of Impact Evaluation in the PROGRESA/Oportunidades Program of Mexico.*

Orazio Attanasio. *Evaluating a Conditional Cash Transfer: The Experience of Familias en Accion in Colombia.*

Emanuela Galasso. *Assessing Social Protection to the Poor: Evidence from Argentina*.

Health

Adam Wagstaff. *An Impact Evaluation of a Health Insurance Scheme in China.*

Pascaline Dupas. *Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment (Bednets – Kenya).*

Michael Kremer. *Evaluating a Primary School Deworming Program in Kenya* [Video presentation only].

Sustainable Development

Dominique Van De Walle. *Making smart Policy: Using Impact Evaluations of Rural Roads (Vietnam).*

Jocelyne Delarue. *The Impact Evaluation of MicroFinance Projects and their Expected Use (Madagascar and Morocco).*

John Hoddinott. *Ethiopia's Food Security Program* [Video Presentation only].

Reporting Back from the Sector Sessions and Lessons Learned

Norbert Schady. *Impact Evaluation of Anti-Poverty Programs and Conditional Cash Transfers.*
http://siteresources.worldbank.org/INTISPMA/Resources/Training-Events-and-Materials/449365-1199828589096/NorbertSchady.pdf

Halsey Rogers. *The Impact of Impact Evaluations: Lessons from the Education Sector.*
http://siteresources.worldbank.org/INTISPMA/Resources/Training-Events-and-Materials/449365-1199828589096/HalseyRogers.pdf

Other Titles in the Doing Impact Evaluation Series

1. Impact Evaluation and the Project Cycle
2. Conducting Quality Impact Evaluation Under Budget, Time and Data Constraints
3. Impact Evaluation for Slum Upgrading Interventions
4. A Guide to Water and Sanitation Sector Impact Evaluation
5. Conducting Impact Evaluation in Urban Transport
6. Data for Impact Evaluation
7. Impact Evaluation for Microfinance
8. Impact Evaluation for Land Property Rights Reform
9. Methodologies to Evaluate Early Childhood Development Programs
10. Impact Evaluation for School-Based Management Reform
11. Evaluation in the Practice of Development
12. Impact Evaluation of Rural Road Projects
13. Methodologies to Evaluate the Impact of Large-Scale Nutrition Projects
14. Making Smart Policy: Using Impact Evaluation for Policy Making

THE WORLD BANK

Poverty Reduction and Economic Management PREM

Thematic Group on Poverty Analysis, Monitoring and Impact Evaluation