WHEN WILL WE EVER LEARN?

IMPROVING LIVES THROUGH IMPACT EVALUATION

REPORT OF THE EVALUATION GAP WORKING GROUP

MAY 2006

Evaluation Gap Working Group

Co-chairs

William D. Savedoff

Ruth Levine

Nancy Birdsall

Members

François Bourguignon*

Esther Duflo

Paul Gertler*

Judith Gueron

Indrani Gupta

Jean Habicht

Dean Jamison

Daniel Kress

Patience Kuruneri

David I. Levine

Richard Manning

Stephen Quick

Blair Sachs

Raj Shah

Smita Singh

Miguel Szekely

Cesar Victora

Project coordinator

Jessica Gottlieb

Note: Members of the Working Group participated in a personal capacity and on a voluntary basis. The report of the Working Group reflects a consensus among the members listed above. This report does not necessarily represent the views of the organizations with which the Working Group members are affiliated, the Center for Global Development's funders, or its Board of Directors.

*Members who submitted reservations, which can be found at the end of the report (p. 44).

WHEN WILL WE EVER LEARN?

IMPROVING LIVES THROUGH IMPACT EVALUATION

Report of the Evaluation Gap Working Group

William D. Savedoff
Ruth Levine
Nancy Birdsall
Co-Chairs

Center for Global Development
Washington, D.C.

"In my eyes, Americans as well as other tax payers are quite ready to show more generosity. But one must convince them that their generosity will bear fruit, that there will be results."

—Paul Wolfowitz, President, World Bank

"Aid evaluation plays an essential role in the efforts to enhance the quality of development co-operation."

—Development Assistance Committee,
Organisation for Economic Co-operation and Development,
Principles for Evaluation of Development Assistance

"As long as we are willing to continue investing in experimentation, research, and evaluation and build on this knowledge base, we will be able to meet the development challenge."

—Nicholas Stern, Second Permanent Secretary,
HM Treasury, United Kingdom

"The Development Committee recognizes the need to increase its focus on performance by ensuring that development results are reviewed through clear and measurable indicators."

—Trevor Manuel, Minister of Finance, South Africa

"Success depends on knowing what works."

—Bill Gates, Co-Chair, Bill & Melinda Gates Foundation

"We urge the [multilateral development banks] to continue to increase their collaboration and the effectiveness of their assistance, including through increased priority on improving governance in recipient countries, an enhanced focus on measurable results, and greater transparency in program decisions."

—G-7 Finance Ministers

"In particular, there is a need for the multilateral and bilateral financial and development institutions to intensify efforts to . . . [i]mprove [official development assistance] targeting to the poor, coordination of aid, and measurement of results."

—Monterrey Consensus

"If development projects are transparent, productive, and efficiently run, I believe that they will enjoy broad support. If they are not, they are likely to fare poorly when placed in competition with domestic priorities or more tangible security-related expenditures."

—Richard G. Lugar, United States Senator
and Chairman of the Senate Foreign Relations Committee

ISBN 1-933286-11-3

Center for Global Development
1776 Massachusetts Avenue, NW
Third Floor
Washington, D.C. 20036
Tel: 202 416 0700
Web: www.cgdev.org

Contents

Acknowledgments

We are grateful for the enthusiasm, support, and interest of the many individuals who contributed to this report. The time and energy given by so many to debate about the nature of the "evaluation gap" and its solutions is testimony to the importance of this issue. It is a reflection of a growing recognition that the international community is doing far less than it could to generate knowledge that would accelerate progress toward the healthy, just, and prosperous world we all seek.

Among those who deserve special mention are members of the Evaluation Gap Working Group, who volunteered, in their individual capacities, to work together over a period of two years. Working Group members actively participated in extensive discussions, gathered data, made presentations, and grappled constructively, patiently, and deliberately. Throughout, they were attentive to both the big picture and the small details, and we appreciate their commitment.

We would also like to express our thanks to some 100 people who participated in consultations in Washington, D.C.; London; Paris; Menlo Park, California; Mexico City; Cape Town; and Delhi, and to another 100 people who agreed to be interviewed, submitted written comments, or participated in an online survey. All these thoughtful contributions helped us understand the problems of learning about what works, provided a wealth of ideas about possible solutions, and pushed us to refine our understanding of the issues and the practicalities. Together, the many responses to an early draft of this report helped us to see that the world is not as simple as we might originally have been tempted to portray it *and* that the solution to the evaluation challenges is far from impossible. We are particularly grateful to Jan Cedergren, who convened a small group of senior officials from development agencies to examine the advantages and disadvantages of a focused effort to improve the quantity and quality of impact evaluations. Special thanks go to Jessica Gottlieb, who provided essential support and continuity to this initiative through her many intellectual and organizational contributions. The report was edited, designed, and laid out by Communications Development Incorporated.

We are grateful for financial support and intellectual contributions from the Bill & Melinda Gates Foundation and The William and Flora Hewlett Foundation, with particular thanks to Dan Kress, Blair Sachs, Raj Shah, and Smita Singh for their insights and encouragement.

When Will We Ever Learn?

Improving Lives through Impact Evaluation

Executive summary

Successful programs to improve health, literacy and learning, and household economic conditions are an essential part of global progress. Yet after decades in which development agencies have disbursed billions of dollars for social programs, and developing country governments and nongovernmental organizations (NGOs) have spent hundreds of billions more, it is deeply disappointing to recognize that we know relatively little about the net impact of most of these social programs. Addressing this gap, and systematically building evidence about what works in social development, would make it possible to improve the effectiveness of domestic spending and development assistance by bringing vital knowledge into the service of policymaking and program design.

When we seize opportunities to learn, the benefits can be large and global

In 2004 the Center for Global Development, with support from the Bill & Melinda Gates Foundation and The William and Flora Hewlett Foundation, convened the Evaluation Gap Working Group. The group was asked to investigate why rigorous impact evaluations of social development programs, whether financed directly by developing country governments or supported by international aid, are relatively rare. The Working Group was charged with developing proposals to stimulate more and better impact evaluations (see appendix A).

The initiative focused on social sector program evaluations because of their high profile in international fora (the Millennium Development Goals are an example). The Working Group deliberated during 18 months and consulted with more than 100 policymakers, project managers, agency staff, and evaluation experts through interviews and meetings (see appendix C). This report is the culmination of that process.

We need to know more about social development

When we seize opportunities to learn, the benefits can be large and global. Rigorous studies of conditional cash transfer programs, job training, and nutrition interventions in a few countries have guided policymakers to adopt more effective approaches, encouraged the introduction of such programs to other places, and protected large-scale programs from unjustified cuts. By contrast, a dearth of rigorous studies on teacher training, student retention, health financing approaches, methods for effectively conveying public health messages, microfinance programs, and many other important programs leave decisionmakers with good intentions and ideas, but little real evidence of how to effectively spend resources to reach worthy goals.

Many governments and organizations are taking initiatives to improve the evidence base in social development policy, but investment is still insufficient relative to the demand, and the quality of evaluation studies is

mixed. Governments and agencies regularly seek ideas and guidance to develop new programs or to improve existing ones, but on time frames and budgets that do not allow rigorous evidence to be developed. These institutions may do well in their normal data collection and evaluation tasks related to monitoring inputs, improving operations, and assessing performance, but largely fail in building knowledge, which requires studies that fall outside normal budget and planning cycles and for which incentives are sorely lacking.

There are many types of knowledge and correspondingly many useful methods. One type of knowledge concerns the net impact of a program or intervention on conditions that the program sought to alter—children's health status, income-generation by households, learning outcomes, for example. Acquiring this knowledge typically demands studies that are different from program monitoring or process evaluations. It requires impact studies.

Governments and agencies regularly seek ideas and guidance to develop new programs or to improve existing ones, but on time frames and budgets that do not allow rigorous evidence to be developed

The knowledge gained from rigorous impact studies is in part a public good—the cost of producing such studies are borne by individual institutions or agencies, yet once the results of such studies are available they can be accessed by anyone to improve policy. The value of an individual institution's activities or studies would be multiplied if complemented by collective efforts to identify policy questions of shared importance, to cluster studies around priority issues, to ensure that studies are reliable and valid, to register and disseminate studies, and to build research capacity in developing countries.

We are facing a costly evaluation gap

The missing puzzle piece in learning about what kinds of social interventions can succeed is impact evaluations, studies that document whether particular programs are actually responsible for improvements in social outcomes relative to what would have happened without them. An "evaluation gap" has emerged because governments, official donors, and other funders do not demand or produce enough impact evaluations and because those that are conducted are often methodologically flawed.

Too few impact evaluations are being carried out. Documentation shows that UN agencies, multilateral development banks, and developing country governments spend substantial sums on evaluations that are useful for monitoring and operational assessments, but do not put sufficient resources into the kinds of studies needed to judge which interventions work under given conditions, what difference they make, and at what cost.

Even when impact evaluations are commissioned, they frequently fail to yield useful information because they do not use rigorous methods or data. A systematic review of the United Nations Children's Fund (UNICEF) estimated that 15 percent of all its reports included impact assessments, but noted that "[m]any evaluations were unable to properly assess impact because of methodological shortcomings" (Victora 1995). Similarly, a review of 127 studies of 258 community health financing programs found that only two studies were able to derive robust conclusions about the impact on access to health services (ILO 2002).

Poor quality evaluations are misleading. No responsible physician would consider prescribing medications without properly evaluating their impact or potential side effects. Yet in social development programs, where large sums of money are spent to modify population behaviors, change economic livelihoods, and potentially alter cultures or family structure, no such standard has been adopted. While it is widely recognized that withholding programs that are known to be beneficial would be unethical, the implicit corollary—that programs of unknown impact should not be widely replicated without proper evaluation—is frequently dismissed.

Why is there an evaluation gap?

An evaluation gap exists because there are too few incentives to conduct good impact evaluations—and too many obstacles

An evaluation gap exists because there are too few incentives to conduct good impact evaluations—and too many obstacles. These obstacles include technical, bureaucratic, and political challenges. While impact evaluations generally have to be designed as an integral part of a new program, politicians and project managers are focused in the early program phases on design and implementation. At this early stage starting an impact evaluation has immediate costs while the benefits of measuring the impact are felt only well into the future. Paradoxically, the same people who would like to have good evidence today about the impact of earlier social programs are unlikely to make the efforts necessary to design and implement the kind of impact evaluation study that would benefit those who follow.

Rising impatience with ignorance

Tolerance for the evaluation gap is waning. Developing country governments are demanding better information about the efficacy of social spending. In 2001, for example, Mexico passed legislation requiring that impact evaluations be conducted on a variety of social programs, explicitly recognizing the value of learning what works and why as a guide for future budget decisions. NGOs have collaborated with leading academic institutions to evaluate the impact of their programs, with the goal of identifying what works and what does not. This is seen as vital for both better programs and more effective fundraising. Donor countries are increasingly concerned that international financial assistance should generate results.

A growing number of examples show that good quality impact evaluations can be feasible, ethical, timely, and useful. The capacity to conduct impact evaluations at research institutions around the world is greater than ever before, using a range of proven methods to measure impacts that can be attributed to a particular program or policy. The technology and networks for sharing information have increased dramatically. Impact evaluations have played critical roles in helping NGOs modify education programs in India to improve student achievement, protected and expanded national conditional cash transfer programs in several Latin American countries, and demonstrated the impact of inexpensive health interventions in improving school attendance in Africa.

Building this knowledge requires that governments and agencies take a strategic view and conduct impact evaluations in projects that can yield important information about what is effective and what is not. It also requires

that evaluations use methods that measure the net impact of programmed activities in a valid way. Better coordination of impact evaluations across countries and institutions would make it possible to cluster some studies around common thematic areas and improve the ability to generalize findings.

Moving forward

Concern about the evaluation gap is widespread, as demonstrated by the many ways that public agencies, intergovernmental commissions, NGO networks, research centers, and foundations are addressing it. Many initiatives are under way to:

- Increase access to existing information through reviews, searchable databases, and policy pamphlets and newsletters.
- Improve regular data collection by developing country governments and develop aggregate indicators.
- Promote specific evaluations with grants and other kinds of funding.
- Conduct research and demonstrate good evaluation practices.

But progress will be slow and investment insufficient without greater effort.

Collective commitments can help to ensure that sufficient investments are made in improving the production and use of knowledge about social program impacts

After deliberation, analysis, and consultation, the Evaluation Gap Working Group recommends that the full range of stakeholders—NGOs, foundations, research centers, bilateral agencies, developing country governments, and multilateral development banks—should both reinforce existing initiatives and collaborate on a new set of actions to promote more and better impact evaluations.

Recommendations for individual action: reinforce existing efforts

At minimum, governments and agencies should **reinforce efforts to generate and apply knowledge from impact evaluations of social programs**. This includes strengthening overall monitoring and evaluation systems; dedicating resources to impact evaluation; ensuring collaboration between policymakers, project managers, and evaluation experts; improving standards for evidence; facilitating access to knowledge; and building capacity in developing countries to conduct rigorous evaluations (see table 1 on p. 33 for a summary of these recommendations).

Though countries and agencies will gain from undertaking these activities independently, incentives are more favorable when other organizations engage in complementary ways. For example, an organization benefits by strengthening its internal evaluation systems, but its ability to interpret and use its evaluation information is vastly enhanced when it benchmarks its performance against that of other organizations and learns from their experiences. Dedicating funds to impact evaluation will benefit an organization's decisions about social programs—all the more so if impact evaluations addressing similar questions are being financed by other groups. Similarly, an organization's credibility and reputation are enhanced when there is transparency in disseminating findings, whether favorable or not. Sustaining transparency is easier when other organizations assume a similar posture. Thus collective commitments can help to ensure that sufficient investments are made in improving the production and use of knowledge about social program impacts.

Recommendations for collective action: commitments to public goods through a new council

Independent actions by individual countries and agencies can reduce the evaluation gap, but progress is likely to be much faster if some countries and agencies **collectively commit to increase the number of impact evaluations and adhere to high standards of quality**. In one form of commitment, similar to a contract, each organization would agree to do its part, to shoulder its fair share of the required tasks. In another form of commitment organizations would support a common infrastructure to carry out functions that are most effectively accomplished jointly. In both cases organizations assume responsibilities and reap benefits by collaborating.

The benefits of collective action are clear, and there are many strategies for implementation

The benefits of collective action are clear, and there are many strategies for implementation. Through wide-ranging consultations, the Working Group identified the following characteristics of a successful new initiative:

- Complementary to existing initiatives.
- Strategic in its choice of topics and studies.
- Opportunistic in its approach to supporting good impact studies.
- Linked directly and regularly engaged with policymakers, governments, and agencies.
- Involving collective, voluntary commitment by a set of governments and public and private agencies to conduct their own studies or contribute funds for contracting such studies by others.
- Committed to independence, credibility, and high standards for evidence.

An initiative that meets these criteria must have clearly identified functions that will both redress the evaluation gap and be more efficient if conducted collaboratively. It also requires an appropriate funding mechanism and an institutional design that is feasible, efficient, and accountable.

The Evaluation Gap Working Group developed a consensus that some entity—whether a committee, standards-based network, secretariat, or other organization—is needed as a focal point for leading such an initiative. For the following discussion, this entity is referred to as a "council."[1] The council's membership would include any set of developing country governments, development agencies, NGOs, foundations, and other public and private entities that volunteer to generate new, policy-relevant knowledge about social programs. The Working Group identified a set of core functions and elaborated ideas on funding and institutional design. They are offered here as a way to facilitate action toward a real-world solution. (See table 2 on pp. 36–37 for a summary of these recommendations.)

Core functions. The Working Group identified functions that would contribute to reducing the evaluation gap, are best carried out collaboratively, and would benefit from the focused attention provided by an entity like a council. Of these functions, the following were judged to be core functions:

- *Establishing quality standards for rigorous evaluations*. It is costly and confusing for each government or agency to create its own standards for rigor in impact evaluation. A council could periodically convene experts to set a common standard or endorse existing standards (see appendix

G for examples). Making standards explicit would facilitate the design of new impact evaluations, serve as a reference in reviewing proposals, and help to build local capacity for conducting and interpreting studies.

- *Administering a review process for evaluation designs and studies*. Reviewing proposals and studies requires time, money, and knowledge. While larger organizations and agencies may have the capacity, smaller ones do not. A council could administer reviews with a rotating panel of experts from different fields on behalf of member organizations—benefiting from economies of scale and scope. By reviewing impact evaluation designs and assessing completed evaluations according to clear and transparent methodological standards, the council can also help members distinguish between stronger and weaker forms of evidence.
- *Identifying priority topics*. No government or agency can initiate studies on every policy question that they would like answered. Nor is it necessary to evaluate every program. A collective effort to identify the most pressing and enduring policy questions would help governments and agencies to cluster evaluations around common topics and to focus efforts on programs that are most likely to yield useful information for future policymaking. By participating in such a collective effort, governments and agencies can influence the questions being asked and benefit from studies done by other institutions on programs like their own.
- *Providing grants for impact evaluation design*. The window of opportunity to design a good impact evaluation on an important question is narrow, occurring just at the moment of program conception and design. Often, the missing ingredient is timely funding to contract an expert to meet with stakeholders to assess whether an impact evaluation would be appropriate and what methods would generate the best evidence and then to design the evaluation. With relatively small amounts of money, the council could act as a powerful catalyst—in some cases making it possible to do impact evaluations that might not otherwise get done and in other cases increasing the likelihood that the money spent on evaluation generates reliable and valid conclusions.

With relatively small amounts of money, the council could act as a powerful catalyst, making it possible to do impact evaluations that might not otherwise get done

Other functions. Other functions identified in this review are either less critical to the council's mission or require substantially more resources. These functions might be delegated to the council in the future, depending on the council's performance, its staffing, and members' interest and financial support. These other functions, explained in the report, are:

- *Organizing and disseminating information*, such as a prospective registry of impact evaluations, databases of completed qualified studies, and systematic reviews.
- *Building capacity to produce, interpret, and use knowledge* by encouraging links between researchers, agency staff, and project managers; recognizing impact evaluation designs that incorporate capacity building; encouraging new people to enter the field of evaluation with fellowships; and disseminating training materials and rigorous evidence.

- *Creating a directory of researchers* for use by members and actively encouraging the use of qualified experts.
- *Undertaking communication activities and public education programs* to explain the benefits and uses of impact evaluation, advocate for appropriate legislation and policies, and generate public support for building knowledge.
- *Administering funds on behalf of members*. Some members might choose to use the council's services to commission and manage impact evaluations on their behalf. Members could also hire the council for specific services, such as to convene an external review panel to assess an impact evaluation design.

Members of the council would commit to finance or contribute to financing impact evaluations that address questions of common interest and enduring importance

Administering a pooled impact evaluation fund. This final function is discussed separately because the Working Group could not reach a consensus on its inclusion as a core function. It would involve delegating responsibility to the council to administer a pooled fund dedicated to conducting rigorous impact evaluations of social development programs in developing countries. The council's role would be clearly defined and directed by the members, through a governing body, to commission independent impact evaluations on topics agreed by the members to be of high priority.

Some Working Group members were concerned that such a fund would divert financial resources from current impact evaluation efforts. They also argued that the initiative should proceed gradually, beginning with more modest and immediately feasible functions.

Others argued that giving the council adequate funds to commission impact evaluations was essential to address the central concerns set out in the analysis: that the knowledge from impact evaluations is partially a public good that will inevitably be underfunded without a collective effort and that quality and credibility are enhanced when impact evaluations are commissioned externally and independently.

Funding options

The best solution to a public good problem such as the generation of sufficient impact studies is often to obtain a collective agreement from all parties to commit some level of funding to the common effort. The funds can then continue to be applied independently by each party to the agreement. Alternatively, a portion of the funds can be pooled for management by a particular entity. Any discussion of funding needs to distinguish between financing impact evaluations and financing a council's core functions and services. These are some of the first questions that prospective members will have to negotiate.

Funding studies. The essential problems posed by public goods are insufficient investment and free-riding. To avoid these problems, members of the council would commit to finance or contribute to financing impact evaluations that address questions of common interest and enduring importance.

For organizations that fulfill their commitment by commissioning their own studies, the council's role would be to receive information on which studies are being started and implemented and their associated spending.

The council would then verify whether the studies meet the collectively agreed-on standards for rigor. For organizations that fulfill their commitment by commissioning others to conduct studies, the council would similarly verify that research designs and studies are rigorous, registering expenditure and reporting. Developing countries should be equal partners, committing to conduct or finance impact evaluations. Impact evaluations conducted within a particular country with grants or loans from abroad would still count toward that country's membership commitment.

Developing countries should be equal partners, committing to conduct or finance impact evaluations

Funding core functions and services. The council's core functions, again, have aspects of public goods and require agreement to ensure sufficient funds and reduce free-riding. Members would share in the financing of these core functions through contributions reflecting differences in scale and financial resources.

Institutional options

A further set of questions concern how to constitute the council to best provide collectively beneficial services. The council can be constituted in many different forms, from an interagency committee, to a network, secretariat, or independent organization.[2] The choice will depend on assessing the relevant tradeoffs, and the institutional design should ultimately be guided by the structure that will best fulfill a range of aims, including the following:

- *High standards of technical quality*, to produce studies that generate strong evidence.
- *Independence and legitimacy*, to help the council develop a reputation for independence and integrity.
- *Operational efficiency*, to exploit economies of scale and scope and avoid duplication of functions.
- *International leadership*, to enhance the council's capacity to provide leadership.

Some ideas are presented schematically in table 3 on page 41.

Will we really know more in 10 years?

Imagining 10 years into the future, when the target date for the Millennium Development Goals has come and gone, the international community could be in one of two situations.

We could be as we are today, bemoaning the lack of knowledge about what really works and groping for new ideas and approaches to tackle the critical challenges of strengthening health systems, improving learning outcomes, and combating the scourge of extreme poverty.

Or we could be far better able to productively use resources for development, drawing on an expanded base of evidence about the effectiveness of social development strategies.

Which of those situations comes to pass has much to do with the decisions that leaders in developing country governments and development agencies make over the next couple of years about conducting impact evaluations.

If a group of leading national governments and development agencies recognizes the tremendous potential of more and better impact evaluation and overcomes the natural institutional resistance to engage in an ambitious new effort, we are convinced that a collective approach will loosen many of the constraints that have contributed to the current situation. Shared agenda setting, high methodological standards, and independent evaluation have the potential to vastly expand and deepen our collective knowledge base. To work, this need not be a compulsory effort by all members of the international community—every international agency, every developing country government—but a pioneering effort *is* required by a few at the leading edge who are ready to seize the opportunity.

Shared agenda setting, high methodological standards, and independent evaluation have the potential to vastly expand and deepen our collective knowledge base

Getting to that collective approach will not be simple. Prospective members will have to choose the functions and institutional design that they think will work best, taking into consideration many tradeoffs. The ideas in this report are offered as a point of departure. The single imperative is to reach agreement on an appropriate institutional design as soon as possible to take advantage of current opportunities to learn about what works in social development programs.

I. The lessons rarely learned

> Knowledge is not a luxury. For our bank to intelligently manage risks in our social investment portfolio, it's essential to understand whether, how, and when complex interventions in education, health, microfinance, and other social areas really work. I'm not as interested in what the models tell us people *should* do as I am in finding out what people *actually* do.
>
> —Nachiket Mor, Executive Director, ICICI Bank, India

The international community is united about the urgent need to improve social and economic conditions in developing countries. There is no doubt about the importance of increasing the proportion of children who make it through school and learn enough to compete in the labor market. Of improving the health of babies, children, young people, and parents. Of expanding the opportunities for households to raise themselves out of poverty. There is also no doubt that the factors required in combination to achieve these goals—money, political will, and knowledge about effective public policies—are in short supply.

This report is about the shortage of knowledge and how to remedy it. Knowledge is lacking about what works: what actions national governments, private actors, development agencies, and others can take to lead to beneficial changes in health, education, and other aspects of human welfare.

While we have a large body of information describing the problems of the poor, a growing stock of research findings on the fundamental causes of the many unfavorable outcomes observed in developing countries, and an ongoing flow of data on program inputs and outputs, the base of evidence about the *impact* of both traditional and innovative social policies and programs across varied contexts is limited indeed. Moreover, an enormous number of

opportunities to acquire new knowledge for the future are missed, as new programs are undertaken without attention to measuring their impact.

Persistent shortcomings in our knowledge of the effects of social policies and programs reflect a gap in both the quantity and quality of impact evaluations. Impact evaluations are studies that measure the impact directly attributable to a specific program or policy, as distinct from other potential explanatory factors. Despite the demand by policymakers and program designers for the type of knowledge that impact evaluations provide, few such evaluations are undertaken and the quality of those that are conducted is highly variable.

Persistent shortcomings in our knowledge of the effects of social policies and programs reflect a gap in both the quantity and quality of impact evaluations

In 2004 the Center for Global Development reviewed existing initiatives to address this problem. That review found that many organizations are working on impact evaluation, but none was asking why good social program evaluations are relatively rare in the first place. Consequently, the Center for Global Development, with support from the Bill & Melinda Gates Foundation and The William and Flora Hewlett Foundation, convened the Evaluation Gap Working Group to investigate the factors that lead to underinvestment in rigorous impact evaluations of social development programs, whether financed directly by developing country governments or supported by international aid.

The Working Group was charged with developing proposals to stimulate more and better impact evaluations (see appendix A). The initiative focused on social sector program evaluations because of their high profile in international fora (the Millennium Development Goals, for example). The Working Group deliberated over 18 months and consulted with more than 100 policymakers, project managers, agency staff, and evaluation experts through interviews and meetings (see appendix C). This report is the culmination of that process.

As we show in this report, the gap in impact evaluation is an almost inevitable outcome of the way developing countries and the international community now organize and fund monitoring and evaluation. Closing the gap will require a new approach that takes advantage of opportunities for learning across countries and agencies and provides clear incentives as well as resources to increase the number of evaluations, adhere to quality standards, and disseminate evaluation results widely. With such a new approach, it is possible to imagine that in 5 or 10 years we will know far more than we do today about the most effective ways to use money and political will to achieve vital improvements in people's lives.

The many meanings of evaluation

To start a discussion about the challenges of evaluation, the multiple meanings of the word have to be teased apart. Many kinds of information and evidence are needed by policymakers who decide how much money to use for what programs; by program designers who are charged with making decisions about target beneficiaries, delivery mechanisms, type of services to be provided, financing arrangements, and other key design features;

and by program implementers who face the daily challenges of operating complex health, education, housing, social welfare, and other programs.

The diverse types of knowledge needed for good policymaking, design, and implementation come from many sources: from monitoring program activities, building knowledge about processes and institutions, providing information necessary for external accountability, and measuring impact.[3]

> Part of the difficulty in debating the evaluation function in donor institutions is that a number of different tasks are implicitly simultaneously assigned to evaluation: building knowledge on processes and situations in receiving countries, promoting and monitoring quality, informing judgment on performance, and, increasingly, measuring actual impacts. Agencies still need their own evaluation teams, as important knowledge providers from their own perspective and as contributors to quality management. But these teams provide little insight into our actual impacts and, although crucial, their contribution to knowledge essentially focuses on a better understanding of operational constraints and local institutional and social contexts. All these dimensions of evaluations are complementary. For effectiveness and efficiency reasons, they should be carefully identified and organized separately: some need to be conducted in house, some outside in a cooperative, peer review, or independent manner. In short, evaluation units are supposed to kill all these birds with one stone, while all of them deserve specific approaches and methods. (Jacquet 2006)

Effective monitoring requires data collection during the entire implementation phase

A portion of this knowledge is generated as a routine part of public sector management and the international development business. However, as described in this report, a major gap persists.

Monitoring programs of governments or international agencies are generally built into each operation as an integral part of an organization's management information system. So, for example, a teacher training program would have a means—albeit imperfect—of knowing the funds expended, the number of courses conducted, and the number of teachers trained. Monitoring involves collecting data and analyzing it to verify whether programs were implemented according to plan, whether financial resources and inputs were applied as intended, whether the expected outputs were realized, whether intended beneficiaries were reached, and whether time schedules were met. Effective monitoring requires data collection during the entire implementation phase and, therefore, is generally conducted as an integrated aspect of project execution. The information from monitoring is an essential tool for quality management and control, for detecting irregularities as well as inefficiencies, and for making corrections in real time.

Building knowledge about processes and institutions incorporates information gathered in monitoring programs and goes further to ask how and

why programs get implemented according to plan. It requires a system within agencies and governments to document experiences and share them among staff and across departments. These types of evaluations use a variety of methods to develop a shared vision of an institution's "best practice"and "lessons learned." Because their purpose is to document and transmit experiences, much of this evaluation work must be done in-house to create feedback and lessons to inform future policy. Independent external process and operational evaluations can provide valuable perspective and checks on views generated internally. Many national or subnational governments and international organizations have created evaluation units with differing degrees of independence; some also commission external evaluations.

Impact evaluation asks about the difference between what happened with the program and what would have happened without it

Being accountable to stakeholders is something that national or subnational governments and international agencies do by providing information about their activities and opening their books with sufficient transparency to allow shareholders, taxpayers, and civil society stakeholders to judge their performance. Annual reports and financial statements are one part of this evaluation task, which can extend to placing design documents, notes from meetings, the results of tenders, and project monitoring information into the public domain.

Generating knowledge about whether a program achieved its basic aims requires impact evaluation, which analyzes and documents the extent to which changes in the well-being of the target population can be attributed to a particular program or policy. Such evaluation tries to answer the question: "What difference did this program make?" Impact evaluation asks about the difference between what happened with the program and what would have happened without it (referred to as the counterfactual). For example, "Are children staying in primary school and learning more than they would have without this particular curriculum or teaching innovation?" This difference is the impact of the program.

All four of these categories of information and knowledge are needed for good decisionmaking about social development programs. But not all of them are generated by existing institutional mechanisms. Ensuring that these information categories are generated, transmitted, and used requires attention to the incentives, decisionmaking processes, and resources applied to each kind of study. Ensuring that they are used appropriately requires attention to context and generalizability, in both analysis and interpretation. Because the different types of information and knowledge differ in purpose, methods, and requirements, strategies for their improvement will vary. Furthermore, organizations will find it easier to do some kinds of evaluation than others. In general, governments and development agencies are better at monitoring and process evaluations than at generating information for accountability or measuring impact.

II. The gap in impact evaluation

> From a policy and development planning perspective, the issues being tackled through the Evaluation Gap work on "When Will We Ever Learn?" are becoming increasingly important for making rational decisions in the effective use of limited human and financial resources. Equally important is continued effort to raise the awareness of decisionmakers on evidence-based good practices to maximize the impact of development initiatives. Selectivity is a critical component in the decisionmaking process of prioritizing investments. Such investments must be underpinned by performance benchmarks against which outcomes and impacts can be best measured. Impact evaluations are, therefore, an important tool for measuring the development effectiveness of investments and should be widely applied.
>
> —Philibert Afrika, Director,
> Operations Policies and Review Department,
> African Development Bank Group

Investments in building knowledge can have tremendous and unpredictable returns

The challenge on which the Evaluation Gap Working Group focused was the quantity and quality of impact evaluation. Impact evaluations differ from other forms of research, reporting, and studies in several ways. First, impact evaluations generate knowledge that has wider benefits and may be more applicable to other settings and over time than the information generated by monitoring activities, process evaluations, or performance assessments.

Second, impact evaluations require different kinds of data collection. Most notably, they require attention to gathering information from appropriate comparison groups so that valid inferences can be made about the impact of a particular program compared with what would have happened without it or with a different program. This type of data collection must be considered from the start—the design phase—rather than after the program has been operating for many years, when stakeholders may ask, "So what is the program really accomplishing?"

Third, impact evaluations are not required for all programs and projects. Rather, they are best targeted to programs that are new or expanding and for which effectiveness has not been established.

Building knowledge for learning

The value of impact evaluation is best understood as part of a broad scientific enterprise of learning, in which evidence is built over time and across different contexts, forming the basis for better policymaking and program design. This type of knowledge is, in part, a public good,[4] in the sense that once the knowledge is produced and disseminated, anyone can benefit from it without depleting it or excluding others from its benefits. In this way, investments in building knowledge can have tremendous and unpredictable returns. For example, the discovery in the late nineteenth century that cholera was transmitted through contaminated water has saved untold lives around the world.

Though the benefits from investing in such knowledge are high, the incentives for any particular individual, organization, or country are insufficient to stimulate investment equal to its full social value. Furthermore, there is a temptation to be a free-rider—to let others invest in building knowledge because one can enjoy the fruits produced by others without investing anything oneself.

When a developing country or development agency learns what works (or what does not) among programs that seek to improve health, enhance education outcomes, or reduce poverty, that knowledge can be invaluable to others facing similar problems. This knowledge is not a global public good in the purest sense. After all, the findings of an impact evaluation for a particular program in a particular context will be most relevant and useful in that specific instance. Nevertheless, some of what is learned will be useful to countries with similar contexts. And well done impact evaluations will provide sufficient information about context to help decide whether findings can be generalized to other situations. Equally important, the replication of impact evaluations in similar programs in different places builds an ever-expanding base of evidence and is the most systematic way of improving public policy and spending public money well.

Well done impact evaluations will provide sufficient information about context to help decide whether findings can be generalized to other situations

Applying the right methods

To improve the evidence base, we need more impact evaluations that are designed to provide valid information about impact and that are relevant to important policy questions. The starting point is defining the policy question and the context. From there it becomes possible to choose the best method for collecting and analyzing data and drawing valid inferences.

Many impact evaluations fail to provide rigorous evidence because, even when they measure changes among beneficiaries, they often cannot demonstrate that the changes were due to the program in question. One of the most common ways to estimate the impact of a program is to compare outcomes before and after a program is implemented. Yet many things change at the same time that a project is implemented, so without further information, it is not correct to assume that observed outcomes are due to the project. For example, population health status may improve after a reform of health service delivery in a particular region, but unless other competing explanations—such as changes in income, agricultural productivity, or infectious disease vectors—are ruled out, evidence of impact itself is lacking.

Another common way to measure impact is to compare outcomes in areas that receive a program with those that do not. To obtain valid measurements of the program's impact, this approach requires that the study can account for any systematic differences between areas with and without programs. If programs are introduced in places where they are more likely to be successful, disentangling these systematic differences may be extremely difficult, if not impossible. Studies will be unable to determine whether better outcomes in targeted areas result from the program or from the better institutional conditions and implementation capacities that led these areas to be targeted in the first place. Such comparisons may look like they provide evidence about a program's effectiveness, but they can-

not give decisionmakers a clear answer to the question: "Will this program improve outcomes in less favorable areas?"

In still other cases program beneficiaries are compared with nonbeneficiaries in the same area. Such comparisons can provide valid information about impact only when systematic differences between the two groups can be ruled out or accounted for. Yet in many social programs beneficiaries choose whether to participate. This can occur when people with greater resources, motivation, or abilities seek out assistance from new programs. In such cases it is difficult to know whether improved outcomes are due to these unobserved characteristics of the individuals who choose to participate or to the intervention.[5] Again, such comparisons may look like they provide evidence about a program's effectiveness, but they cannot give decisionmakers a clear answer to the question: "Can this program help people who have not made efforts to participate?"

It is usually necessary to build an impact evaluation into the design of a program from the start so that appropriate comparison groups can be identified

When the selection of beneficiaries is influenced by any of these factors, it is difficult to know (and usually impossible to test) whether statistical controls for observable difference will fully account for the potential bias. In fact, both the extent and the direction of the bias may be unknown, although numerous studies have shown that such bias can be quite prevalent (Glazerman and Levy 2003; Lalonde 1986).

Improperly conducted evaluations are misleading. They present conclusions that are unsubstantiated. This means that the risk of wasting public resources or even harming participants is real. This is why clinical trials of medications have become a standard and integral part of medical care. No physician would consider prescribing strong medications whose impact and potential side-effects have not been properly evaluated. Yet in social development programs, where huge sums can be spent to modify population behaviors, change economic livelihoods, and potentially alter cultures or family structure, no such standard has been adopted. While it is widely recognized that withholding programs that are known to be beneficial would be unethical, the implicit corollary—that programs of unknown impact should not be widely replicated without proper evaluation—is frequently dismissed.

To avoid these problems it is usually necessary to build an impact evaluation into the design of a program from the start so that appropriate comparison groups can be identified. Remembering that the point of departure is always the policy question and context, with the methodological choice following, it is usually worth asking whether a random-assignment approach—that is, randomly choosing which individuals, families, or communities will be offered a program and which will not—is appropriate and feasible. Where this method can be applied, it ensures that impact measurements are not confounded by systematic differences between beneficiary and control groups.

Where random assignment cannot be applied, either because it is not appropriate to the policy question or because it is not feasible, other approaches can be applied, such as controlled before-and-after studies, interrupted time-series studies, and various types of matched comparison studies (such as difference-of-differences and propensity score matching). Using any of these approaches also requires assessing whether they are

feasible, appropriate to the context, relevant to the policy question, and will yield valid inferences about impact.

All impact evaluations require advance planning, careful attention to confounding factors and systematic bias, and adequate time and money (see, for example, Habicht, Victora, and Vaughan 1999; Victora, Habicht, and Bryce 2004; and Altman and others 2001). It is particularly important to highlight the need to build impact evaluations in at the design stage, when the expected impact of the program is being defined and appropriate comparisons can be established. If this is not done, it is easy to neglect impact evaluations or do ones that yield invalid results. The number of well done studies remains limited. Part of the challenge for addressing the evaluation gap is to do better studies. Another part is to improve the methods for addressing these limitations.

Opportunities to learn from social development programs are being systematically neglected

Uncovering gaps in quantity and quality

Despite the need to build an evidence base for designing new programs, the quantity of impact evaluations is small and the quality is highly variable. Substantial resources are usually applied to designing a program, monitoring its implementation, and measuring its outputs. Very little is done to measure a program's impact. Ultimately, this means that many good programs are not expanded as widely or as rapidly as they should be, and funds continue to be wasted on approaches that are failing.

Opportunities to learn from social development programs are being systematically neglected. In developing countries resources for social development programs are often overstretched, and immediate problem solving dominates long-term knowledge building, leading to neglect of information-gathering activities. In countries that do have effective information gathering, most of these resources are directed toward monitoring the use of funds, deploying and managing personnel, and producing outputs and services. By contrast, relatively little is spent to rigorously assess whether programs are having the desired impact beyond what would have occurred without them. Bilateral and multilateral development assistance agencies generally dedicate a share of program funds to monitoring implementation and disbursement, but efforts to extract lessons about program effectiveness from these data sources are regularly disappointed. The kinds of data collected do not lend themselves to measuring the net impact of programs.

The results are clear: a lack of knowledge about the effectiveness of programs in which developing country governments, NGOs, international donors, and philanthropists are spending scarce funds, and a weak base of evidence for future decisionmaking. Consider the history of programs that promote voluntary community health insurance schemes as a way to build sustainable financing for health services. Such programs have been proposed and encouraged for decades (see, for example, WHO 1978). Millions of dollars have been spent in dozens of countries, and reviews of the literature evaluating such programs give the impression that we know a great deal about them and that they are beneficial (see, for example, WHO Commission on Macroeconomics and Health 2001).

However, reviews that explicitly discount studies that are methodologically weak find very little rigorous evidence on whether these strategies are effective. The International Labour Organization's Universitas Programme reviewed 127 studies covering 258 community health schemes and found that only two studies had "internal validity"—that is, were designed in such a way that they could distinguish impacts on the relevant population that were specific to the program from changes attributable to other factors:

> even for utilization of services the information and analysis is scarce and inconclusive mostly due to the few studies that address the question . . . and due to the lack of internal validity for most of those studies that address the question. The main internal validity problems are related to, inter allia, lack of base lines, absence of control groups, problems in sampling techniques, control for confounding variables . . . and sources of data for utilization analysis. (ILO 2002, p. 47)

Of the hundreds of evaluation studies conducted in recent years, only a tiny handful were designed in a manner that makes it possible to identify program impact

Another review found that only 5 of 43 studies that considered the impact of community health insurance on mobilizing funds and improving financial protection for members used statistical controls to support their findings (Ekman 2004).

The problem appears across many sectors:

- The impact of payment mechanisms on healthcare providers was the subject of a Cochrane Review. After searching 17 bibliographic databases, including ISI, Econlit, and MEDLINE, the review found only four studies that could draw valid conclusions (Gosden and others 2004).
- The What Works Working Group at the Center for Global Development reviewed 56 public health interventions that were nominated by leading international experts as examples of major successes. Of these, 12 were excluded because they were too new to be properly evaluated or were small scale. But 27 were excluded because the impact of the public health interventions could not be documented (Levine and others 2004).
- A systematic review of the United Nations Children's Fund (UNICEF) reports found that 44 of 456 were impact evaluations. The review estimated that 15 percent of all UNICEF reports included impact assessments, but noted that "[m]any evaluations were unable to properly assess impact because of methodological shortcomings" (Victora 1995).
- At the Inter-American Development Bank, only 97 of 593 projects active as of July 2004 reported they had collected data on beneficiaries and of these only 18 had data on nonparticipants—information that is necessary for evaluating impact.[6] Similar results can be found at most other regional development banks and bilateral agencies, though recent initiatives are beginning to address this gap.

A further indication that too few impact evaluations are being conducted, despite requests and financing, comes from the shortcomings listed in the evaluation reports themselves. The following selection from such studies is not a systematic survey, but it is recognizable to anyone who has delved into the literature of evaluation reports. (See appendix E for further examples.)

- "[T]his review revealed that, with the exception of Jalan and Glinskaya, none of the studies could qualify as true impact evaluations." [An evaluation of a $1.3 billion primary education program in India with support from the World Bank, the European Commission, UK Department for International Development, UNICEF, and the Dutch government.]
- "[T]here is no proper baseline survey with which the present-day economic situation of the trained farm women and their families can be compared." [A Danida review of four training projects for farm women.]
- "The original plan to collect pre- and post-quantitative data to measure the change in learners' reading, writing, and numeracy skills over the course of the project proved impossible for a variety of reasons." [An assessment of an NGO program to use computer technology in literacy training for adults in Zambia and India.]

Several countries and organizations are working to redress the problem of underinvestment in impact evaluation

Signs of progress

Of course, impact evaluation has not been barren in all fields or at all times, and both methodological developments and an increase in technical capacity have contributed to improvements over time. Good evaluations do happen, and when they are disseminated, they stand out. The RAND health insurance experiment and the Job Training Partnership Act (JTPA) evaluation in the United States remain important points of reference for designing health insurance and job training programs (Newhouse 2004; Gueron 2002; Wilson 1998). The evaluation of Mexico's conditional cash transfer program, Progresa/Oportunidades, has influenced the design of similar programs throughout the world (Morley and Coady 2003). It is striking, in fact, that the few impact evaluations that are widely recognized as adhering to good methodological standards are cited repeatedly and serve as highly visible reference points for our understanding of the impact of social programs.

The Working Group's review found that several countries and organizations are working to redress the problem of underinvestment in impact evaluation in a variety of ways: through advocacy, dissemination of guidelines, training programs, literature reviews, and promotion or conduct of specific evaluations. Mexico passed legislation in 2001 to require impact evaluations of its social development programs and recently created the National Council on Social Policy Research. Chile has used impact evaluations to inform budget decisions and policymaking. Several NGOs working on education in developing countries have undertaken impact evaluations to improve the effectiveness of their work and to assist their advocacy for applying more resources to effective programs.[7] The World Bank's Development Research Group is currently engaged in an initiative to make impact evaluation a more systematic endeavor within the Bank, focused around six thematic areas: school-based management, community participation in education, information for accountability in education, teacher contracting, conditional cash transfer programs to improve education outcomes, and slum upgrading programs. The Inter-American Development Bank

has financed evaluations of several conditional cash transfer programs and training programs. The Agence Français de Développement recently initiated its first rigorous impact evaluation of a microfinance program. India's ICICI Bank is engaging in impact evaluation of microfinance and other programs. The United States Agency for International Development has conducted impact evaluations of insecticide-treated bednets and a variety of cost-effective medical interventions. And this list is by no means exhaustive (see appendix D).

These initiatives are a beginning, concentrated in particular topics and regions. Relatively little is known about many important questions confronting social development programs, such as how best to extend healthcare services to poor families, how to modify health behaviors to discourage the spread of HIV/AIDS, how to increase attendance and completion of schooling by girls, or how microfinance programs affect household incomes and well-being. Studies are also geographically concentrated in certain regions, with more of them in Latin America than in Sub-Saharan Africa or Asia. Perhaps more important, no current initiative appears to tackle two fundamental issues:

Little is known about many important questions, such as how best to extend healthcare to poor families, how to modify behaviors to discourage the spread of HIV/AIDS, or how to increase attendance and completion of schooling by girls

- Understanding whether and how the type of knowledge sought can be generated, within real-world contexts, for the programs that attempt to change social conditions.
- Understanding and addressing the underlying reasons for the persisting evaluation gap—despite the recognition of the value of evidence about impact—so that a solution can be designed that takes these reasons into account.

These are addressed in turn in the sections that follow.

III. Impact evaluations in the real world

> The hardest part of any evaluation is how to quantify the counterfactual. Any retrospective evaluation involves asking whether one could have achieved better results if one had done it some other way, and it is obviously very difficult to be sure of what would have been the outcome of an alternative strategy.
>
> —Montek Singh Ahluwalia, Deputy Chairman, Planning Commission, India

> Incentives for making evidence-based decisions are still lacking. You don't see discussions in Congress based on evaluation results; this is still a predominantly politically driven process. Instead, incentives exist to spend funds allocated at the beginning of the year by the year's end in whatever way possible. The way forward will depend on moving from individual leadership in recognizing the need for policymaking based on rigorous evidence to an institutional way of operating.
>
> —Gloria Rubio, Acting Director General of Monitoring and Evaluation, Ministry of Social Development, Mexico

The observation that more impact evaluations are needed is often met with skepticism—if not outright rejection. Many of the main objections reflect an underlying concern that we cannot learn about what works in social programs in a meaningful way within real-world constraints. It is commonly held that impact evaluations:

- Tell us little about social programs that we do not already know. *But good studies can avoid costly mistakes and prevent doing harm.*
- Are not needed to demonstrate success. *But good studies can identify successes even under adverse circumstances, where success means doing less badly.*
- Are not necessary to know which programs work. *But good studies distinguish real successes from apparent successes.*
- Cannot address important issues. *But current methods can answer questions that are important to social policy decisions.*
- Cannot be ethically implemented. *But ethical issues can be managed.*
- Are too costly. *But ignorance is more expensive than impact evaluations.*
- Produce results too late to be of use to decisionmakers. *But impact evaluations can provide timely information.*
- Do not provide important information about how programs operate. *But impact evaluations complement other studies, they do not replace them.*
- Are too complex and do not influence policymaking. *But findings from impact evaluation can be simple and transparent.*

Many of the main objections reflect an underlying concern that we cannot learn about what works in social programs in a meaningful way within real-world constraints

Good studies avoid costly mistakes and prevent doing harm

Findings from impact evaluations can help to avoid costly mistakes. For example, an Indian NGO (Seva Mandir) decided to hire a second teacher for their nonformal education centers in the hopes that it would increase attendance and attainment levels. Of the NGO's 42 centers, 21 were randomly selected to receive a second teacher. Intermediate indicators—such as the number of days the school was in session—improved, but test scores remained the same. The NGO was able to see that the benefits of the two-teacher initiative were not justified by the cost and redirected its funds to expand other, more promising programs (Banerjee, Jacob, and Kremer 2001).

The risk of wasting funds on ineffective programs is particularly acute for programs that are expected to be scaled up to a national level—and the value of proper evaluations in the early stages is correspondingly high. For example, the Drug Abuse Resistance Education (DARE) program had been adopted in 75 percent of US school districts because it was believed to be effective; however, evaluations with random assignment demonstrated that the program was ineffective, wasting financial resources and school time (Lynam and others 1999; Rosenbaum and Hanson 1998).[8]

The value of impact evaluations takes on particularly urgency in social programs, where real harm can be done. For example, a US program called Scared Straight sought to reduce juvenile delinquency by taking at-risk youths to visit prisons. Program advocates asserted that rigorous evaluations were not necessary, but evidence from nine randomized experiments

subsequently showed not only that this popular and well intentioned program was not effective but that it was harmful, leading to higher delinquency among participants (Petrosino, Turpin-Petrosino, and Finckenauer 2000).

Good studies can identify success even under adverse circumstances

For those convinced of the efficacy of their programs, money spent on demonstrating impact through comparisons of participants and nonparticipants may seem unnecessary. However, without such comparisons beneficial programs that mitigate negative trends might be mistakenly viewed as failures. For example, numerous programs to prevent the spread of HIV/AIDS are being financed around the world, but the best they can hope for in the short run is to slow the rate at which prevalence is increasing. Therefore, unless the programs can demonstrate that the rate at which the disease has spread in their target group is lower than in appropriate comparison populations, they will look like failures.

Poorly done evaluations may mistakenly attribute positive impacts to a program when the positive results are due to something else

This ability to identify a successful program under adverse circumstances was demonstrated clearly with a US Department of Labor Summer Training and Education program. A random-assignment study found that disadvantaged teens lost half a grade in reading ability—apparently a complete failure. However, nonparticipants lost a full grade of reading ability. The evaluation demonstrated that the program mitigated the loss of reading ability that naturally occurred during the summer vacation months (Grossman 1994).

Good studies distinguish real successes from apparent successes

Poorly done evaluations may mistakenly attribute positive impacts to a program when the positive results are due to something else. For example, retrospective studies in Kenya erroneously attributed improved student test scores to the provision of audiovisual aids. More rigorous random-assignment studies demonstrated little or no effect, signaling policymakers of the need to consider why there was no impact and challenging program designers to reconsider their assumptions (Glewwe and others 2004). Similarly, a US program to assist poor families through social service visits found that families receiving the program experienced improvements in welfare—but so did the families that were randomly assigned to a control group that did not receive the visits (St. Pierre and Layzer 1999). In both cases, a good study helps avoid spending funds on ineffective programs and redirects attention to improving designs or to more promising alternatives.

Current methods can answer questions that are important to social policy decisions

It is difficult to design high-quality impact evaluations that can answer such policy questions as, “Under what circumstances should a country have a fixed exchange rate?” Nevertheless, the range of questions that can be answered by well designed impact evaluations is much wider than is generally recognized.

Studies aimed at learning the best way to assist individuals can be relatively easy to design but still require time and money. For example, certain questions can be easily studied by comparing participants and nonparticipants because they relate to how individuals respond to specific interventions:

- Do vitamin A supplements reduce infant mortality (Sommer and others 1986)?
- Do textbooks increase students learning (Glewwe, Kremer, and Moulin 2001)?
- Do microfinance programs improve child nutrition (MkNelly and Dunford 1998)?

Even questions that might be considered particularly difficult to answer—such as the impact of gender on political decisionmaking—can be rigorously studied

But questions regarding the best ways to "produce" services—requiring comparisons across classrooms, facilities, or districts—can also be addressed in a relatively straightforward fashion:

- Does hiring an additional teacher in nonformal schools improve attendance and performance (Banerjee et al 2003)? Does class size affect student learning (Angrist and Levy 1999; Mosteller and Sachs 1996)?
- Does rewarding teachers or children for improved test scores lead to sustained boosts in students' learning? Which is more effective: incentives for teachers or for students (Glewwe, Ilias, and Kremer 2003; Kremer 2003)?
- Does community monitoring of development programs reduce corruption? Is it more or less effective than audits (Olken 2004)?

It is feasible, though more resource intensive, to use impact evaluation to assess programs that have externalities—that is, to measure the net impact on a person (or community) from a program that was delivered to a neighboring person or community:

- Does school-based mass treatment of children for intestinal parasites, in a high prevalence area, improve health and schooling even for those not receiving the treatment (Miguel and Kremer 2001)?
- Does agricultural extension have effects beyond the farmers who are directly reached through the diffusion of their learning to their neighbors (Conley and Udry 2000)?

In other cases measuring the impact of a national program delivering new social services is possible by contrasting changes across districts or municipalities as they are introduced in successive waves:

- Do cash transfers to poor families that are conditional on school attendance and use of preventive healthcare services improve children's health and schooling (Gertler 2000; Schultz 2000; Buddlemeyer and Skofias 2003)?
- Does building schools lead to improved schooling and earnings (Duflo 2001)?

Even questions that might be considered particularly difficult to answer—such as the impact of gender on political decisionmaking—can be rigorously studied:

- Do quotas for women's participation in political decisionmaking improve allocations of public funds (Chattopadhyay and Duflo 2001)?

In short, the only real limitation on this type of impact evaluation is in addressing questions for which no credible counterfactual can be constructed. But even in these cases there are usually underlying questions that need to be answered through impact evaluation.

All of that said, it is clear that much room remains to improve evaluation methods, compare random assignment and other approaches, and strengthen operational knowledge about the "how to's" of impact evaluation.

Ethical issues can be managed

Impact evaluations that rely on collecting data from control groups are sometimes thought to be unethical because they exclude people from program benefits. But this criticism applies only when resources are available for serving everyone as soon as the program starts. In fact, whenever funds are limited or programs need to be expanded in phases, only a portion of potential beneficiaries can be reached at any time. Choosing who initially participates by lottery is no less ethical (and perhaps even more so) than many other approaches. Some programs are allocated by lottery when they are oversubscribed (school choice in the United States or voucher programs in Colombia) or for transparency and fairness (random rotation of local government seats to be set aside for women in the Indian elections).

Whenever there is reasonable doubt of a program's efficacy or concerns with unforeseen negative effects, ethics demands that the impact be monitored and evaluated

Furthermore, whenever there is reasonable doubt of a program's efficacy or concerns with unforeseen negative effects, ethics demands that the impact be monitored and evaluated. For example, in Mexico opponents of a conditional cash transfer program in the mid-1990s argued that giving funds to poor mothers might increase their vulnerability to domestic abuse. A well designed impact evaluation was able to put those serious concerns to rest.[9] The simple truth is that many well intentioned social programs are like promising medical treatments—we cannot really know if they do more good than harm until they are tested.

Finally, starting with a properly evaluated pilot program can greatly increase the number of eventual program beneficiaries, because the evidence of success will provide support for continuing and expanding an effective program.

Ignorance is more expensive than impact evaluations

It is argued that impact evaluations are too costly or difficult. This argument is often made by comparing the cost of an evaluation with the program that is its subject. But the appropriate comparison is not the program cost but the value of the knowledge it would produce.

For example, evaluations of demonstration training programs in Latin America and the United States have sometimes exceeded a third of the initial program costs, but the evaluation results affected decisions about the rollout of much larger national programs. In these cases the value of scaling up programs that worked and avoiding or redesigning those that were ineffective was extremely high. To the degree that these findings were generalizable, they yielded benefits to other countries as well. Thus a few well selected impact evaluations can generate knowledge that influences the design and adoption of an entire class of interventions around the world.

Sometimes the additional cost of doing a good impact evaluation is quite small. When projects are results oriented and require baseline data, an intelligent design for gathering data can determine whether an impact evaluation will be feasible—sometimes without any additional cost for data collection. Some costs may even be lower in studies of developing country programs because the field costs of surveys and local researchers is generally lower than in higher income countries.

The principal cost of an impact evaluation is the cost of data collection; and the cost of collecting data can be just as expensive for an inconclusive study as for a compelling and rigorous one. For example, a large primary education program in India (the District Primary Education Program) spent millions of dollars collecting data on all the districts in which the program was implemented. But this kind of data collection does not serve as the basis for valid inferences about impact. A proper data collection strategy (for example, randomly choosing which districts would be offered the program and then conducting surveys in a sample of districts that were offered the program and those that were not) might have cost less and would have provided useful information about the program's impact (Duflo and Kremer 2003; Duflo 2004).

The cost of collecting data can be just as expensive for an inconclusive study as for a compelling and rigorous one

This is not meant to imply that rigorous impact evaluation comes cheap. Depending on the question being asked and the associated evaluation design, studies may cost millions of dollars over several years. The point is that the relevant way to assess those costs is relative to the value of the knowledge that will be generated—a value that is measured both in avoiding harm and in reaching more people with proven programs. This kind of questioning arises precisely because the knowledge from impact evaluations is a public good; the incentives to finance these studies do not reflect their full social benefits and, consequently, insufficient investment goes into them.

Impact evaluations can provide timely information

Some challenge the value of impact evaluation studies on the grounds that they take too long to produce results so that by the time the findings are available the programmatic approach has already evolved. However, the time taken to produce results depends a great deal on the questions being studied. Some rigorous impact evaluations produce results within a matter of months. Others take longer, but are still available in time to affect important policy decisions.

For example, the initial findings of Mexico's impact evaluations of its national conditional cash transfer program were available in time to convince a new administration to preserve it. A rigorous impact evaluation comparing different kinds of teachers provided valuable information to the Indian NGO Pratham in time for it to expand a program of community-based teachers who had been shown to be at least as effective as new teachers, but at less cost.

It is also possible to design impact evaluations that generate useful feedback during implementation. For example, a multiyear study of the impact

of HIV education in Kenya was designed to assess intermediate outcomes, such as the accuracy of knowledge about HIV transmission, as well as to monitor the long-term impact. The transfer of knowledge is only a necessary, not a sufficient, condition for the program to have an impact; but measuring the success or failure of reaching such intermediate goals can help program managers make necessary adjustments to improve implementation.

Though useful information can often be collected in shorter time frames, most questions about the impact of social programs require collecting data over years. Valid evidence of a program's effectiveness often cannot be produced in less time. Decisions based on invalid findings are likely to be poor ones, no matter how timely they may be. If a program is still functioning when the impact evaluation is completed, the findings will be useful to those making decisions about it in the future. If the program ends, the findings will still be of value to those who are contemplating similar interventions in other places or interventions that rely on similar principles.

Decisions based on invalid findings are likely to be poor ones, no matter how timely they may be

Criticizing impact evaluations for not providing timely information simply confuses its purposes with other kinds of evaluations and ignores how knowledge about social programs is built: in a time frame that is linked to but extends beyond the normal project cycle. This is why impact evaluations should be done strategically, to answer policy questions that are likely to have enduring importance while focusing on programs likely to generate information that can guide policymakers and managers in the future.

Impact evaluations complement others studies

Critics sometimes claim that impact evaluations can only tell whether something has an impact, not why and how. But a good impact evaluation can provide evidence about the mechanism through which the outcome is achieved when it simultaneously collects information on processes and intermediate outcomes. Impact evaluations are not a replacement for sound theories and models, needs assessment, monitoring, and operational evaluations. All of these elements are necessary to complement the analysis of impact. But knowledge gained from impact evaluations is a necessary complement to these other kinds of analyses. Ideally, different forms of evaluation should not be seen as competitors, but as mutually reinforcing parts of a virtuous cycle.

Findings from impact evaluation can be simple and transparent

The final critique is that impact evaluations are too complex for policymakers and do not influence policymaking. In fact, good impact evaluations, especially randomized evaluations, are relatively easy to present to policymakers. MDRC conducted randomized control trials of numerous state welfare programs in the United States.[10] Because the findings were readily conveyed to policymakers, these studies had a significant impact on US welfare reform legislation in the mid-1980s (Wiseman and others 1991; Gueron 1997, 2002).

Latin America offers other examples of impact evaluations that affected policy. In the 1980s evaluations of radio-assisted education programs in Nicaragua led to widespread replication of this promising intervention

(Jamison 1978). The impact evaluation of Progresa in the mid-1990s is widely credited with preserving that social program in the transition to an opposition administration (the program was retained and expanded, and the name was changed to Oportunidades). Furthermore, the Progresa evaluation influenced the adoption of similar conditional cash transfer programs in many other countries (Morley and Coady 2003).

If we do not start now, then when will we ever learn?

Building the kind of knowledge generated by impact evaluations is one of the best investments we can make

Most important, if impact evaluations are not started today, then we will never have access to the information needed for evidence-based decisions. This point has been made in recent declarations associated with the creation of the Global Fund to Fight AIDS, Tuberculosis, and Malaria; the replenishment of International Development Association and African Development Bank concessional funds; and the formulation of the Millennium Development Goals. In each case attention to measurement of results makes it imperative to lay down the foundations today so that we can learn about the effects of our actions tomorrow. Any investment takes time to yield benefits, and building the kind of knowledge generated by impact evaluations is one of the best investments we can make.

IV. Why good impact evaluations are rare

> I totally agree on the need to close the so-called "Evaluation Gap." Not doing so has been and will continue to be costly in terms of inefficient decisionmaking involved in the design and implementation of social policies. I also agree on the need for a coordinated international response. Knowledge from impact evaluation studies, at least in some of its elements, can be considered a global public good.
>
> —Julio Frenk, Minister of Health, Mexico

The Evaluation Gap Working Group looked carefully at why, amid growing demand for knowledge about what works, the resources and attention devoted to impact evaluation are still so limited. A central cause of the shortfall is the nature of the benefits of impact evaluations.

The benefits of the knowledge generated by impact evaluations go well beyond the organization or place in which they are generated. That means that the collective interest in seeing that a body of impact evaluation research gets built, reviewed, disseminated, and improved is greater than the individual interest of any particular group, agency, or country. Because knowledge from impact evaluations is a public good, the incentives for any individual organization to bear the costs are much lower than the full social benefits would justify. Because the knowledge from impact evaluations is not a "pure" public good—there are often substantial benefits to the country or agency that commissions the study for learning about its own programs—such studies do get done, but not in the numbers or with the quality that are justified by the potentially global benefits.

The situation is complicated by the fact that impact evaluations are not regular, ongoing activities of most organizations, but are applied strategically to programs from which important knowledge can be gained. Impact evaluations do not have to be conducted in-house. Indeed, their integrity, credibility, and quality is enhanced if they are external and independent. They do not take place in step with normal budget and planning cycles. Rather, they must be initiated when the right combination of factors arises—an important policy question, the opportunity to integrate impact evaluation into a new or expanding program design, or the active interest and collaboration of policymakers, funders, researchers, project managers, and beneficiaries.

Structures of decisionmaking generate diffuse demands for knowledge from impact evaluations, not synchronized with the time frame required by such studies, forcing them to compete for resources and attention with project implementation, and even, at times, discouraging the gathering of information that could be potentially constraining or embarrassing. Thus the general problem created by the public good nature of impact evaluation can be further elucidated by looking at some particular obstacles to conducting impact evaluations—those related to the structure of demand, the absence of timely funding, and the patterns of incentives.

When a new program is being designed is precisely when an impact evaluation can most easily be developed to provide useful answers about the program's impact

Demand, money, and incentives

Demand for the knowledge produced by impact evaluations tends to be spread out across many actors and over time.[11] Demand arises every time someone in government, an NGO, a multilateral development bank, or a bilateral agency asks: "What programs are effective at . . .?" This can occur when a new program is being designed, when additional funding is being requested, or when internal reviews of institutional performance are being conducted.

When a new program is being designed is precisely when an impact evaluation can most easily be developed to provide useful answers about the program's impact. That is the moment when program designers want the benefit of prior research, yet have few incentives to invest in a new study. Yet if they do not invest in a new study, the same program designers will find themselves in the same position four or five years later because they missed the opportunity to learn whether the intervention had an impact (O'Donoghue and Rubin 1999). Other institutions and governments that might have learned from the experience also lose when these investments in learning about impact are neglected.

It is in such circumstances that timely availability of funding can make a big difference. Despite a lack of incentives to conduct impact evaluations, many program designers and managers still have an interest in measuring the impact of their programs. When funding for impact evaluation studies is not readily available, it makes it more difficult to act on their interest. For example, the Familias en Acción project in Colombia began implementation in some communities before baseline information was collected. A rapid-response trust fund might have allowed data activities to advance

independently of project approval.[12] If funding is readily available, it can make the difference between doing and not doing a rigorous study.

Other incentives exist at the institutional level to discourage impact evaluations. Government agencies involved in social development programs or international assistance need to generate support from taxpayers and donors. Since impact evaluations can go any way—demonstrating positive, zero, or negative impact—a government or organization that conducts such research runs the risk of findings that undercut its ability to raise funds (Pritchett 2002). Policymakers and managers also have more discretion to pick and choose strategic directions when less is known about what does or does not work. This can even lead organizations to pressure researchers to soften or modify unfavorable studies or simply to supress the results—despite the fact that knowledge of what does not work is as useful as knowledge of what does.

Studies that demonstrate that programs are successful are more likely to be publicized by the participating institutions and also are more likely to be published in academic journals

When such pressures hold sway, a noticeable bias appears in the body of published findings. Studies that demonstrate that programs are successful are more likely to be publicized by the participating institutions and also are more likely to be published in academic journals. A publication bias emerges that provides an unfairly positive assessment of social programs (Dickersin and Min 1993). One way to counter this publication bias is to establish a prospective registry of impact evaluations, that is, to record impact evaluations when they start. Then, future literature reviews can better assess whether published findings are representative or not.

V. Closing the evaluation gap—now

> Every program for improving the lives of poor people in developing countries begins with an intuition about what will work. However, the hopes and good intentions of program implementers, coupled with the human tendency to seek only confirming evidence, leads to programs being initiated and even replicated without learning whether they actually work. In the arena of social change, the null hypothesis is more than an abstract statistical concept. It reflects the reality that many intuitively obvious theories do not in fact produce their intended outcomes. The Evaluation Working Group's proposals offer the beginning of a remedy to this pervasive problem.
>
> —Paul Brest, President,
> The William and Flora Hewlett Foundation,

The time is right for a major push to close the evaluation gap. Why? First, there is a clear personal commitment from a growing community of professionals who recognize the value of impact evaluations, many of them employed by governments and international agencies.

Second, the capacity around the world to collect data and conduct research is growing. Many more people than ever before are trained in impact evaluation methods and in fields that allow them to interpret the findings of such studies, and technological advances have reduced the costs and time of collecting and processing data.

Box 1 Requests from developing countries for systematic production and use of impact studies

- Flexible funding from outside domestic budget procedures to contract experts to assess whether impact evaluations can be conducted and to design them.
- Research centers and training for government staff.
- Linking local research centers to international experts and institutions to build local capacity.
- Providing independent grant review processes or certification for studies that meet internationally accepted standards of reliability and validity.
- Advocating for producing and using impact studies in public policy and educating legislators, journalists, and the public about the value and appropriate use of research findings.
- Documenting specific experiences to show how impact studies are started, implemented, written up, disseminated, and integrated into policymaking.
- Facilitating contacts across countries working on similar issues.
- Registering prospective studies and facilitating access to systematic reviews and high-quality studies.

Third, there appears to be growing recognition in many agencies of the need to measure results; this cannot be done well without complementary studies that get at the issue of attribution.

Finally, skepticism about the use of funds for development assistance puts pressure on agencies to measure impact. When the overriding risk is closure of a program or severe reduction in funding, then the downside risk of negative findings is less problematic. Both managers and project designers see greater benefit in measuring the impact of their programs in the hope that they can demonstrate that the programs should continue.

Beyond these initiatives policymakers in developing countries have expressed interest in gaining access to better information, building knowledge, and incorporating evidence in policy decisions. In interviews, surveys, and meetings officials in developing countries have requested more support for systematic production and use of impact studies. In particular, they have identified a number of services that would benefit their efforts to inform policymaking (see box 1).

For the international community, too, demand for knowledge about impact is growing as a consequence of commitments to substantially increase aid flows in novel ways, efforts to hold agencies accountable for the use of public funds, and emphasis on results and performance. International commitments such as the Millennium Development Goals create both a challenge for impact evaluation and an opportunity to learn.

Starting with what we have

Concern about the evaluation gap is widespread, as demonstrated by the many ways that public agencies, intergovernmental commissions, NGO networks, research centers, and foundations are addressing it. In particular, initiatives are under way to:

- Increase access to existing information through reviews, searchable databases, policy pamphlets, and newsletters.
- Improve regular data collection by developing country governments and produce aggregate indicators.

- Promote specific evaluations with grants and other kinds of funding.
- Conduct research and demonstrate good evaluation practices.

This discussion of initiatives, far from comprehensive, is intended to demonstrate the range of existing efforts.

Access to data and information

Numerous organizations are trying to make information and data more readily accessible

Numerous organizations are trying to make information and data more readily accessible. The Organisation for Economic Co-operation and Development's (OECD) Development Assistance Committee has a searchable Evaluation Inventory of studies by its member bilateral assistance agencies. The Institute of Development Studies, with support from the UK Department for International Development, maintains ID-21, a database of studies, with an associated strategy for outreach and dissemination by e-newsletter. Other initiatives aimed at increasing the exchange of information include the Development Gateway and the Global Development Network and official channels such as the United Nations Evaluation Forum and ECG Net, the Evaluation Cooperation Group's Web site to foster collaboration among the evaluation units of the multilateral development banks.

Some initiatives aim to provide access to knowledge by synthesizing the results of multiple studies. The Campbell Collaboration has established a process to generate systematic reviews of programs in education, crime and justice, and poverty reduction. The Cochrane Collaboration has taken the lead in systematic reviews of medical studies. The Robert Wood Johnson Foundation also has the Synthesis Project on health policy. The Canadian Health Services Research Foundation is currently analyzing methods for the synthesis of social policy studies.

Better data collection

Multiple initiatives aim to improve data collection in developing countries by conducting surveys or building local capacity to establish ongoing data collection efforts. Examples include the Demographic and Health Surveys sponsored by the US Agency for International Development, the Living Standard Measurement Study household surveys sponsored by the World Bank, the Regional Program for Improving Household Surveys and Measurement of Living Conditions in Latin America and the Caribbean program sponsored by the Inter-American Development Bank to support improvement of government statistical offices, and a recent initiative to improve data collection by PARIS21 (Scott 2005).

Other initiatives aim to increase capacity for local stakeholders or researchers to conduct good quality evaluations, including programs sponsored by the World Bank's Independent Evaluation Group, Canadian Health Services Research Foundation, and many bilateral agencies.

In addition, international efforts aim to standardize and systematize the collection and interpretation of indicators, such as the Health Metrics Network, the Partnership for Maternal, Newborn & Child Health, and the UN Millennium Project.

Financing and conducting impact evaluations

Every bilateral and multilateral agency and almost every government has contracted an impact evaluation at some time or another. Some agencies and private foundations have also established grant programs that are open to unsolicited proposals (for example, Canadian Health Services Research, the Bill & Melinda Gates Foundation, and Development Gateway).

Many developing countries are taking their own initiatives to learn from social development programs through better impact evaluations. Agencies in Chile, India, and Kenya have all started or actively collaborated in designing good impact evaluations because they recognize the value of the information they will build. Mexico has even passed legislation requiring impact evaluations of a wide range of social development programs.

The fact that knowledge about the impact of social programs is a public good means that collective actions are most likely to succeed in generating sufficient investment

Research centers worldwide, such as the Institute for Fiscal Studies (London), the National Institute of Public Health (Mexico), the Group for Analysis of Development (Peru), the Institute for Financial Management and Research (India), and the International Food Policy Research Institute (Washington, D.C.) have established reputations in supervising, conducting, and advising impact evaluations of social programs in developing countries. Several international programs aim specifically to increase the number of skilled evaluators in low- and middle-income countries and to contribute to building a supply of researchers and to promoting an appreciation of evaluation findings within public policy debates.

Most international agencies also have internal initiatives aimed at improving impact evaluation. Interviews with staff at multilateral development banks and bilateral agencies indicate that they are aware of the need for better impact evaluations and that several initiatives are under way to improve the number and quality of these studies. The World Bank's Development Impact Evaluation (DIME) program illustrates the kinds of steps that institutions can take to better link their operational and research capacities, in partnership with developing countries, to generate knowledge from impact evaluations on selected thematic areas.

Recommendations

Generating knowledge about what kinds of social programs work best requires actions to increase the effective demand for such knowledge as well as its supply. This means finding ways to create positive incentives to use knowledge in decisionmaking and incentives to conduct appropriate impact evaluations. While much can be done by governments and agencies on their own, the fact that knowledge about the impact of social programs is a public good means that collective actions are most likely to succeed in generating sufficient investment. Organizations can make commitments to one another to do things independently, but they may also find it useful to create new networks, committees, or institutions to carry out particular tasks.

In the sections that follow the Evaluation Gap Working Group lays out two broad recommendations: doing more and better through existing individual efforts and making a major advance through a collective endeavor.

Recommendations for individual action: reinforce existing efforts

At a minimum, governments and agencies should reinforce existing initiatives to generate and apply knowledge from impact evaluations of social programs. This includes strengthening their overall monitoring and evaluation systems; dedicating resources to impact evaluation; ensuring collaboration between policymakers, project managers, and evaluation experts; improving standards for evidence; facilitating access to knowledge; and building capacity in developing countries to conduct rigorous evaluations (see table 1 for a summary of these recommendations).

Countries and agencies need to adequately finance and staff all forms of monitoring and evaluation because they are complementary

- *Strengthen internal monitoring and evaluation systems.* Countries and agencies need to adequately finance and staff all forms of monitoring and evaluation because they are complementary. This means ensuring that programs are monitored, that institutional learning from documenting experiences with processes and implementation takes place, and that information is provided to allow external scrutiny and improve accountability. Bilateral and multilateral agencies should vigorously pursue the channels opened by the Development Assistance Committee Evaluation Network and the Development Assistance Committee/multilateral development bank Joint Venture on Managing for Development Results and follow their recommendations to improve these evaluation functions. Developing countries should recognize that allocating time and resources to building these evaluation functions is an investment with large payoffs.
- *Dedicate resources to impact evaluation.* Countries and agencies need to dedicate sufficient funds to strategically selected impact evaluations. Because knowledge building is a long-term strategic process, it requires that funds be budgeted at the highest levels of an organization so that these tasks will be implemented without interfering with the more visible, day-to-day demands of project management and monitoring.
- *Involve policymakers, project managers, and evaluation experts.* Countries and agencies can improve the content and use of impact evaluations by involving policymakers, project managers, and evaluation experts in their design. This process ensures that relevant questions will be asked, useful data will be collected, and that reliable evidence will be produced. It increases the likelihood that results will be used to improve the social programs in its given context and that the findings will be of interest to people making decisions about similar social programs in other places.
- *Improve standards for evidence.* Countries and agencies should set high standards for evidence on social program impact because this will benefit their own decisionmaking about social programs and build a reputation that will lead other organizations to pay greater attention to their experience and findings. High standards can be established and upheld by submitting impact evaluation proposals to review by independent experts, and to a prospective registry of studies, to increase transparency and assist the research community in identifying publication bias. Finally, completed impact evaluations should be submitted to an external peer-review process to ensure that only valid evidence enters the public domain, signal the quality of the study to policymakers, and act as an incentive for rigor in other studies.

- *Facilitate access to knowledge and its use.* Countries and agencies should also strive to publish their peer-reviewed impact evaluations, regardless of the conclusions. Though it may be difficult to openly disseminate reports that show problems in social programs, organizations will still benefit from showing that they are serious about learning from their programs and acting on good evidence. Furthermore, an organization's favorable reports will gain greater credibility to the extent that the public sees that unfavorable reports are also openly published and debated. Part of this dissemination should include making primary data publicly available for reanalysis and encouraging the production of systematic reviews of evidence.
- *Build capacity in developing countries.* Countries and agencies need to invest in the capacity for producing and using impact evaluations in developing countries through training, collaboration, and information exchanges. Developing countries directly benefit from investing in capacity building because it increases the pool of evaluators and expertise on which they can draw. It also increases the chances that domestic researchers will be hired to evaluate programs in other countries, from which much can be learned. International agencies benefit to the extent that they can contract evaluations more easily and with greater assurance of their quality when a country's local research community can be engaged. Greater local capacity for impact evaluation is likely to have further benefits by creating a culture of decisionmaking that values rigorous evidence. Finally, a strong local research community can give sustained attention to feeding relevant information into domestic policymaking debates.

Table 1 Reinforce existing efforts through reciprocal commitments and benefits

Independent actions to be undertaken by social ministries of developing countries, bilateral agencies, multilateral development banks, research institutions, philanthropic foundations, and nongovernmental organizations

My organization commits to:	*My organization benefits by:*
Strengthen our overall internal evaluation system.	Verifying that inputs are purchased and properly applied; learning how to improve processes; documenting and sharing institutional experience; creating a context for interpreting and acting on the findings of impact evaluations.
Dedicate funds for impact evaluation that are meaningful in relation to our portfolio of activities.	Learning from our own impact evaluations; rationalizing the use of impact evaluation funds by directing them to select topics; reducing free-riding.
Involve policymakers, project managers, and evaluation experts in the design of impact evaluations.	Ensuring that studies ask relevant questions and are designed to generate rigorous evidence and be more useful.
Set high standards for evidence, have independent external reviewers evaluate the quality of proposed impact evaluations, register impact evaluations prospectively, and submit completed impact evaluations to peer review.	Yielding valid inferences about the impact of a specific social program; conferring a reputation for contributing to the evidence base; assessing information that we seek for answering our policy questions in light of potential publication bias; providing our staff with incentives to supervise and produce high-quality studies; making our studies more likely to be read and used when they have been externally validated.
Disseminate studies, publish primary data, and encourage production of systematic reviews.	Encouraging other organizations to collaborate more openly in sharing evidence; subjecting data to reanalysis to allow for corrections and build further knowledge; facilitating the use of information by nonexperts.
Build capacity for producing and using impact evaluations in developing countries through training, collaboration, and information exchange.	Making it more likely to find local researchers with requisite skills to conduct impact evaluations; improving the likely quality of resulting studies; increasing the likelihood that partners will be better informed about research quality, interpretation, and uses.

Countries and agencies will gain from undertaking these activities independently, but the incentives to proceed with such tasks are far more favorable when other organizations engage in complementary ways. An organization benefits by strengthening its internal evaluation systems, but its ability to interpret and use its evaluation information is vastly enhanced when it benchmarks its performance against other organizations and learns from their experiences. Dedicating funds to impact evaluation will benefit an organization's decisions about social programs—all the more so if impact evaluations addressing similar questions are being financed by other groups.

A collective approach permits agencies and governments to take advantage of economies of scale and scope

While an organization benefits by involving policymakers, addressing relevant questions, and following high standards of evidence, it benefits all the more when other organizations do the same because the information they generate will be more relevant and reliable. Transparency in disseminating findings, whether favorable or not, can enhance an organization's credibility and reputation, but sustaining transparency is easier when other organizations assume a similar posture. Finally, capacity building in developing countries is such a large task that any efforts taken individually will be marginal, while the sum of activities by all organizations can have substantial impact. Thus, collective commitments are necessary to ensure that sufficient investments are made to improve the production and use of knowledge about social program impacts.

Recommendations for collective action: commitments to public goods through a new council

While specific actions by individual countries and agencies can reduce the evaluation gap, genuine progress is likely to be faster and more lasting if those countries and agencies collectively commit to increase the number and quality of impact evaluations. A collective approach permits agencies and governments to take advantage of economies of scale and scope (for example, efficiencies associated with having a place to turn to for review of evaluation designs against agreed standards) and helps to share the costs so that individual agencies or governments do not bear the full burden of producing knowledge that others use.

Such collective commitments can take two distinct forms. One form of commitment would resemble a contract, in which each organization agrees to shoulder its share of the required tasks. The second form of commitment would be to support a common infrastructure or platform to carry out functions that are most effectively accomplished jointly. In both cases organizations are assuming responsibilities and reaping benefits by collaborating.

While the benefits of collective action are clear, collective action could take various forms. Through wide-ranging consultations, the Evaluation Gap Working Group identified the following characteristics of a successful new initiative:

- Complementary to existing initiatives.
- Strategic in choice of topics and studies.
- Opportunistic in its approach to supporting good impact studies.

- Linked directly and regularly engaged with policymakers, governments, and agencies.
- Involving collective, voluntary commitment by a set of governments and public and private agencies to conduct their own studies or contribute funds for contracting such studies by others.
- Committed to independence, credibility, and high standards for evidence.

An initiative that meets these criteria must have clearly identified functions that will both redress the evaluation gap and be more efficient if conducted collaboratively. It also requires an institutional design that is feasible, efficient, and accountable and an appropriate funding mechanism.

The Working Group identified a set of core functions and elaborated ideas on funding and institutional design

The Evaluation Gap Working Group debated these points, obtained broad input, and developed a consensus that some entity—whether a committee, network, secretariat, or other organization—is needed as a focal point for leading such an initiative. For the following discussion, this entity is referred to as a "council."[13] The council's membership would include any set of developing country governments, development agencies, NGOs, foundations, and other public and private entities that volunteer to generate new, policy-relevant knowledge about social programs. The Working Group identified a set of core functions and elaborated ideas on funding and institutional design. The recommendations represent what was learned from many different perspectives and are offered as a way to facilitate action rather than to impose a particular solution. Ultimately, the participating members will determine whether to follow this particular set of recommendations and what institutional form to adopt for this collaborative effort (see table 2 for a summary of these recommendations).

Core functions. The Working Group identified functions that would contribute to reducing the evaluation gap, are best carried out collaboratively, and would benefit from the focused attention provided by an entity such as a council. Of these functions, the following were judged to be core functions:

- *Establishing quality standards for rigorous evaluations.* It is costly and confusing for each government or agency to create its own standards for rigor in impact evaluation. A council could periodically convene experts to set a common standard or endorse existing standards (see appendix G for examples). Making standards explicit would facilitate the design of new impact evaluations, serve as a reference in reviewing proposals, and help build local capacity to conduct and interpret studies.
- *Administering a review process for evaluation designs and studies.* Reviewing proposals and studies requires time, money, and knowledge. While larger organizations and agencies may have the capacity, smaller ones do not. Even larger organizations and agencies cannot have expertise in every topic. A council could administer reviews with a rotating panel of experts from different fields on behalf of member organizations—benefiting from economies of scale and scope. The reviews would assess the relevance, validity, and feasibility of the proposed designs or studies, based on criteria established in consultation with participating stakeholders to ensure high standards of evidence.

Table 2 Make collective commitments

To be signed by several social ministries of developing countries, bilateral agencies, multilateral development banks, research institutions, philanthropic foundations, and nongovernmental organizations and executed by an international council

Function	*The council would:*	*My organization would:*	*How my organization would benefit:*
Establish quality standards for rigorous evaluations	• Convene experts to establish or endorse standards for methodological rigor and evidence. • Support advances in evaluation methodologies.	• Participate in establishing internationally agreed standards for rigorous impact evaluations by which we will abide.	• Our studies will gain external legitimacy and we will find it easier to evaluate the quality of evidence coming from other organizations. • We will be informed of methodological advances.
Administer a review process for evaluation designs and studies	• Administer a review process with rotating panels of experts for impact evaluation proposals and completed studies. • Assess whether submitted impact evaluations meet agreed standards of reliability and validity through endorsing existing peer review processes or, when necessary, submitting for separate expert review. • Publicize clear standards for quality of evidence. • Promote production of systematic reviews. • Offer prizes for high-quality research.	• Contract the council to conduct reviews of proposals and studies when we do not have in-house capacity. • Provide staff to act as reviewers in areas of their expertise. • Facilitate linking researchers and project managers when opportunities for impact evaluation arise. • Participate in establishing standards. • Submit impact evaluations for review by council or some other accredited independent process.	• This will reduce our requirements for in-house expertise and improve our ability to identify rigorous evidence in decisionmaking and evaluation design. • We can more readily distinguish between strong and weak evidence. • There will be greater incentives for our staff to produce or supervise rigorous studies. • We will get external legitimacy, credibility, and prestige for the studies we finance or contract.
Identify priority topics	• Convene committees of stakeholders to identify enduring questions for social program decisionmaking to guide members and the council's own programs toward policy-relevant research.	• Participate in committees to identify enduring questions and priority topics.	• We will have the opportunity to influence the focus of impact evaluations in other organizations in light of our own demands for information. • We will learn about the most pressing concerns of other organizations. • We will be more likely to find impact evaluations on issues that are of importance to us.
Provide grants for impact evaluation design	• Manage funds that would be used for assessing whether an impact evaluation is feasible and relevant.	• Contribute to a collective pool of funds for impact evaluation designs.	• Our staff and managers will have access to flexible, timely funding to seize opportunities for initiating rigorous impact evaluations.
Organize and disseminate information	• Maintain a comprehensive database of impact studies, with associated information about the quality of evidence; prospective registry of studies. • Provide open access to primary data. • Document evaluation experiences. • Disseminate training materials.	• Send our documents, prospective studies, training materials, and data to the council.	• We will have access to knowledge generated by other organizations, learning materials, and data. • We will be able to distinguish between strong and weak evidence.
Build capacity to produce, interpret, and use knowledge	• Facilitate access to training materials, programs, and experts. • Encourage institutional collaborations and use of local researchers. • Optionally, finance fellowships, courses, and activities to strengthen local research institutions	• Engage in or provide technical support for capacity building activities for local research.	• We will be more likely to find local researchers with requisite skills to conduct impact evaluations. • Resulting studies are more likely to be of good quality. • Partners will be better informed about research quality, interpretation, and uses.
Create a directory of researchers	• Maintain and provide access to a directory of qualified researchers.	• Submit information regarding qualified researchers. • Make use of references as required.	• We will gain ready access to a list of qualified researchers

Communication and public education	• Raise awareness and advocate for changes in legislation and institutions to establish strong incentives to produce and incorporate knowledge from impact evaluations into policymaking. • Educate legislators, journalists, and civil society in proper uses and value of impact evaluation.	• Coordinate communication and public education activities with the council.	• We will have an external ally to encourage the production and use of quality research.
Administer funds for impact evaluation on behalf of members	• Provide services to members who request and finance them, such as prepare terms of reference, contract and monitor research teams, and convene external review panels to assess research designs and final reports.	• Choose whether to use the council's services.	• We would have the option of contracting the council to manage impact evaluations, taking advantage of its network of review panels, expertise, quality assurance mechanisms, and reputation for independence and credibility.
Administer a pooled fund for impact evaluations	• Solicit, contract, monitor, and ensure the quality of impact evaluations to answer questions identified by members as of enduring importance and of high priority.	• Contribute funds to a pool managed by the council. • Participate with other members in setting priorities for which programs and questions will be addressed by the impact evaluations financed by the pooled fund.	• Our contributions will leverage substantially more impact evaluation than we can achieve on our own, complementing and expanding the value of our impact evaluation work. • We will have access to a growing body of rigorous impact evaluations on questions of importance to us.

By reviewing evaluation designs and assessing completed evaluations according to clear and transparent standards of methodological rigor, the council can help members distinguish between stronger and weaker evidence. By rating the quality of proposals and research, the council would enhance the generation of knowledge from impact evaluations in several ways. Researchers would have greater incentives to do rigorous studies, knowing that the effort would be recognized. Project managers, policymakers, and the public could more easily direct their attention to better evidence. And efforts to build capacity could more easily identify models for emulation.

- *Identifying priority topics.* No government or agency can initiate studies on every policy question that they would like answered. Nor is it necessary to evaluate every program. A collective effort to identify the most pressing and enduring policy questions would help governments and agencies to cluster evaluations around common topics and to focus efforts on programs that are most likely to yield useful information for future policymaking. By participating in such a collective effort, governments and agencies can influence the questions being asked and benefit from studies done by other institutions on programs like their own.
- *Providing grants for impact evaluation design.* The window of opportunity to design a good impact evaluation on an important question is narrow, occurring just at the moment of program conception and design. Often, the missing ingredient is timely funding to contract an expert to meet with stakeholders to assess whether an impact evaluation would be appropriate and what methods would generate the best evidence and then to design the evaluation. With small amounts of money, the council could act as a powerful catalyst—in some cases

making it possible to do impact evaluations that might not otherwise get done; in other cases increasing the likelihood that the money spent on evaluation generates reliable and valid conclusions.

Other functions. Other functions identified in this review are either less critical to the council's mission or require substantially more resources. These functions might be delegated to the council in the future, depending on the council's performance, its staffing, and members' interest and financial support:

A council could collaborate with other organizations to set up a prospective registry of impact evaluations

- *Organizing and disseminating information.* With rapid changes in technology a wide range of databases and search engines are available to find information about what works in social development. However, the sheer number of studies and data that appears in a Web search is daunting without the ability to easily identify which information is most relevant and rigorous. A council could collaborate with other organizations to set up a prospective registry of impact evaluations to address publication bias, maintain databases of completed qualified studies, and encourage the production of systematic reviews.
- *Building capacity to produce, interpret, and use knowledge.* Efforts to build local research capacity and evaluation systems should continue. The creation of a council could enhance these individual efforts by establishing a network of expert reviewers who can also serve as technical consultants and trainers; by rewarding proposals that are led by developing country evaluators or that incorporate genuine partnering with local research institutions; by encouraging new people to enter the field of evaluation with fellowships or involvement in proposals; by briefing public officials, journalists, and civil society organizations on the benefits and uses of impact evaluation; and by disseminating training materials and rigorous evidence.
- *Creating a directory of researchers.* Governments and agencies often have difficulty finding qualified research partners and use the same consultants repeatedly because of the costs of identifying new ones. The council's endorsement of standards for impact evaluation, its network of reviewers, and its database of rigorous studies could generate a directory of researchers with proven skills and expertise. With little additional effort, the council can make this information available to its members and actively encourage the use of qualified experts.
- *Undertaking communication activities and public education.* Politicians and the public do not readily understand impact evaluations, yet impact evaluations are critical to informing public debate and policymaking. The council could explain the benefits and uses of impact evaluation, advocate for legislation and policies to support the production and application of such knowledge, and build public support. The council's network of experts, representatives from member organizations, and its own staff can give impetus to domestic initiatives to strengthen evaluation systems.
- *Administering funds on behalf of members*. Some members might choose to use the council's services to commission studies on their

behalf. Members who have identified a particular topic or project to be evaluated could provide the council with funds for this purpose. The council would issue terms of reference, seek proposals from qualified teams, ensure engagement of local policymakers, convene an external review panel, award the contract, monitor the research phase, and assess the final product using external reviewers. Members could also hire the council to provide one or more of these services independently (for example, to convene an external review panel for assessing the design of an impact evaluation).

Administering a pooled impact evaluation fund. This final function is discussed separately because the Working Group could not reach consensus on its inclusion as a core function. It would involve delegating responsibility to the council to administer a pooled fund dedicated to conducting rigorous impact evaluations of social development programs in developing countries. The council's role would be clearly defined and directed by the members, through a governing body, to commission independent impact evaluations on topics agreed by the members to be of high priority. The studies would be carried out only with the full agreement of the country or countries involved and of any other members financing or involved in implementing the intervention. The impact evaluations itself would be contracted to third parties.

The studies would be carried out only with the full agreement of the country or countries involved and of any other members financing or involved in implementing the intervention

Some Working Group members were concerned that such a fund would divert financial resources from current impact evaluation efforts and might displace responsibility for impact evaluation work that needs to continue within member organizations. Some also expressed preference for a gradual approach, beginning with more modest and immediately feasible functions, in recognition of the difficulties involved in starting any new international initiative.

Others argued that giving the council adequate funds to commission impact evaluations was essential to address the central concerns set out in the analysis: that impact evaluation of development interventions is a public good that will inevitably be underfunded in the absence of a collective effort, that quality and credibility are enhanced when impact evaluations are commissioned externally and independently, and that a pooled approach would facilitate clustering studies around common themes and across different contexts. They also preferred a more ambitious approach out of concern that the sustainability of the current momentum for learning from social development programs may flag in the absence of a collective, albeit voluntary, agreement to create an entity with some measure of funding insulated from immediate political and bureaucratic pressures.

Several Working Group members argued that, given its recognition of the benefits of randomized evaluation, the report should also recommend earmarking some funds from a future initiative for such studies. They were concerned that without such earmarking, the task of "establishing quality standards for rigorous evaluation" remains ill-defined and uncertain.

Funding options

The best solution to a public good problem such as the generation of impact studies is often to obtain a collective agreement from all parties to commit some level of funding to the common effort. Those committed funds can then continue to be applied independently by each party to the agreement. Alternatively, a portion of the funds can be pooled for management by a particular entity. Furthermore, any discussion of funding needs to distinguish between financing impact evaluations and financing a council's core functions and services. These are some of the first questions that prospective members will have to negotiate.

Funding should be negotiated as a package so that each party making a commitment sees that others are also contributing

Funding for core function and services. The council's core functions, again, have aspects of public goods and require agreement to ensure sufficient funds and reduce free-riding. Members would share in the financing of these core functions through contributions reflecting differences in scale and financial resources.

Funding for studies. The essential problems posed by public goods are insufficient investment and free-riding. To avoid these problems, members of the council would commit to finance or contribute to financing impact evaluations that address questions of common interest and enduring importance.

For organizations that fulfill their commitment by commissioning their own studies, the council's role would be to receive information on which studies are being started and implemented and their associated spending. The council would determine whether standards of methodological rigor were being met. The council would report back to members on how much had been accomplished relative to collectively set commitments. (Reporting should include data on spending, but the focus should be on progress in implementing rigorous studies since the goal is obtaining valid evidence.)

For organizations that fulfill their commitment by commissioning others to conduct studies, the council would similarly verify that research designs and studies are rigorous, registering expenditure and reporting.

Developing countries should be equal partners, committing to conduct or finance impact evaluations. Impact evaluations conducted within a particular country with grants or loans from abroad would still count toward that country's membership commitment.

Prospective members will have to negotiate a common agreement to apportion funding based on an interplay of factors. In the final agreement, organizations should be paying for studies or services that match their own internal mandates but that, as part of a common agreement, contribute to the full range of core functions and studies necessary to reduce the evaluation gap. Funding should be negotiated as a package so that each party making a commitment sees that others are also contributing, thereby ensuring that the sum of all efforts is greater than the individual parts.

Institutional options

A further set of questions concern how to constitute the council to best provide collectively beneficial services.[14] The council can be constituted in

many different forms, from an interagency committee to a network, secretariat, or independent organization.[15] The choice will depend on assessing the relevant tradeoffs, and the institutional structure should ultimately be guided by the structure that will best fulfill a range of aims, including high standards of technical quality, independence and legitimacy, operational efficiency, international leadership, and mobilization of additional resources for impact evaluation. Some ideas are presented schematically in table 3.

High standards of technical quality. The greatest risk to this initiative is that it might mobilize additional resources for impact evaluation and end up financing studies that fail to generate strong evidence. Institutional designs that give the council greater autonomy and involve evaluation experts in its governance are more likely to set and maintain high standards. Institutional designs that engage members in negotiating and setting standards

Table 3 Implementation options

	Interagency committee	*Special program within an existing organization*	*Secretariat*
		Structural characteristics	
Governance	Members appoint staff to act as liaison	Members elect a supervisory committee	Members elect a board
Resources required from members	Mainly staff time	Staff time and funds	Staff time and funds
Funding for core functions	None	Membership dues	Membership dues
Funding for impact evaluations	Members participate in funding activities on the basis of voluntary independent commitments that they manage themselves	Members make commitments to spend a specific share of their social program budget, grants, or loans on their own impact evaluations or impact evaluations commissioned by the council	Members make commitments to spend a specific share of their social program budget, grants, or loans on their own impact evaluations or impact evaluations commissioned by the council
Staffing	No specialized staff	Staff dedicated to managing technical review and support	Staff dedicated to managing technical review and support and some administrative functions
		Tradeoffs	
Direct costs	Lowest	Medium	Highest
Indirect costs	Relies on borrowed staff for technical, financial, and administrative functions	Relies on borrowed staff for financial and administrative functions	Lowest; staff time required to participate in committees, reviews, and governance
High standards of technical quality	Difficult; least agile decisionmaking structure and limited autonomy and engagement of technical experts	Moderate difficulty; focused managerial attention but limited autonomy and engagement of technical experts	Least difficult due to focused managerial attention and dedicated technical experts
Independence and legitimacy	Low	Low	High
International leadership	Middle	Low	High
Operational efficiency	Low cost but correspondingly low output; depends critically on efficiency of coordination mechanisms and fulfillment of commitments by members	Moderate costs but commensurately larger output; depends critically on efficiency of host organization and dynamic between members and host organization	Greater direct costs but correspondingly greater output; depends critically on scale economies and coordination with members
Ability to mobilize additional funds	Moderate, depending on how engaged members are and how actively they focus on the initiative	Low, depending on how high a priority is given to the initiative within the host organization	Moderate to high, depending on engagement of members in policy decisions and demonstration of the initiative's value to stakeholders

are more likely to achieve full collaboration in supporting and following such standards.

Independence and legitimacy. The legitimacy of impact evaluations depends, ultimately, on their rigor. If studies are done well, then their findings will be recognized as legitimate through any expert review—whether for publication in an academic journal or inclusion in a systematic review (Campbell Collaboration 2005a, b). However, impact evaluations play a role in spheres of debate that do not necessarily involve expert reviews. In particular, when findings are debated in the press, in civil society, and in government, technical quality is not necessarily self-evident. Other signals may be used to assess a study's legitimacy—such as how it was financed, who conducted it, and how it was approved or published. Institutional designs that provide the council with greater autonomy will benefit the initiative to the extent that the council develops a reputation for independence and integrity. Members will also benefit when they can rely on the council's legitimacy to review and independently validate any impact evaluations they have conducted.

Members will also benefit when they can rely on the council's legitimacy to review and independently validate any impact evaluations they have conducted

Operational efficiency. The council's operational efficiency will be enhanced by exploiting economies of scale and scope and by avoiding duplication of functions. An institutional design that takes advantage of administrative capacities in member organizations or networks is likely to be less costly and more productive than one that establishes new administrative capacities.[16] Institutional designs that rely on existing activities and structures rather than duplication may also be more efficient whenever the marginal costs of assuming additional responsibilities are low.

International leadership. The initiative will be more successful to the extent that it promotes a wider appreciation of the value of rigorous impact evaluations. Being a collaboration of pioneers—committed developing countries, international agencies, foundations, NGOs, and research centers—the initiative will have a natural platform for educating the public, promoting rigorous standards of evidence, and encouraging better use of information in policymaking. The council's ability to lead will be affected by its identification with member organizations, its perceived independence and integrity, and the scale of its capacity to assume new tasks and be proactive. Institutional designs that increase the involvement of members in the council's operation can enhance the council's standing when it takes on leadership tasks; however, such involvement can also reduce the council's agility and flexibility, its capacity to develop focused messages, and its ability to respond quickly to new opportunities.

Will we really know more in 10 years?

Imagining 10 years into the future, when the target date for the Millennium Development Goals has come and gone, the international community could be in one of two situations.

We could be as we are today, bemoaning the lack of knowledge about what really works and groping for new ideas and approaches to tackle the critical challenges of strengthening health systems, improving learning outcomes, and combating the scourge of extreme poverty.

Or we could be far better able to productively use the resources for development, based on an expanded base of evidence about the effectiveness of social development strategies.

Which of those situations comes to pass has much to do with the decisions that leaders in developing country governments, NGOs, and development agencies make over the next couple of years about conducting impact evaluations.

A pioneering effort is required by a few at the leading edge who are ready to seize the opportunity

If a group of leading national governments and development agencies recognizes the tremendous potential of more and better impact evaluations and overcomes the natural institutional resistance to engage in an ambitious new effort, we are convinced that a collective approach will loosen many of the constraints that have contributed to the current situation. Shared agenda setting, high methodological standards, and independent evaluation have the potential to vastly expand and deepen our collective knowledge base.

To work, this need not be a compulsory effort by all members of the international community—every international agency, every developing country government. But a pioneering effort *is* required by a few at the leading edge who are ready to seize the opportunity.

Getting to that collective approach will not be simple. Prospective members will have to choose the functions and institutional design that they think will work best, taking into consideration many tradeoffs. The ideas contained in this report are offered as a point of departure. The single imperative is to reach agreement on an appropriate institutional design as soon as possible to take advantage of current opportunities to learn about what works in social development programs.

Reservations

While the report represents a consensus of the Working Group and its overall conclusions are broadly supported by all the members in their individual capacities, two members requested that their dissenting opinions be registered on issues of particular importance that they feel are not adequately treated in the document.

François Bourguignon and Paul Gertler

The report makes a compelling case that filling the knowledge gap is a critical next step for the development community to take to improve the lives of people living in developing countries. Filling the gap will require a sustained effort to increase the number of impact evaluations of development interventions. Moreover, it will require that similar interventions be evaluated in various settings, with the consequent need for a concerted evaluation effort.

We fully support this message and believe that the report and the consultation process that produced it have served a major role in bringing attention to this issue. We also agree with most of the report's conclusions and recommendations. However, we believe that the report's recommendations need to be strengthened in two areas. First, closing the evaluation gap should not only be about providing more and better information, but also about making sure that the new knowledge is used in ways that improve the lives of people living in developing countries. Second, the success of any effort to improve impact evaluation will depend on the nature of the institutional arrangements created. Below we propose a number of recommendations along these two lines:

- We believe that one of the keys to filling the knowledge gap is strong partnerships with developing country governments, NGOs, and researchers. Countries should both benefit from the new knowledge generated and be full partners in its generation. This requires an effort to build capacity within developing countries. Countries must have both the interest and capacity for quality evaluation to be systematically conducted and institutionalized.
- In our view one of the most critical challenges requiring collective action is related to the *global public goods* aspects of the evaluation gap. These concern mainly the need for systematic information sharing and dissemination within developing countries and development agencies, as well as coordination in the identification of priority topics. We are concerned that these issues have not received the attention they deserve in the recommendations of the report—and indeed have been left out of the core functions of the proposed council.
- We strongly recommend that any proposals for international collective action should encourage and not stifle or otherwise create undue burdens to initiatives to conduct impact evaluations by individual organizations (national or international). National and international groups are heterogeneous in their interest in impact evaluations and their ability to conduct them. Participation in the council could prove burdensome for organizations that have capacity and have already scaled up quality evaluations. This would be the case, for example, for the

establishment of "quality standards" and the associated function of "reviewing" evaluation designs and studies. We believe that it would be better to let individual organization decide which services to use as needed.

- We also recommend that any proposal for establishing a new body for the purpose of allocating funds ensure that the body actually *increase* the total amount of funding available for evaluation and not increase the transaction costs in the allocation of such funding. In particular, we remain worried about creating a new bureaucracy with its own interests in managing funds that should be directed to global public goods.

Reply by Co-Chairs William Savedoff, Ruth Levine, and Nancy Birdsall to Bourguignon and Gertler

We are puzzled by this reservation because the points it makes do not differ from the Working Group consensus and are incorporated into the report. First, we state that the purpose of generating knowledge is to improve policy decisions that can make real differences to the well-being of people in developing countries (for example, pp. 9–10). Second, we explicitly address the issue of different institutional designs (pp. 40–42). Third, we are clear that improving impact evaluations and their use requires full partnership with developing country governments, NGOs, and researchers (throughout the report; see, for example, pp. 31, 40, and 42), noting that such partnerships ensure good design and local relevance. We also highlight the need to strengthen capacity building (p. 33) and propose it as a function for the council (p. 38). Fourth, we agree that information sharing and dissemination are important (pp. 29–30). Indeed, identifying priority topics is included as a proposed core function (p. 37). Fifth, we agree that new resources should be additional to and complementary to existing initiatives (p. 34) and support fully the logic of reinforcing existing activities within member organizations (pp. 32–33). Finally, we nowhere recommend that any member have any obligation to submit any evaluation designs or reports to the proposed council. Membership itself is envisioned as voluntary. Members would decide on how the council would be governed and how such functions as "establishing quality standards" would be made operational (p. 35).

Appendices

Appendix A. Objectives of the Working Group

The Evaluation Gap Working Group was convened by the Global Health Policy Research Network, an initiative of the Center for Global Development, to address the lack of information about the effectiveness of social programs in low- and middle-income countries. Donors, developing country leaders, and development program implementers need to know whether these programs work. But impact measurement is rare. And quality measurement is rarer still. The result? An evaluation gap. Lack of information about what works—and what does not—leaves donors and decisionmakers with little basis on which to defend the wisdom of their investments or make adjustments if needed. In part, this problem reflects the methodological difficulties of measuring the expected impacts of social programs. But experience shows that even when good methodologies exist, disincentives to use them reduce the likelihood that evaluations will be undertaken.

The main objective of the Evaluation Gap Working Group was to develop practical recommendations to solve this problem. The Working Group sought to understand the reasons for the lack of good impact evaluation, with a focus on health and education sectors, and the possible ways to make significant progress toward closing the evaluation gap.

To do this, the Working Group:

1. Reviewed the current status of impact evaluation in social sector programs.
2. Explored the impediments to sustaining good impact evaluations.
3. Consulted with stakeholders, including governments, the research community, private foundations, multilateral and bilateral agencies, and major international nongovernmental organizations.
4. Developed recommendations to address the evaluation gap problem, taking account of other complementary initiatives.

Appendix B. Profiles of Working Group members

The Evaluation Gap Working Group comprised the following members, who served in their individual capacity and not as representatives of their institutions.

Nancy Birdsall is the founding president of the Center for Global Development. She served for three years as senior associate and director of the Economic Reform Project at the Carnegie Endowment for International

Peace, where her work focused on issues of globalization and inequality and reform of the international financial institutions. During 1993–98 she was executive vice president of the Inter-American Development Bank, the largest of the regional development banks, where she oversaw a $30 billion public and private loan portfolio. Before that she spent 14 years in research, policy, and management positions at the World Bank, most recently as director of the Policy Research Department. She is the author, co-author, or editor of more than a dozen books and monographs, and she has written more than 75 articles for books and scholarly journals published in English and Spanish. Shorter pieces of her writing have appeared in dozens of Latin American and US newspapers and periodicals.

François Bourguignon, chief economist and senior vice president for Development Economics at the World Bank, ensures that the Bank develops knowledge that helps guide policy on trade and poverty, economic growth and poverty, aid effectiveness, globalization, international migration, and efforts to achieve the Millennium Development Goals, among others. Since his appointment in October 2003 he has helped put economic growth and impact evaluation of programs and policies at the center of the Bank's research agenda. He previously served as director of the Bank's Development Economics Research Group. Since 1985 he has been a professor of economics at École des Hautes Études en Sciences Sociales in Paris and held academic positions at the University of Chile, University of Toronto, and Bocconi University. He is a fellow of the Econometric Society, was president of the European Economic Association for Population Economics, and received the Silver Medal for Academic Achievements from the French National Center for Scientific Research in 1999. In addition to being the managing editor of the *World Bank Economic Review* (2000–03) and *European Economic Review* (1999–2000), he has authored and edited several books and more than 100 articles in leading journals, including the *American Economic Review*, *Econometrica*, *Review of Economic Studies*, *Journal of Political Economy*, *Journal of Economic Theory,* and *Journal of Development Economics*.

Esther Duflo is the Abdul Latif Jameel Professor of poverty alleviation and development economics in the Department of Economics at the Massachusetts Institute of Technology. She is a co-founder and director of the Poverty Action Lab, research associate at the National Bureau of Economic Research, and on the board of directors of the Bureau for Research and Economic Analysis of Development. She received a master's in economics from DELTA (Paris) in 1995 and a Ph.D. in economics from the Massachusetts Institute of Technology in 1999. She was awarded the Bronze Medal from the Centre National de la Recherche Scientifique (2005), Le Monde's Cercle des économistes Best Young French Economist Prize (2005), and the Elaine Bennett Prize for Research (2003). She is co-editor of the *Journal of Development Economics* and associate editor of the *Journal of Economic Perspectives*. Duflo specializes in development economics and the design and evaluation of

effective antipoverty policy. She has studied household behavior, educational choice and returns to education, decentralization, industrial organization in developing countries, and credit constraints.

Paul Gertler is chief economist of the World Bank's Human Development Network, which works to create better development outcomes in education, health, HIV/AIDS, social protection, children and youth, and disability. He leads the Network's research agenda with the goal of developing evidence-based policy advice focusing on impact and evaluation. Before joining the Bank in 2004, he was Distinguished Professor of Economic Analysis and Policy at the Haas School of Business at the University of California, Berkeley, and professor of health economics and finance at the School of Public Health at the University of California, Berkeley. He has held positions in the Department of Economics at the State University of New York, Stony Brook; the Department of Health Policy and Management and the Department of Economics at Harvard University; and the RAND Corporation. Gertler has experience in consulting and policymaking with the Asian Development Bank, Inter-American Development Bank, Joint United Nations Programme on HIV/AIDS, United Nations Development Programme, World Bank, and World Health Organization, as well as governments in Asia and Latin America and private sector corporations. He earned a Ph.D. in 1985 from the University of Wisconsin, Madison. His awards include the Kenneth Arrow Award in Health Economics (1996); Academic Career Leadership Award, National Institutes of Health (1998); and a Global Development Network Award (2002). He has published more than 75 journal articles and books.

Judith Gueron is a scholar in residence and president emerita at MDRC, a nonprofit, nonpartisan social policy research organization. She joined MDRC as research director at its founding in 1974 and was president from 1986 through August 2004. Under her leadership MDRC became one of the most prominent US policy research organizations, with a mission to design and evaluate education, employment, and social welfare programs to improve the well-being of low-income Americans and enhance the effectiveness of policy and practice. She directed many of the largest federal and state evaluations of interventions for low-income adults, youths, and families and was a pioneer in developing research methods that have made it possible to base social programs on rigorous evidence of effectiveness. Gueron was president of the Association for Public Policy Analysis and Management, served on several National Academy of Sciences committees and federal advisory panels, and frequently testified before Congress. In 2004–05 she was a visiting scholar at the Russell Sage Foundation. Gueron received her Ph.D. in economics from Harvard University in 1971. She has been awarded the American Evaluation Association's Myrdal Prize for Evaluation Practice and the Richard E. Neustadt Award from the John F. Kennedy School of Government. She is a widely published expert on employment and training, poverty, and family assistance.

Indrani Gupta is professor and head of the Health Policy Research Unit of the Institute of Economic Growth, Delhi, India. She has been instrumental in setting up a unit for health economics and policy research, which remains one of the few places in India that undertakes policy-oriented research on the health sector. Her work experience includes teaching economics in India and abroad, working in the government of India where she was a career economist, and being a consultant at the World Bank. She has worked extensively on issues of demand for health and healthcare, health insurance and financing, costing and cost-effectiveness, economics of HIV/AIDS, poverty and health, and implications of global agreements on the health sector in India. She received her Ph.D. in Economics from the University of Maryland.

Jean-Pierre Habicht is graduate professor of nutritional epidemiology at Cornell University. He served as medical officer for the Pan American Health Organization/World Health Organization at the Institute of Nutrition for Central America and Panama (1969–74), where he was in charge of implementing a randomized control trial of the impact of nutritional supplementation on the outcomes of pregnancy and on the health and growth of children. He was a special assistant to the director for the Health and Nutrition Examination Survey, National Center for Health Statistics, US Public Health Service (1974–77). As James Jamison Professor at Cornell University since 1977, he has continued research on child and maternal health and he has used randomized trials and other methods to evaluate the impact on child health of interventions, programs, and policies in Africa, Asia, Australasia, Latin America, and the United States. He served on and was chairman of World Health Organization expert and other committees in family planning, child health, and nutrition. He was chairman of the technical group (AGN/SCN) that advised the United Nations on nutrition. He earned his M.D. (1962) and Doktorat der Medezin (1964) from the University of Zurich, Switzerland, an M.P.H. from Harvard School of Public Health (1968), and a Ph.D. in nutritional biochemistry from the Massachusetts Institute of Technology (1969). He has received numerous prizes for his work, the latest the 2006 McCollum Lectureship in International Nutrition.

Dean T. Jamison is professor of health economics in the School of Medicine at the University of California, San Francisco, and an affiliate of UCSF Global Health Sciences. He is also an adjunct professor at the Peking University Guanghua School of Management and at the University of Queensland School of Population Health. Before joining the University of California, Los Angeles, faculty in 1988, he spent many years at the World Bank as a senior economist in the research department; division chief for education policy; and division chief for population, health, and nutrition. In 1992–93 he was director of the World Bank's World Development Report Office and lead author of *World Development Report 1993: Investing in Health*. In 1994 he was elected to membership in the Institute of Medicine of the US National Academy of Sciences. In 1994 he was elected to membership in the Institute of Medicine of the US National

Academy of Sciences. Jamison studied at Stanford (A.B., philosophy; M.S., engineering sciences) and at Harvard (Ph.D., economics, under K. J. Arrow). His publications cover economic theory, public health, and education. Most recently Jamison served as the senior editor for the Disease Control Priorities Project, where he was involved with preparation of *Disease Control Priorities in Developing Countries*, second edition, and *The Global Burden of Disease and Risk Factors*—both published by Oxford University Press in early 2006.

Daniel Kress, with more than 15 years of experience in international health policy and finance, is a senior health economist at the Bill & Melinda Gates Foundation. He served as senior health economist for the World Bank in the Middle East and North Africa Region, where he was responsible for health projects in excess of $200 million and health sector strategy and dialogue in Algeria, Iran, and Morocco. In addition, he served on many technical review panels at the World Bank and was peer reviewer for "The Millennium Development Goals for Health: Rising to the Challenges," a recent Health, Nutrition, and Population strategy document. Before working at the World Bank, Kress held several assignments in international health. At Abt Associates he served as project director for the Sustainability and Financing of Immunizations Project and as director of research for a $90 million US Agency for International Development (USAID) project to improve health outcomes in the developing world through innovative private sector approaches. Kress has extensive experience in health systems and health reform, including strategies for financing and delivering health services such as community-based insurance, gained from his experience in USAID projects such as Partnerships for Health Reform and PRIME II and under Department for International Development–funded projects in Pakistan. Kress received his Ph.D. in economics from the University of North Carolina at Chapel Hill.

Patience Kuruneri is a public health and development expert at the African Development Bank (AfDB), where she has served since 1992 in various capacities to support its work in health sector investments in more than 40 African countries. Her expertise ranges from developing policy guidance and designing programs and projects to ensuring investment quality. She led the formation of health sector and postconflict investment packages in Ethiopia, The Gambia, Ghana, Nigeria, and Sierra Leone, building on her country-level experience with the United Nations Children's Fund in Ghana (1986–91). She has also worked for the World Health Organization (WHO), Regional Office for Africa (1997–98), and at WHO headquarters (2002–04) on secondment from the AfDB—her last assignment was senior adviser for the Roll Back Malaria Partnership Secretariat on finance and resource mobilization matters involving reviews of global resource tracking systems, costing tools, and innovative financing modalities. Her contributions to the Roll Back Malaria agenda also include helping to define the framework used to establish the Malaria Medicines and Supplies Service.

David I. Levine is a professor at the Haas School of Business at the University of California, Berkeley. He is also research director of the Center for Responsible Business and chair of the Center for Health Research. His research focuses on labor markets and workplaces, particularly what combinations of management policies lead to effective workplaces with high levels of employee skill and decisionmaking. His current work includes how to improve learning about what policies work in economic development and how industrialization has affected children in newly industrializing nations. Levine was an undergraduate at Berkeley, and he has taught at the Haas School since receiving his Ph.D. in economics from Harvard University in 1987. He has also had visiting positions at the Sloan School of Management at the Massachusetts Institute of Technology, the US Department of Labor, and the Council of Economic Advisers. His publications include five books as well as articles in *The American Economic Review, Administrative Science Quarterly,* and *The Review of Economics and Statistics*.

Ruth Levine is a health economist with 15 years of experience working on health and family planning financing issues in Eastern Africa, Latin America, the Middle East, and South Asia. At the Center for Global Development she manages the Global Health Policy Research Network. Before that Levine designed, supervised, and evaluated health sector loans at the World Bank and the Inter-American Development Bank. She also conducted research on the health sector and led the World Bank's knowledge management activities in health economics and finance from 1999 to 2002. From 1997 to 1999 she served as the advisor on the social sectors in the Office of the Executive Vice President of the Inter-American Development Bank. Levine has a doctoral degree from Johns Hopkins University, has published on health and family planning finance topics, and is co-author of *The Health of Women in Latin America and the Caribbean* (World Bank 2001), *Making Markets for Vaccines: Ideas to Action* (CGD 2005), and *Millions Saved: Proven Successes in Global Health* (CGD 2004).

Richard Manning is chair of the Organisation for Economic Co-operation and Development's Development Assistance Committee (DAC), where he took up his duties in 2003. Manning was former director general for policy at the UK Department for International Development (DFID). He worked for DFID and its predecessor agencies from 1965 to 2003, including periods in West Africa and Southeast Asia. He was an alternate executive director at the World Bank. Before becoming chair, Manning worked with the DAC over many years, and from 2001 to early 2003 was chair of the DAC Task Force on Aid Practices, which produced a report on "Harmonising Donor Practices for Effective Aid Delivery" ahead of the High-Level Forum in Rome in February 2003. In March 2005 he was co-chair at the Paris High-Level Forum on Aid Effectiveness.

Stephen A. Quick is director of the Office of Evaluation and Oversight at the Inter-American Development Bank (IDB), where he has served since June 2000. The office reports directly to the IDB's board of executive directors and

contributes to improving the IDB's developmental effectiveness by conducting independent evaluations of its projects and programs. Before that he was manager of the Department of Strategic Planning and Budget at the IDB and advisor to the president on hemispheric affairs. Before joining the IDB, he was the executive director and a chief economist for the Joint Economic Committee of the US Congress, a chief economist for the Senate Democratic Policy Committee, and a senior economist for the House Banking Committee. He has taught at the university level and worked as a private consultant. Quick holds a Ph.D. from Stanford University. He has worked primarily in the areas of international macroeconomics, trade, debt, and finance.

Blair Sachs is a program officer in the Policy and Finance team at the Bill & Melinda Gates Foundation. She develops and manages grants that explore and drive innovative policy and finance solutions for sustainable improvements in increasing access to global health technologies. She leads in developing the global health policy research portfolio and provides policy and finance guidance and analytics to the global health divisions. Originally, her finance efforts focused on the Vaccine Fund and Global Alliance Vaccine Initiative, major grantees in the immunization program. More recently, she has concentrated on policy and financial issues related to the HIV and reproductive health programs. Before working at the foundation, Sachs operated in the business development team of a microbicide biotechnology firm. She also managed health programs with CARE International in Ecuador and assisted a US Agency for International Development project, the Juhudi Women's Association, to initiate a medical dispensary in a rural ward in Tanzania. At Johns Hopkins University Sachs earned an M.B.A. from the School of Professional Studies in Business and Education and an M.P.H. from the School of Public Health.

William D. Savedoff is a senior partner at Social Insight, an international consulting firm with expertise in economic and political analysis of public policy. For more than 15 years he has worked on improving the quality of social services in developing countries in Africa, Asia, and Latin America. Before joining Social Insight, he was an associate researcher at the Instituto de Pesquisa de Economia Aplicada (Rio de Janeiro), a senior economist at the Inter-American Development Bank, and a senior economist at the World Health Organization. He is the author and editor of books, articles, and studies, including *Diagnosis Corruption* (Inter-American Development Bank 2001) and *Wages, Labour, and Regional Development in Brazil* (Aldershot 1995).

Rajiv Shah is the director for Strategic Opportunities at the Bill & Melinda Gates Foundation, where he leads efforts to explore new strategic areas of giving and manages the foundation's special projects portfolio. With an annual budget of more than $100 million, Shah and his team develop foundation strategy and make and manage grants around the world to extend financial services to the poor; expand access to improved water, sanitation, and hygiene; and improve agricultural productivity to reduce poverty and hunger.

Shah previously served as the foundation's deputy director for Policy and Finance for Global Health and as its senior economist. He helped start the foundation's advocacy office, craft its global health strategy, and manage its largest grantee relationship, the $1.5 billion Vaccine Fund. Shah also initiated efforts to develop a portfolio of innovative policy and financing initiatives. Before joining the foundation, he served as a healthcare policy advisor on the Gore 2000 presidential campaign. He co-founded Health Systems Analytics and Project IMPACT and currently serves on boards for the Global Development Network, City Year Seattle, and Time to Vote. Shah has co-authored articles, book chapters, and working group reports on topics ranging from the quality of domestic cardiovascular care to the effective financing and implementation of global health and development initiatives. He earned his M.D. from the University of Pennsylvania Medical School and his M.Sc. in Health Economics at the Wharton School of Business, where he received a National Institutes of Health Medical Scientist Training Grant.

Smita Singh is the director of the Global Development Program at The William and Flora Hewlett Foundation, where she is developing the new philanthropic program to address major global development challenges. Before joining the foundation, she was a scholar at the Harvard Academy of International and Area Studies. Her research interests focus on the comparative political economy of developing countries. She also worked for the Commission on National and Community Service (now the Corporation for National Service), where she developed higher education initiatives and funding strategies for dispersing grants to community service and service-learning projects at more than 200 colleges and universities. Singh also worked at ABC News Nightline and with community-based women's organizations in India.

Miguel Székely was Under-Secretary for Planning and Evaluation at the Ministry of Social Development of Mexico between March 2002 and January 2006. He worked as chief of the Office of Regional Development at the Office of the President of Mexico during 2001 and as research economist at the Inter-American Development Bank from 1996 to 2001. He has a Ph.D. in economics and a master's in economics for development from the University of Oxford and a master's in public policy from the Instituto Tecnológico Autónomo de México. He has lectured on development economics for Latin America at El Colegio de México, Instituto Tecnológico Autónomo de México, and the University of Oxford. His is a specialist on social and economic problems in Mexico and Latin America and has researched widely on inequality and poverty. He has 55 academic publications, including 6 books, journal articles, and chapters in edited volumes.

Cesar G. Victora is a professor of epidemiology at the Federal University of Pelotas in Brazil, which he joined in 1977 after obtaining his M.D. from the Federal University of Rio Grande do Sul. In 1983 he obtained a Ph.D. in healthcare epidemiology at the London School of Hygiene and Tropical Medicine. He has conducted research on maternal and child health and nutrition, equity issues,

and the evaluation of health services, work resulting in more than 300 publications. He works closely with the United Nations Children's Fund and with the World Health Organization, where he is a consultant to the Department of Child and Adolescent Health and Development and was a member of the Advisory Committee on Health Research (2000–04). He is also an honorary professor at the London School of Hygiene and Tropical Medicine, international associate editor of the *American Journal of Public Health*, and editorial consultant at *The Lancet*. He won the Conrado Wessel Prize for Medicine in 2005.

Jessica Gottlieb is a program coordinator at the Center for Global Development, where she manages several working groups in the Global Health Policy Research Network. A graduate of Yale University, she has previously worked at the Academy for Educational Development on international health programs and conducted research on health systems in Mali and France.

Appendix C. Individuals consulted and consultation events

Arnab Acharya, London School of Hygiene and Tropical Medicine; David Adams, US Agency for International Development; Philibert Afrika, African Development Bank; Martha Ainsworth, World Bank; Negar Akhavi, Avahan India AIDS Initiative and Bill & Melinda Gates Foundation; Jeff Albert, Aquaya; Eduardo Amadeo, formerly with the Ministry of Social Development, Argentina; Venkatesh Athreya, M. S. Swaminathan Research Foundation; Orazio Attanasio, Institute for Fiscal Studies; Ganesan Balachander, Ford Foundation; Moses Banda, Government of Zambia; Abhijit Banerjee, Massachusetts Institute of Technology; Rukmini Banerji, Pratham; Cynthia S. Bantilan, International Crops Research Institute for the Semi-Arid Tropics; Owen Barder, Center for Global Development; Geoff Barnard, University of Sussex; Douglas Barnett, African Development Bank; Jon Baron, Coalition for Evidence-Based Policy; Jere Behrman, University of Pennsylvania; Fantahun Belew, Ministry of Finance and Economic Development, Ethiopia; Sharon Benoliel, US Agency for International Development; Bob Berg, United Nations Association; Stephano Bertozzi, Institute of Public Health, Mexico and University of California, Berkeley; Suman Bery, National Council for Applied Economic Research, India; Shashanka Bhide, National Council for Applied Economic Research, India; Ties Boerma, World Health Organization; David Bonbright, Keystone; Bob Boruch, University of Pennsylvania; Tom Bossert, Harvard University; Paul Brest, The William and Flora Hewlett Foundation; Mayra Buvinic, Inter-American Development Bank.

Catherine Cameron, UK Department for International Development; Agulhas, Inc.; Siobhan Carey, UK Department for International Development; Karen Cavanaugh, US Agency for International Development; Jan Cedergren, Ministry of Foreign Affairs, Sweden; Madhav Chavan, Pratham; Cynthia Clapp-Wincek, US Department of State; Warren Clark, Washington National

Cathedral; Daniel Clay, University of Iowa; Paul Clements, University of Michigan; Jorge Coarasa, Ministry of Social Develompent, Mexico; Amy Coen, Population Action International; Neils Dabelstein, Danish International Development Agency; Samantak Das, National Council for Applied Economic Research, India; Phil Davies, Government Social Research Unit, United Kingdom; Antonie de Kemp, Ministry for Foreign Affairs, Netherlands; Paul Delay, Joint United Nations Programme on HIV/AIDS; Dorothy DeMoya, Campbell Collaboration; Elizabeth Docteur, Organisation for Economic Co-operation and Development; Andrew Donaldson, National Treasury, South Africa; Krista Donaldson, Aquaya; Jean Duff, Washington National Cathedral; Joe Eichenberger, Asian Development Bank; Sally Ethelston, Population Action International; Richard Feachem, Global Fund to Fight AIDS, Tuberculosis, and Malaria; Steven Feldstein, Office of the US Under Secretary for Economic, Business, and Agricultural Affairs; Ariel Fiszbein, World Bank; Marilyn Sheldon Flynn, School of Social Work, University of Southern California; Esther Forgan, UK Department for International Development; Birger Forsberg, Karolinska Institutet; Tamara Fox, The William and Flora Hewlett Foundation; Julio Frenk, Ministry of Health, Mexico; Linda Frey, The William and Flora Hewlett Foundation.

Diane Gagnon, Canadian Health Services Research Foundation; Coralie Gevers, World Bank; Gourisankar Ghosh, Water Supply and Sanitation Collaborative Council; Rachel Glennerster, Massachusetts Institute of Technology; David Goldsbrough, International Monetary Fund; Patrick Grasso, World Bank; Robert Greenhill, Canadian International Development Agency; Charles Griffin, World Bank; Rajat Gupta, McKinsey & Co., Inc.; Geoffrey Gurd, Public Health Agency of Canada; Libby Haight, University of California, Santa Cruz, and CETA-IFAI; Andy Haines, London School of Hygiene and Tropical Medicine; Jeffrey Hammer, World Bank; Karin Hannes, Belgian Centre for Evidence-Based Medicine; Ricardo Hausman, Harvard University; John Heath, UK Department for International Development; Jim Heiby, US Agency for International Development; Gonzalo Hernandez, Ministry of Social Development, Mexico; Gonzalo Hernandez, National Council on Evaluation of Social Programs, Mexico; C. R. Hibbs, The William and Flora Hewlett Foundation; Jeremy Hurst, Organisation for Economic Co-operation and Development; Joy Hutcheon, UK Department for International Development; Gregory Ingram, World Bank; Paul Isenman, Organisation for Economic Co-operation and Development.

Krista L. Jacobs, University of California, Davis; Ruth Jacoby, Ministry of Foreign Affairs, Sweden; Pierre Jacquet, Agence Française de Développement; Calestous Juma, Harvard University; Margaret M. Kakande, Ministry of Finance, Uganda; Robert N. Kaplan, Inter-American Development Bank; Mudit Kapoor, India School of Business; Mihira Karra, US Agency for International Development; Rupinder Kaur, National Council for Applied Economic Research, India; Patrick Kelley, Institute of Medicine; Janet Kerley, US Agency for International Development;

Neelima Khetan, Seva Mandir; Leonard Njunge Kimani, National Economic and Social Council, Kenya; Kristi Kimball, The William and Flora Hewlett Foundation; Geeta Kingdon, University of Oxford; Urbanus Kioko, University of Nairobi; Daniel Klagerman, Ministry of Economy, France; Bongiwe Kunene, Department of Economic Development, South Africa; Mylene Lagarde, London School of Hygiene and Tropical Medicine; Carol Lancaster, George Washington University; Dan Levy, Harvard University; Eduardo Lora, Inter-American Development Bank; Santiago Levy, formerly with the Instituto Mexicano del Seguro Social, Mexico; Eva Lithman, Development Assistance Committee Evaluation Network; Virginia Loo, Bill & Melinda Gates Foundation; Hans Lundgren, Organisation for Economic Co-operation and Development.

Sarah Macfarlane, University of California, San Francisco; Gokul Mandayam, School of Social Work, University of Southern California; Delene Mark, HOPE Africa; Oswald Mashindano, Economic and Social Research Foundation; Lynn McDonald, University of Toronto; Di McIntyre, University of Cape Town; Carol Medlin, University of California, San Francisco; Nicolas Meisel, Agence Française de Développement; Guadalupe Mendoza, The William and Flora Hewlett Foundation; Jim Michel, formerly with the Organisation for Economic Co-operation and Development; Lew Miller, Wentz/Miller & Associates; Global CME Newsletter; Anne Mills, London School of Hygiene and Tropical Medicine; Marc Mitchell, Harvard School of Public Health; Namhla Mniki, African Monitor; Jean-Paul Moatti, University of the Mediterranean, Aix-Marseille; Bruce Montador, Canadian International Development Agency; Nachiket Mor, ICICI Bank; Grant Morrill, US Agency for International Development; Ricardo Mújica, Ministry of Social Development, Mexico; Anit Mukherjee, National Institute of Public Finance and Policy, India; Bruce Murray, Asian Development Bank; Vidhya Muthuram, Institute for Financial Management Research.

Revd. Njongonkulu Archbishop Ndungane, African Monitor; Jodi Nelson, International Rescue Committee; Binh Nguyen, Asian Development Bank; Walter North, US Agency for International Development; Ngozi Okonjolweala, Minister of Finance, Nigeria; Doug Owens, Stanford University; Andy Oxman, University of Oslo; Natasha Palmer, London School of Hygiene and Tropical Medicine; Rodrigo Parot, Inter-American Development Bank; Alan Pearson, The Joanna Briggs Institute; Carol Peasley, US Agency for International Development; Emily Pelton, independent consultant; Eduardo Gonzalez Pier, Ministry of Health, Mexico; Max Pulgar-Vidal, Inter-American Development Bank; Rosemary Preston, University of Warwick; Lant Pritchett, World Bank; Laura Rawlings, World Bank; Steve Rhee, Aquaya; James Riccio, MDRC; David Rich, Aquaya; Juan Rivera, Instituto Nacional de Salud Pública, Mexico; Michael Roeskau, Organisation for Economic Co-operation and Development; Robin Roizman, US House of Representatives Committee on International Relations; Howard Rolston, US Department of Health and Human Services;

David Ross, London School of Hygiene and Tropical Medicine; Amanda Rowlatt, UK Department for International Development; Nilmini Rubin, US Senate Foreign Relations Committee; Gloria M. Rubio, Ministry of Social Development, Mexico; Jim Rugh, CARE; Inder Ruprah, Inter-American Development Bank.

Michael Schroll, World Health Organization; Bernhard Schwartlander, Global Fund to Fight AIDS, Tuberculosis, and Malaria; Jeffrey Searle, University of Durham; Claudia Serrano, Asesorías para el Desarrollo; Charles Sherman, US National Institutes of Health; Sara Sievers, Bill & Melinda Gates Foundation; Goberdhan Singh, Canadian International Development Agency; Tara Sinha, Self Employed Women's Association; Haluk Soydan, School of Social Work, University of Southern California; Lyn Squire, Global Development Network; Ray Struyk, Urban Institute.

Charles Teller, US Agency for International Development; Anisya Thomas, Fritz Institute; Ranjeetha Thomas, Global Development Network; Vinod Thomas, World Bank; Erin Thornton, DATA; Brian Trelstad, Acumen Fund; Rob D. van den Berg, Global Environment Facility; Ann Van Dusen, Washington Area Women's Foundation; Prashanth Vasil, McKinsey & Co., Inc.; Anthony Venables, UK Department for International Development; Jeroen Verheul, Ministry for Foreign Affairs, Netherlands; Swati Vyas, Self Employed Women's Association; John Wallace, MDRC; Andrea Wang, University of Oxford; Andrew Warner, Millennium Challenge Corporation; Delia Welsh, Millennium Challenge Corporation; Howard White, World Bank; Barbara Wynn, RAND Institute; Sarah Zalud, Brookings Institution; Joan Zlotnik, Institute for the Advancement of Social Work Research; Eliya Zulu, African Population and Health Research Center.

The Evaluation Gap Working Group process and findings were discussed at the following meetings:

Health Metrics Network. May, 2004. Johns Hopkins University, Baltimore, Md.

VII Meetings of the LACEA/IADB/WB Research Network on Inequality and Poverty. November 3, 2004. San Jose, Costa Rica.

World Health Organization. Staff involved in GAVI and Health Metrics Network. November 9, 2004. Geneva, Switzerland.

Global Fund to Fight AIDS, Tuberculosis, and Malaria. November 8, 2004. Geneva, Switzerland.

Development Assistance Committee Evaluation Network. November 10, 2004. Paris.

2ème conférence AFD/EUDN. "Aide au développement: Pourquoi et Comment. Quelles stratégies pour quelle efficacité?" November 25, 2004. Paris.

Center for Global Development. "What's Next for the World Bank?" September 23, 2005. Washington, D.C.

Center for Global Development. "Commitment to Development Index Consortium: First General Meeting." December 5, 2005. Washington, D.C.

The Sixth International Campbell Collaboration Colloquium. February 22–24, 2006. Los Angeles, Calif.
US Senate Committee on Foreign Relations hearing on "Multilateral Development Banks: Promoting Effectiveness and Fighting Corruption." March 28, 2006. Washington, D.C.
Development Assistance Committee Evaluation Network Meeting. March 30–31, 2006. OECD Headquarters, Paris.
Annual Retreat of the Independent Evaluation Group of the World Bank. May 11, 2006. Washington, D.C.

The Center for Global Development convened meetings to discuss the consultation draft at the following:
The William and Flora Hewlett Foundation. July 25, 2005. Menlo Park, Calif.
Center for Global Development. August 1, 2005. Washington, D.C.
Center for Global Development. September 23, 2005. Washington, D.C.
Ministry of Social Development (SEDESOL). February 9–10, 2006. Mexico City.
London School of Hygiene and Tropical Medicine. March 6, 2006. London.
National Council of Applied Economic Research. April 7, 2006. New Delhi, India.
African Monitor. May 18–19, 2006. Cape Town, South Africa.
Communiques of the Mexico, India, and South Africa meetings (www.cgdev.org/section/initiatives/_active/evalgap).

Appendix D. Existing initiatives and resources

Existing initiatives

The Coalition for Evidence-Based Policy offers "Social Programs That Work," a Web site providing policymakers and practitioners with clear, actionable information on what works in social policy, as demonstrated in scientifically valid studies. Specifically, the Web site summarizes the findings from a select group of well designed randomized controlled trials ("gold standard" studies) that have particularly important policy implications—they show, for example, that a social program greatly affects life outcomes of participants or that a widely implemented social program has little or no effect. [www.evidencebasedprograms.org/].

The International Organization for Cooperation in Evaluation (IOCE), a loose alliance of regional and national evaluation organizations from around the world, builds evaluation leadership and capacity in developing countries, fosters the cross-fertilization of evaluation theory and practice around the world, addresses international challenges in evaluation, and assists the evaluation professionals to take a more global approach to identifying and solving problems. It offers links to other evaluation organizations; forums that network evaluators internationally; news of events and important initiatives; and opportunities to exchange ideas, practices, and insights with evaluation associations, societies, and networks. [http://ioce.net].

The Abdul Latif Jameel Poverty Action Lab (J-PAL) fights poverty by ensuring that policy decisions are based on scientific evidence. Located in the Economics Department at the Massachusetts Institute of Technology, J-PAL brings together a network of researchers at several universities who work on randomized evaluations. It works with governments, aid agencies, bilateral donors, and NGOs to evaluate the effectiveness of antipoverty programs using randomized evaluations, disseminate findings and policy implications, and promote the use of randomized evaluations, including by training practitioners to carry them out. [www.povertyactionlab.com/].

The Campbell Collaboration, a nonprofit organization, helps people make well informed decisions about the effects of social, behavioural, and educational interventions. Its objectives are to prepare, maintain, and disseminate systematic reviews of studies of interventions. It acquires and promotes access to information about trials of interventions. And it develops summaries and electronic brochures of reviews and reports of trials for policymakers, practitioners, researchers, and the public. [www.campbellcollaboration.org].

The DAC Network on Development Evaluation, bringing together representatives from 30 bilateral and multilateral development agencies, works to improve evaluation for more effective development assistance. [www.oecd.org/site/0,2865,en_21571361_34047972_1_1_1_1_1,00.html].

Inter-American Development Bank's EvalNet is a Latin American evaluation network created by the Office of Evaluation and Oversight. EvalNet serves as a forum for practitioners and academics interested in evaluation of development projects in Latin America and the Caribbean. [www.iadb.org/ove/Default.aspx?Action=WUCHtmlAndDocuments@EvalNet].

The International Program for Development Evaluation Training offers online course modules aimed at increasing evaluation capacity among senior and midlevel audit and evaluation professionals working in developed and developing country governments, bilateral and multilateral development agencies, and NGOs. It offers a two-week core course covering development evaluation basics, followed by two weeks of 26 freestanding workshops on specific development evaluation topics. [www.ipdet.org/].

OECD-DAC Evaluation Abstracts Inventory provides summaries of evaluations available throughout the international development donor community. Abstracts are provided along with the full text of the report if it is available. Evaluations can be retrieved by donor, country/region, sector, evaluation type, date, or keyword. The Web site is supported by the DAC Working Party on Aid Evaluation of the OECD and managed by the Canadian International Development Agency. [www.dac-evaluations-cad.org/].

The Development Impact Evaluation (DIME) Initiative of the World Bank aims to overcome pre-existing bottlenecks—insufficient resources, inadequate incentives, and, in some cases, lack of knowledge and understanding—that limit the Bank's ability to conduct impact evaluations at the necessary scale and continuity. The DIME Initiative is a Bank-wide collaborative effort under the leadership of the Chief Economist. It is oriented at increasing the number of projects with impact evaluation components, particularly in strategic

areas and themes, increasing the ability of staff to design and carry out such evaluations, and building a process of systematic learning on effective development interventions based on lessons learned from completed evaluations. [http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/0,,contentMDK:20381417~menuPK:773951~pagePK:64165401~piPK:64165026~theSitePK:469372,00.html].

Organizations with evaluation material

The United Nations Children's Fund (UNICEF) maintains a database of abstracts and full reports of evaluations, studies, and surveys related to UNICEF programs. [www.unicef.org/evaldatabase/index_13364.html].

The United Nations Development Programme's (UNDP) Central Evaluation Database contains summaries of evaluation reports. [http://stone.undp.org/undpweb/eo/cedab/eotextform.cfm]. The Evaluation Plan Database provides information about the UNDP's planned and ongoing evaluations. [http://www.undp.org/eo/database/evp/evp.html].

The United Nations Educational, Scientific, and Cultural Organization's (UNESCO) Internal Oversight Service has links to UNESCO's evaluation reports. [www.unesco.org/ios/eng/ios_intermed5evnreports.htm].

The United Nations Population Fund's (UNFPA) Monitoring and Evaluation Resources section includes full evaluation reports and findings of UNFPA-supported projects and programs. [www.unfpa.org/monitoring/reports.htm].

The Active Learning Network for Accountability and Performance in Humanitarian Action's (ALNAP) Evaluative Reports Database on evaluations of humanitarian action is fully searchable, with key sections and summary information. [www.alnap.org/database.html].

The Development Assistance Committee Evaluation Group's Evaluation Inventory contains evaluation abstracts in English and French that various international development organizations have agreed to make available to the general public. [www.dac-evaluations-cad.org/].

The Asian Development Bank's (ADB) Web site includes a database of full reports since 1995 and a catalogue of online and print publications. [www.adb.org/Evaluation/reports.asp].

The Inter-American Development Bank's (IDB) Office of Evaluation and Oversight provides access to IDB reports on thematic and country evaluations and abstracts of evaluations. Many publications are available in Spanish. [www.iadb.org/ove/DefaultNoCache.aspx?Action=WUCPublications@evaluations].

The World Bank Group's Operations and Evaluation Department publishes a variety of document series, sorted by type or available through its online database. [http://lnweb18.worldbank.org/OED/OEDDocLib.nsf/OEDSearch?openform].

The Impact Evaluation Thematic Group (PovertyNet) Web site provides access to selected evaluations. [www.worldbank.org/poverty/impact/practice/introevl.htm].

And the WBI Evaluation Group carries out evaluations of all training activities. [http://info.worldbank.org/etools/WBIEG/publications/index.cfm?pg=getPubs&category=Publications&Intro=yes&instructions=no&showDetails=no&ID].

The International Development Research Centre's (IDRC) Evaluation Unit offers evaluation reports, electronic resources, and links related to each of the IDRC's three programming areas and the corporate level. The IDRC Library (www.idrc.ca/library/) provides public access to development information through two databases. The BIBLIO database provides information on the IDRC's collection of research materials. [http://idrinfo.idrc.ca/scripts/minisa.dll/144/LIBRARY?DIRECTSEARCH].

The IDRIS database provides information on IDRC-funded development research activities. [http://idrinfo.idrc.ca/scripts/minisa.dll/144/IDRIS?DIRECTSEARCH]. Some sections also available in French and Spanish.

The UNICEF Innocenti Research Centre (IRC) offers access to several resources, including internal bibliographic resources and databases, and external databases, online resources, and organizations. [www.unicef-icdc.org/resources/].

The Australian Agency for International Development's (AusAID) Evaluation and Quality Assurance section includes access to its Lessons Learned Database. [http://akwa.ausaid.gov.au/akw.nsf/4f1db83b64c26548ca256cef0009afa0?OpenForm].

German Development Cooperation (BMZ) materials on evaluation can be found in its Specialist Information section. [www.bmz.de/en/service/infothek/fach/index.html].

The Danish International Development Agency (Danida) has a Web site providing access to ongoing evaluations and information on its evaluation policy and programme. [www.um.dk/en/menu/DevelopmentPolicy/Evaluations/].

The UK Department for International Development (DFID) provides a Web site for searching its publications, including country reports, project evaluations, and strategy reports. [www.dfid.gov.uk/pubs/].

Finnida's abstracts of evaluations of country and sectoral programs s are available online. [www.jica.go.jp/english/evaluation/report/index.html].

The Japan International Cooperation Agency's (JICA) Web site contains online publications, including full evaluation reports, up to 2002. [www.jica.go.jp/english/evaluation/report/index.html].

The Swedish International Development Agency's (Sida) evaluation Web site includes access to full reports of evaluation series and studies in evaluation. [www.sida.se/Sida/jsp/polopoly.jsp?d=2269&a=17840].

The United States Agency for International Development (USAID) has launched an initiative to revitalize its evaluation work; its Web site includes links to impact evaluation reports. An annotated bibliography of impact evaluations conducted by USAID is available on request. [www.dec.org/partners/evalweb/evaluations/index.cfm].

Appendix E. Selected examples of program evaluations and literature reviews

The following are examples of program evaluations that were unable to achieve their goals due to insufficient availability of impact evaluation studies or appropriate information on nonprogram participants.

Karim, R., S. A. Lamstein, M. Akhtaruzzaman, K. M. Rahman, and N. Alam. 2003. "The Bangladesh Integrated Nutrition Project Community-Based Nutrition Component: Endline Evaluation, Final Report." University of Dhaka, Dhaka, Bangladesh, and Tufts University, Medford, Mass.

The Bangladesh Integrated Nutrition Project spent some Tk 652 million between 1996 and 2001 from international funds (from the Dutch Government, Canadian International Development Agency, United Nations Children's Foundation, World Bank, and others) and the Bangladesh government to reach 16 percent of the rural population, about 16 million people.

> Although the baseline survey and midterm evaluation were seasonally consistent, the endline survey was, out of necessity, undertaken at a different season of the year. These differences, and the problems encountered in reconciling the data sets, underline the importance of *contracting one organization to conduct all evaluations of such major projects and of means to assure a consistent methodology.* (p. 2, emphasis added)

World Bank. 2003. "A Review of Educational Progress and Reform in the District Primary Education Program (Phases I and II)." Human Development Sector, South Asia Region, Washington, D.C.

The World Bank committed $1.3 billion to the two phases of the District Primary Education Program (DPEP), which aimed at improving primary education in disadvantaged communities. The Dutch government, European Commission, UK Department for International Development, and United Nations Children's Fund also committed large sums to this program over a seven-year period. Ultimately, the program came to serve more than 30 million children.

> The original intent of this report was to evaluate the impact of DPEP I and II based on an exhaustive literature review of the many studies conducted under the aegis of the program. A genuine impact evaluation would assess the magnitude of the change in development objectives of the project that can be *clearly attributed* to the project itself, net of the effect of other programs and external factors. Such an evaluation study would attempt to construct a counterfactual to answer the question, "What would have happened if DPEP had not been implemented?" Typical impact evaluation studies, for programs such as DPEP, which are not nation wide but have partial coverage, and where certain pre-determined criteria were used to select the project districts (i.e. selection was non-random), use statistical methodologies (quasi-experimental or

> non-experimental(to compare project and non-project districts. These statistical techniques attempt to control for other factors that could affect project outcomes. This report is, however, limited to research already done as evident in the literature review. Unfortunately, however, this review revealed that, with the exception of Jalan and Glinskaya, none of the studies could qualify as true impact evaluations. The literature review suggests that DPEP has certainly inculcated a spirit of doing research on primary education, which did not exist in the country prior to the program. However, most studies are limited to studying trends in processes and outcomes in DPEP districts. A few studies do compare DPEP and non-DPEP districts in terms of achievement against outcomes (for example, Agarwal, 2000). However, even these studies are not impact evaluations since they do not statistically control for non-project related factors when comparing outcomes across project and non-project districts. Thus, this report is unable to measure accurately the magnitude of the net impact of DPEP based on this literature review, except to a limited extent for DPEP based on Jalan and Glinskaya. It has thus evolved to become an assessment of the progress made by DPEP I and II in achieving its objective and understanding the successes and limitations of its program of interventions in order to inform future initiatives in educational reform. (p. 9)

Danish International Development Agency. 2004. "Evaluation, Nepal, Joint Government—Donor Evaluation of Basic and Primary Education Programme II." Denmark, Ministry of Foreign Affairs, Copenhagen.

This study evaluated a nationwide primary education program in Nepal that involved more than $150 million, funded primarily by Denmark, the European Union, Finland, Norway, Nepal, and the World Bank, with smaller contributions from the Asian Development Bank, the Japan International Cooperation Agency, and the United Nations Children's Fund.

> Some of the BPEP II activities were launched nation-wide, while others were limited pilots in a few districts. The objective of the pilot projects was to test targeted interventions and new methodologies designed to provide education opportunities for socially disadvantaged groups and girls, as well as develop a sound planning process. Based on the outcomes of the pilot testing, decisions would then be made to expand those that proved cost-effective and relevant for implementation.
>
> The idea of supplementing the core programme with pilot initiatives is a useful strategy, because it provides a platform for supplementing experimental and flexible activities alongside pre-designed core activities.
>
> While it is clear that a wide array of pilot projects was carried out during BPEP II, the evaluation found that it was difficult to obtain a full overview of the number/type of pilot projects initiated. The original plan to have one unit coordinating all pilot activities (BPEDU) was never realised. The absence of effective coordination and firm management of pilots resulted in different implementing agencies/institutions, as well as a range of external agencies, getting involved in the planning, implementing and evaluating of pilot projects, often in a random manner.

> A systematic and standardised monitoring mechanism for pilots does not seem to have been applied. The evaluation had to rely extensively on initial programme documentation (e.g. PIP) and individuals who could recall the history of the BPEP I/II as the primary sources of information on pilot programmes. Due to the high staff turn over in most DOE institutions, however, institutional memory in some cases proved to be limited. (p. 58ff)

Danish Institute for International Studies, Department of Development Research. 2004. "Farm Women in Development Impact Study of Four Training Projects in India: Main Report." Denmark, Ministry of Foreign Affairs, Copenhagen.

> Evaluating the . . . project is much more difficult. Basically, the problem is that there is no proper baseline survey with which the present-day economic situation of the trained farm women and their families can be compared. Scattered impact assessments with limited scope have been carried out by some of the projects, but these are not sufficiently uniform or consistent to be used as a basis for evaluation. The present evaluation provides a comprehensive picture of agricultural activities, income and the overall economic situation, including assets owned, etc. But in terms of changes from year to year, the evaluation has had to rely on information provided by the trained farm women (and in some cases their husbands). This raises the usual questions about reliability as well as the problem of attribution (see below). Thus it has not been possible to quantify precisely the economic benefits to women of their participation in the project.
>
> A special problem in this context is that the interview-based data on yields turned out to be inconclusive. Increased yields are a central intended outcome of the training and extension activities. There is no doubt that, all things being equal, some of the methods and skills taught lead to higher yields. But since widespread drought in some parts of the project areas have had a negative impact on yields, it has not been possible to document this expected positive effect quantitatively. (p. 31)

Note: If a control group had been pre-identified, then a comparison of the decline in yields in the with-project group might have demonstrated that the program had been successful at mitigating the negative impact of the drought. Without a control group, however, the before and after comparison could be used to argue that the program failed.

Farrell, Glen M., ed. 2004. *ICT and Literacy: Who Benefits? Experience from Zambia and India.* Vancouver Commonwealth of Learning.

This three-and-a-half-year project used computers to educate adults in India and Zambia. The president of the foundation that supported the project and its evaluation wrote in the preface that the report's "lessons are highly relevant to the world's ambitious campaign to reduce the scourge of literacy by half in the next decade." (p. viii)

But the report states that it had no evidence on which to measure the program's impact.

> The original plan to collect pre- and post-quantitative data to measure the change in learners' reading, writing and numeracy skills over the course of the project proved impossible for a variety of reasons, such as the delay in getting the project underway, the lack of adequate test instruments, the view that initial testing would "scare off" prospective learners, and the fact that people dropped in and out of the programmes at the centres as other circumstances in their lives dictated. Tests were administered at most of the centres in India as the project ended, and these did provide an indication about learners' skills at that point. There was no end-of-project testing done in Zambia. (pp. 73–74)

Victora, C. G. 1995. "A Systematic Review of UNICEF-Supported Evaluations and Studies, 1992–1993." Evaluation & Research Working Paper Series 3. United Nations Children's Fund, New York.

> In UNICEF, 1,338 reports were completed during 1992 and 1993. This review was restricted to 456 reports available at Headquarters. Of these, a total of 144 reports were selected for the final review: all 44 reports classified in the database as dealing with impact and a random sample of 100 reports (50 studies and 50 evaluations) out of 412 classified as not dealing with impact. (p. v)
>
> The reviewers found that only 20% of reports classified as impact evaluations truly were and that 14% of reports categorized in the 'non-impact' category were in fact impact evaluations. From these results one may assume that 15% of all reports and 35% of all evaluations dealt with impact.
>
> The reviewers felt that 91% of the non-impact evaluations and 31% of the studies had relevant findings for possible reformulation of UNICEF-supported projects or programmes, again a positive finding. Other reports were judged to be relevant for other purposes, such as advocacy. Some 10% of all reports were deemed to be worthless. Over one third (37%) of all reports—including the 10% mentioned above—were judged to be unjustified in terms of costs relative to objectives and actual outcomes. (p. vi)
>
> Based on the data in Table 4, it is possible to estimate that 15% of all reports in the database (and 35% of all evaluations) include impact assessments. The reviewers also noted that, in about 25 reports, the authors had attempted impact evaluations but did not succeed, particularly due to methodological shortcomings.
>
> Only one in five impact evaluations had been correctly classified. Six out of seven non-impact reports were properly classified. By extrapolating these findings to the database as of March1993, one may estimate that 35% of all evaluations, or 15% of all reports, included impact assessments. Many evaluations were unable to properly assess impact because of methodological shortcomings. (p. 10)

> Six in seven studies or evaluations used quantitative approaches. However, most of these employed quantitative data to provide useful qualitative insights. (p. 13)
>
> In almost 60% of the reports, the findings were clearly linked to the objectives and methods. However, in 18% of the reports this linkage was unsatisfactory. Common flaws included that the described methodology could not have produced the findings being reported and that no data were presented relevant to some of the stated objectives (mainly those on assessing impact). (p. 16)

Bellamy, M. 2000. "Approaches to Impact Evaluation (Assessment) in Agricultural Information Management: Selective Review of the Issues, the Relevant Literature and Some Illustrative Case Studies." CTA Working Document 8021. Technical Centre for Agricultural and Rural Cooperation, Wageningen, Netherlands.

> *While there is a burgeoning literature on theory and methodology, it is more difficult to find examples of impact studies in real situations or applied to real projects.* There are some examples of post-hoc evaluations, but few of impact studies, and even fewer where the methodology and process have been set up in advance.
>
> Reference has already been made to studies undertaken by CABI and CTA to evaluate their own information delivery projects. These were essentially evaluation rather than impact studies, designed to improve the services rather than specifically measure impact. (p. 16)

Bernard, A. 2002 "Lessons and Implications from Girls' Education Activities: A Synthesis from Evaluations." United Nations Children's Fund, Evaluation Office, New York.

> Finally, the scope of the synthesis is limited in that *levels of analysis in the evaluations overall are not especially strong*. Most concentrate more on inputs (what projects delivered and the activities they undertook), than on results (the changes which were realised as a consequence of those inputs). Also, only a few explore the factors influencing project implementation, or the implications of these factors for the continued validity of the assumptions guiding the projects. In consequence, while the evaluations provide valuable detail on what is or was happening from the perspective of project delivery, they are rather less rich in terms of the 'value-added' of those actions in making a difference to the situation of girls' education more widely. (p. 26)

Buchan, J., and M. R. Dal Poz. 2002. "Skill Mix in the Health Care Workforce: Reviewing the Evidence." *Bulletin of the World Health Organization* 80 (7): 575–80.

> There are significant limitations to the current evidence on skill mix in the health workforce. Many published studies in this area are merely

> descriptive accounts, which add little in terms of use of methods or interpretation of results. Where studies do move beyond description, their usefulness is often constrained by methodological weaknesses, lack of appropriate evaluations of quality/outcome and cost, and/or use of small sample sizes. Moreover, many of the studies were undertaken in the USA, and the findings may not be relevant to other health systems and countries. The results may therefore be suspect, and of little use for comparative purposes or in drawing general conclusions. (p. 578)

Appendix F. Results from the consultation survey

In an online and email survey disseminated between September 2005 and January 2006, the Center for Global Development (CGD) asked about respondents' experiences using and conducting impact evaluations. The purpose of the survey was to test some of our hypotheses about the problem and to solicit reactions to our ideas and recommendations.

Results from the survey's 61 respondents—workers in research institutes, nongovernmental organizations, international agencies, and governments—are summarized in the following analysis. Many respondents reside in Japan, North America, or Western Europe, though some reside in Africa, Asia, Eastern Europe, or Latin America. The majority of respondents work in Africa, though all other regions were well represented.

Knowledge of existing impact evaluations

Respondents were familiar with social development programs with impact evaluations, but were most familiar with those done in microcredit and conditional cash transfer programs.

Respondents identified the following impact evaluations as high quality:

- IMCI Joint Donor Evaluation of Rwandan Refugee Crisis (2005)
- US Self-Employment Program (experimental design)
- Evaluations of Conditional Cash Transfer programs in Latin America
- Stifel and Alderman's evaluation of a feeding program in Peru (2003)
- Nutrition programs in Argentina
- Morris, S.S., R. Flores, P. Olinto, and J. M. Medina. 2004. "Monetary Incentives in Primary Healthcare and Effects on Utilization and Coverage of Preventive Healthcare Interventions in Rural Honduras: Cluster-Randomized Trial." *The Lancet* 364 (9450): 2030–37.
- Esther Duflo's randomized evaluations
- MkKelly and Lippold's "Microfinance: USAID/AIMS Impact Assessment in Mali."
- AusAID's Health Services Support Program in Papua New Guinea
- Cognitive and biomedical impact measurements on children receiving deworming medication
- OED evaluation of the Bangladesh Integrated Nutrition Project

Knowledge of existing resources on social development policy

Of the current initiatives dedicated to improving social development, some respondents were familiar with Development Gateway and Measure DHS. However, most were unfamiliar with:

- Development Assistance Committee Evaluation Network
- Development Impact Evaluation Initiative (DIME)
- ID-21 (Institute for Development Studies)
- Campbell Collaboration
- Cochrane Collaboration
- Health Metrics Network

Respondents indicated additional initiatives that aim to improve learning about social development policy:

- UN Research Institute for Social Development's (UNRISD) program on Social Policy in a Development Context [www.unrisd.org/unrisd/website/projects.nsf/0/9DBC873B99D850E180256B4F005D6460?OpenDocument]
- Development Ethics [www.development-ethics.org]
- Eldis [www.eldis.org/]
- United Nations Development Programme (UNDP) [www.undp.org/]
- Active Learning Network for Accountability and Performance (ALNAP) [www.alnap.org/]
- The Poverty Action lab at MIT [www.povertyactionlab.com]
- Center for International Development (CID) [www.cid.harvard.edu/]
- Institute of Development Studies University of Sussex [www.ids.ac.uk/ids/]
- Groupe Initiatives [www.groupe-initiatives.org/uk/default.htm]
- Development in Practice (Oxfam group) [www.developmentinpractice.org/]
- CIDA's Capacity Development Extranet [http://web.acdi-cida.gc.ca/cd]
- Enterprise Development Impact Assessment Information Service [www.enterprise-impact.org.uk/index.shtml]
- Microfinance Gateway Impact Assessment Centre [www.microfinancegateway.org/section/resourcecenters/impactassessment/]
- Imp-Act @ IDS [www.ids.ac.uk/impact/index.html]
- USAID Poverty Assessment Tools [www.povertytools.org/Project_Documents/project.htm]
- Institute for Fiscal Studies at University College, London [www.ifs.org.uk/index.php]
- USAID Initiative to Revitalize Evaluations [www.dec.org/partners/evalweb/]
- Developing country governments
- National evaluation associations in developing countries
- USAID Evaluation Database [www.dec.org/]
- Eurodad [www.eurodad.org]
- World Bank Live Database in Africa [http://www4.worldbank.org/afr/stats/ldb.cfm]
- World Bank Living Standards Measurement Study [www.worldbank.org/lsms/]

- MandE News [www.mande.co.uk/]
- Overseas Development Institute [www.odi.org.uk/]
- InterAction Evaluation and Program Effectiveness Working Group [www.interaction.org/evaluation/]

Enduring questions to be answered by impact evaluations

Respondents converged around the following major enduring questions that challenge social policy in developing countries:

Who is responsible for implementing social policy, and who is most effective?

- Role of government vs. civil society.
- Public vs. private enterprise roles in social change and social policy.
- What are the appropriate roles of international aid institutions?
- Impact of local religious organizations on improvements in health, education, and poverty.

How does promoting local ownership over the development process improve:

- Mitigating the unexpected consequences of social development efforts?
- Consideration of historical context within a country when creating new social policy?
- Harmonizing aid supply with demand?

Are the poorest of the poor targeted in development programs?

How to improve the sustainability of development programs?

What works in development?

- Poverty alleviation efforts vs. social development programs.
- Safety net/social security approach of European donors vs. funding realities and capacities of host governments.
- Are successes replicable?

Concerns about local governments:

- What is the impact of decentralization?
- How to eliminate corruption in the political process?

Sector-specific concerns:

- How to improve the quality of social services (education, health, etc)?
- How to effectively deliver health services?
- How to define the right problem in education?

Suggested literature reviews on enduring questions:

Estrella, M., and J. Gaventa. 1998. "Who Counts Reality? Participatory Monitoring and Evaluation: A Literature Review." IDS Working Paper 70. Institute of Development Studies, Brighton, UK. [www.ids.ac.uk/ids/bookshop/wp/Wp70.pdf].

Carlsson, J., and L. Wohlgemuth, eds. 2002. "Learning in Development Cooperation." Swede, Expert Group on Development Issues, Stockholm. [www.egdi.gov.se/pdf/20002pdf/2000_2.pdf].

Oxfam. 2005."Programme Impact Report: Oxfam GB's Work with Partners and Allies around the World." London. [www.eldis.org/cf/rdr/rdr.cfm?doc=DOC20049].

Tendler, J. 1997. *Good Government in the Tropics*. Baltimore, Md.: Johns Hopkins University Press.

Missed opportunities for impact evaluation

The following are a few anecdotes among many from respondents on missed opportunities to implement impact evaluations that may have yielded beneficial knowledge.

> Sustainability of immunization programs following polio eradication in Latin America. I have a protocol, have applied for funds to do such a study but cannot locate funding. Donors are either amnesic or funding is too tight.
>
> Bangladesh, the impact of government policy and NGO's, notably BRAC, on the notable increases in enrollment, and the erasure of urban/rural [gaps] as well as inequality. Uganda and Malawi, on how removing school fees impacted enrollment, retention and school learning. I think the opportunities were lost because most of the benefits of learning the impacts would accrue to other countries.
>
> In nearly all of our work in conflict and post-conflict countries, we have used 'urgency' to justify not devoting resources to collecting the necessary information to do an impact analysis, leaving us to start over in the next conflict situation, with only anecdotal ideas of what works.
>
> I know of impact evaluations that took place, but were then filed away and the learning was never embedded, which can be worse. This is often because the results of the impact evaluations threaten the interests of powerful people and sections of the community.

Suggestions for avoiding missed opportunities in the future include:

- "Specific instructions from the granting authority that an impact evaluation needed to be proposed and implemented as part of the grant award. Clients are reluctant to spend borrowed funds on evaluations. They feel that the money should be used for program implementation. Moreover, the results of the impact evaluation benefit other countries and, therefore, the cost of evaluations should be spread out among many beneficiaries."
- "Improving evaluation expertise of program designers."
- "Communicate demand for use of impact knowledge by developing countries."
- "More funds earmarked for international health impact evaluations."
- "Change the organizational/bureaucratic/career incentives in favor of doing more and more rigorous impact evaluations."
- "Seeing impact evaluations not as a cost bust as a high-return investment for development."
- "Political commitment to counteract entrenched interests."
- "More conscious, up-front planning of the impact evaluation."
- "Learning from success in other countries."
- "It should be the policy of donor organizations that appropriate mechanisms be included in projects by implementing partners and that

resources be made available for rigorous, independent impact analysis. Development agencies are constantly reinventing the wheel because we rarely know exactly what works and doesn't work and much more importantly, why."
- "If donors started to actually pay for performance and not for project expenses, project metrics would be part of everyday activities."
- "Creating a platform where evaluators might safely post the results of evaluations, i.e. others could use the information, but the evaluator would not be penalized for negative results."

Solutions for improving impact evaluations

Respondents indicated that the following actions are important to improving impact evaluations:
- "Increased pressure on governments and agencies to measure results."
- "Training for evaluators."
- "Increased exchange of existing information."
- "Timely funds."

Respondents added several other ways to promote more and better impact evaluations:
- "Less ideologically driven agendas."
- "Consistency with what results mean."
- "Appropriate career incentives for doing impact evaluations."
- "More effective aid as in OECD Paris Declaration, starting with government ownership and alignment, harmonization and management for results."
- "Continuity of the study teams."
- "Greater desire of governments to get evidence on impact."
- "Creating incentives within projects and programs to engage impact evaluation."
- "Better understanding of what motivates people to plan based on evidence, or what motivates them to use 'alternative' sources of information."
- "A 'feedback culture.'"
- "Not just to measure results but analyze and publicize them."
- "The involvement of development project participants in the impact evaluation, particularly in defining the indicators of success."
- "'Safe' and transparent platforms for posting information."
- "Time afforded to carry [out] impact evaluations."

Most respondents thought the following actions were very important to improving learning about social development (in order of importance):
- "Encouraging developing countries to conduct evaluations."
- "Coordinating research among agencies."
- "Creating a new fund for impact evaluation."

Participants proposed alternative ways to improve learning about social development:
- "Changing career incentives in major development institutions to favor impact evaluations."

- "Making sure that multilateral development banks provide evaluations as a free good in loans."
- "Disseminating information and sharing research among stakeholders and promoting feedback."
- "Increasing pressure on governments and agencies to measure results and to be honest, transparent, and accountable."
- "Increasing media coverage of development issues."
- "Encouraging timelines that are amenable to impact evaluations."

Appendix G. Ideas for setting impact evaluation quality standards

The following examples show how organizations and authors have sought to establish standards of evidence in relation to different forms of impact evaluation. One of the most critical tasks for the proposed council would be to review these approaches to standard setting, facilitate a debate among stakeholders, and endorse or develop the standards that the council would apply. Setting standards should not be viewed as a one-shot task. Rather, it should be viewed as a process of adopting, testing, assessing, and modifying standards in light of members' experiences and methodological progress.

Example 1: From D. Levine (2005)

What is a "rigorous" evaluation?

No recipe will ensure that an evaluation produces the correct result. Nevertheless, rigorous, credible, and transparent evaluations are more likely to be valid, useful, and used. Such evaluations typically meet the following standards:

Process

- Rigorous evaluation almost always requires integration into a project design. At a minimum, it is usually important to identify the baseline status of those groups receiving the program and their comparison (or "control") groups. Even better, early integration of program evaluation and design often permits the rollout of a program to be randomized, leading to more convincing results.
- The evaluation is supervised by an independent research team.
- Evaluation designs undergo a peer review by experts in evaluation (preferably prior to decisions about funding).
- All evaluation results are disseminated publicly.
  - This requirement avoids the problem of selectively publishing favorable evaluations.
- To reduce the costs of other evaluations, the survey instrument is made available on the Web as soon as possible (in the lan-

guage of the country where the program is implemented and also preferably also in English).
- To permit re-analysis and increase transparency, the data are available on the Web in a timely fashion (consistent with privacy considerations).

Substance
- Because randomized trials are more convincing than other research methods, most evaluations should involve randomization. When randomized trials are not feasible, examining changes over time between carefully matched comparison groups or other rigorous study designs can be reasonable alternatives.
  - Simple before-and-after comparisons are not sufficient to count as "rigorous," because they do not show whether the program itself was responsible for any changes.
- Proposals for a rigorous evaluation include calculations showing the study is of large enough size to identify with high probability effects large enough to matter for policy purposes ("power calculation").
- Because some impacts take years to appear, some rigorous studies need long lives relative to most program funding cycles. When spillovers are potentially important (e.g., when studying infectious diseases), evaluations of community-level effects are included.
- To be fully convincing, evaluations provide evidence that the purported causal channels were affected by the intervention (Victora, Habicht, and Bryce 2004). For example, it is important to determine whether a nutrition education program led to faster child growth; at the same time, the findings are far more convincing if the evaluation study also documents changes in food purchases. Studies look for unintended consequences and study the process of implementation.

How do rigorous impact evaluations relate to other forms of evaluation?

Most aid organizations, like many other organizations, engage in a variety of forms of evaluations. For example, operations evaluations create qualitative feedback of how programs are operating. Such evaluations are important in their own right, and also important complements to most rigorous impact evaluations (Victora, Habicht, and Bryce 2004).

Evaluations for continuous improvement provide rapid feedback about new innovations, permitting decision makers to spot problems and try out solutions. Such feedback can often be integrated into rigorous impact evaluations in two fashions. For example, the data from the treatment sites can be fed back to decision makers to improve the program. The impact evaluation, then, is not of the prototype program as originally designed, but of the combination of the prototype program plus the improvements made along the way as information arrived.

Example 2: From Policy Hub, Government Social Research Unit, HM Treasury, United Kingdom

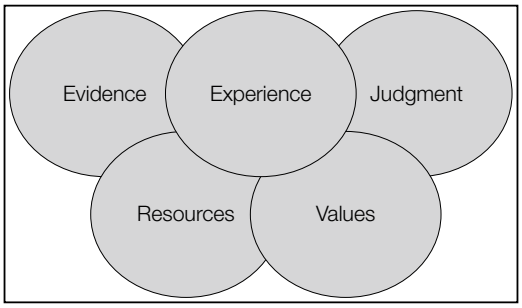
See www.policyhub.gov.uk/evaluating_policy/how_res_eval_evid.asp#systematic.

How research and evaluation evidence contributes to policymaking

Introduction	Experimental and quasi-experimental evidence	Survey and administrative evidence
Qualitative research evidence	Economic evaluation evidence	Philosophical and ethical evidence
Systematic evidence		Back to evaluating policy

Introduction

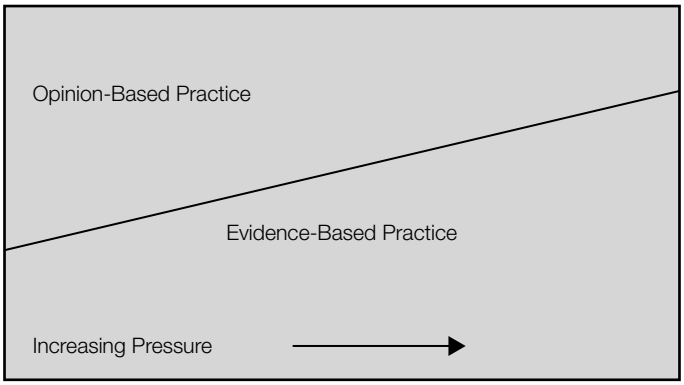
Evidence is one factor that contributes to policy making, implementation and delivery. The following diagram indicates other important factors:



From opinion-based policy to evidence-based policy

Evidence-based policy has been defined as 'the integration of experience, judgement and expertise with the best available external evidence from systematic research' (Davies, 1999). This involves a balance between professional judgement and expertise on the one hand and the use of valid, reliable and relevant research evidence on the other.

Gray (1997) has suggested that evidence-based policy and practice involves a shift away from opinion-based decision making to evidence-decision making.



(Source, Gray, J.A.M., 1997)

Evidence-based decision making draws heavily upon the findings of scientific research (including social scientific research) that has been gathered and critically appraised according to explicit and sound principles of scientific

inquiry. The opinions and judgements of experts that are based upon up-to-date scientific research clearly constitute high quality valid and reliable evidence. Those opinions that are not based upon such scientific evidence, but are unsubstantiated, subjective and opinionated viewpoints do not constitute high quality, valid and reliable evidence.

Different types of evidence are generated by different types of research methods and research designs.

Experimental and quasi-experimental evidence

Experimental and quasi-experimental evidence is generated by research methods such as:

- Randomised controlled trials (including area-based, cluster randomised trials)
- Controlled before-and-after studies
- Interrupted time series studies
- Various types of matched comparison studies (such as difference-of-differences, and propensity score matching)

These methods provide valid and reliable evidence about the relative effectiveness of a policy intervention compared with other policy interventions, or doing nothing at all (sometimes called the counterfactual). They provide appropriate evidence about questions such as: is a personal adviser service more, or less, effective than providing skills training, or doing nothing at all, in terms of advancing low paid people in the labour market?

Survey and administrative evidence

Evidence from social surveys and administrative data are often used in experimental and quasi-experimental studies. However, they also provide valuable information in their own right about the nature, size, frequency, and distribution of a problem or a topic under investigation. The General Household Survey, for instance, provides a wealth of evidence about income, employment, housing, lifestyles, health and illness and many other aspects of everyday living based upon a sample of households throughout the UK Administrative data, such as the New Deal Database collected and maintained by the Department of Work and Pensions, provides valuable evidence about people's entry to, exit from, and participation in the labour market. Other important sources include 'raw' administrative data sets such as Housing Benefits Claimants, reported crime, claimant counts, and applications as homeless people.

Qualitative research evidence

Evidence is also often required about why a policy works (or fails to work), how it works, for whom, and under what conditions it works or fails to work. This involves eliciting evidence of the opinions, attitudes and perceptions of different stakeholders in the policy process. Such evidence is particularly important for the successful implementation and delivery of policies, especially across a range of populations and sub-groups. The following qualitative research methods provide such evidence:

- Theory-based methods (including theories or change, programme theory evaluation, 'realistic' evaluation methods)

- Goals-based evaluation methods
- Goals-free evaluation methods
- In-depth interviews
- Focus groups (including stakeholder analysis)
- Consultative techniques (including user satisfaction studies)
- Ethnographies
- Observational and participant-observational studies
- Conversation and discourse analysis

Economic evaluation evidence

Policy making, implementation and delivery, inevitably involves decisions about the use and allocation of scarce resources. Consequently, evidence is required about the most cost-effective way of achieving a given objective, and about the greatest benefit and utility that can be achieved from the available resources. Such evidence is provided by economic evaluation methods which include:

- Cost-effectiveness analysis
- Cost-benefit analysis
- Cost utility analysis
- Opportunity cost appraisal
- Deadweight and counterfactual appraisal

Philosophical and ethical evidence

Policy making takes place against a background of values, including beliefs, ideologies and aspirations. Consequently, evidence is often required about the range of values involved in a policy decision or initiative, and about ways of adjudicating between competing values. Such evidence requires the methods of political and moral philosophy and ethics. These include:

- Consultative techniques
- Needs analysis
- Critical incidence analysis

Systematic review evidence

Evidence from single studies have the limitations of being sample specific and often time- and context-specific. Also, not all research studies are carried out to the highest standards of validity, reliability, analysis and presentation. Consequently, not all single studies are of equal value and some can be statistically and scientifically biased.

Systematic reviews of research literature attempt deal with these problems by establishing standards of inclusion and exclusion of single studies, separating high quality from low quality research evidence, and providing syntheses of what the high quality evidence is telling us about a topic or policy area. There are different types of review evidence including:

- Narrative Reviews
- Vote Counting Review
- Systematic Reviews
- Meta-Analyses
- Best-Evidence Synthesis

- Meta-Qualitative Reviews
- Rapid Evidence Assessments

These are some of the types of evidence used in evidence-based policy making, implementation and delivery. More detailed information about these different types of evidence are found in the Magenta Book. These types of evidence are not exhaustive and this page will be updated as other types of evidence used in policy making are identified.

The Government Social Research Unit provides seminars and training in most of the types of evidence presented above.

References

Davies, P. T. 1999. 'What is Evidence-Based Education?', British Journal of Educational Studies, 47, 2, 108–121.

Gray, J. A. M. 1997. Evidence-Based Health Care: How to Make Health Policy and Management Decisions, New York and London, Churchill Livingstone, 1997.

Example 3: From US Department of Education's Institute of Education Sciences (IES) "What Works Clearinghouse"

See http://w-w-c.org/reviewprocess/study_standards_final.pdf.

WWC study review standards

The What Works Clearinghouse (WWC) reviews studies in three stages. First, the WWC screens studies to determine whether they meet criteria for inclusion within the review activities for a particular topic area. The WWC screens studies for relevance on the following dimensions: (a) the relevance of the intervention of interest, (b) the relevance of the sample to the population of interest and the recency of the study, and (c) the relevance and validity of the outcome measure. Studies that do not meet one or more of these screens are identified as "Does Not Meet Evidence Screens."

Second, the WWC determines whether the study provides strong evidence of causal validity ("Meets Evidence Standards"), weaker evidence of causal validity ("Meets Evidence Standards with Reservations"), or insufficient evidence of causal validity ("Does Not Meet Evidence Screens"). Studies that "Meet Evidence Standards" include randomized controlled trials that do not have problems with randomization, attrition, or disruption, and regression discontinuity designs without attrition or disruption problems. Studies that "Meet Evidence Standards with Reservations" include quasi-experiments with equivalent groups and no attrition or disruption problems, as well as randomized controlled trials with randomization, attrition, or disruption problems and regression discontinuity designs with attrition or disruption problems.

Third, all studies that meet the criteria for inclusion and provide some evidence of causal validity are reviewed further to describe other important characteristics. These other characteristics include: (a) intervention fidelity; (b) outcome measures; (c) the extent to which relevant people, settings, and measure timings are included in the study; (d) the extent to which the study allowed for testing of the intervention's effect within subgroups; (e)

statistical analysis; and (f) statistical reporting. This information does not affect the overall rating.

Studies that "Meet Evidence Standards" and "Meet Evidence Standards with Reservations" are summarized in WWC Reports. The WWC produces three levels of reports: study reports, intervention reports, and topic reports. WWC Study Reports are intended to support educational decisions by providing information about the effects of educational interventions (programs, products, practices, or policies). However, WWC Study Reports are not intended to be used alone as a basis for making decisions because (1) few, if any, studies are designed and implemented flawlessly and (2) all studies are tested on a limited number of participants and settings, using a limited number of outcomes, at a limited number of times. Therefore, generalizations from one study should, in most cases, not be made.

The WWC Study Reports focus primarily on studies that provide the best evidence of effects (e.g., primarily randomized controlled trials and regression discontinuity designs and, secondarily, quasi-experimental designs) and describe in detail the specific characteristics of each study. The WWC also conducts systematic reviews of multiple studies on one specific intervention and summarizes the evidence from all studies in the intervention reports.

Finally, the WWC summarizes the evidence of all interventions for a topic in the topic report. Neither the What Works Clearinghouse (WWC) nor the U.S. Department of Education endorses any interventions.

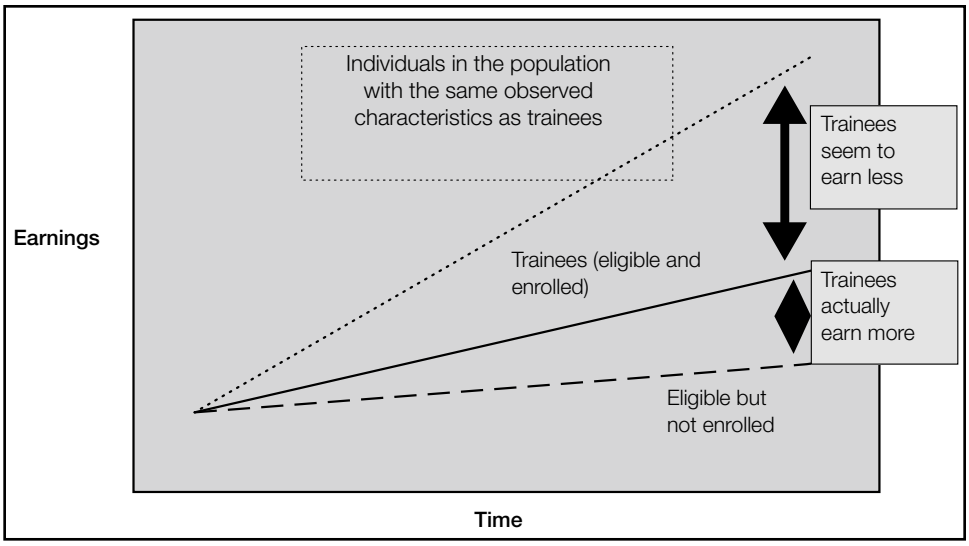
Appendix H. Advantages and limitations of random-assignment studies

The key advantage of random-assignment studies is that they can effectively reduce unobserved bias, giving greater confidence that the measured impact of a program is attributable to that program and not to some other factor. In general, methods that do not use random assignment can account only for systematic biases that are related to observable differences between treatment and control groups. The exceptions are studies that take advantage of "natural" experiments, such as those using regression discontinuity.

Consider job-training programs financed by the United States in the 1970s and 1980s. Several studies compared the earnings of job trainees with those of individuals in the general population who had similar characteristics but did not participate in the job-training program. The results were disappointing, finding that job trainees earned less than others with similar characteristics. What these studies could not address is that the individuals who entered the public job-training programs had already failed to find work—some unobserved factors accounted both for their need for the program and for their subsequently poorer earnings.

A prospective study that randomly assigned which applicants could participate in the program, however, yielded opposite results, finding that job

Figure H.1 Impact evaluation using comparisons based on random assignment



training led to improved earnings. The random-assignment study reached the correct conclusion because it compared only individuals eligible for the program, thereby eliminating the unobserved differences that skewed the other studies (figure H.1).

The US government commissioned a review of youth employment and training programs to determine what had been learned. The experts who participated in that commission found:

> Our review of the research on YEDPA [Youth Employment and Demonstration Programs Act] shows dramatically that control groups created by random assignment yield research findings about employment and training programs that are far less biased than results based on any other method. . . . The fact that some studies successfully used random assignment suggests that this procedure is feasible and presents no serious technical difficulties in execution. It is evident that if random assignment had consistently been used in YEDPA research, much more would have been learned (Betsey, Hollister, and Papgeorgiou 1985, p. 18).

Random-assignment studies are not a panacea, and they cannot substitute for other approaches in every situation. Random-assignment studies should be used only when the questions being addressed are amenable to this research approach. In addition, researchers must follow well established standards for assuring that assignment is random and that results are not biased by attrition. The results of a random-assignment study, as with most other methods, cannot be generalized without considering differences across contexts; consequently, such studies need to be replicated in different places to accumulate a broader body of knowledge about the intervention. Finally, the use of random-assignment studies to analyze complex social policies is still developing. Researchers need to pay careful attention to the underlying mechanisms and models of behavior being tested to be sure that the method is applied where it is appropriate.

A few well conducted random-assignment studies can complement other kinds of research in useful ways. The attention to sampling and data collection for a random-assignment study can generate data that is useful for analysis by other methods (Todd and Wolpin 2004). Random-assignment studies can also help to assess the reliability of evidence from other research. For example, the evaluation of Integrated Management of Childhood Illness included a range of process and operational evaluations along with a random-assignment study in Bangladesh (Bryce and others 2004).

Comparing studies that use different methods can boost confidence in the reliability of findings and provide insights into which methods are more or less effective at reaching valid conclusions. Reviews comparing studies with and without random assignment have now been conducted on many topics—the impact of neighborhood poverty on individuals (Liebman, Katz, and Kling 2004); the effectiveness of training programs (Fraker and Mayard 1984; Lalonde 1986; Friedlander and Robins 1995; Friedlander, Greenberg, and Robins 1997); social welfare policy in Sweden (Bratberg, Grasdal, and Risa and others 2002); welfare, job training, and employment services programs (Glazerman and Levy 2003); conditional cash transfer programs (Diaz and Handa 2004); and improving test scores and reducing dropout rates (Wilde and Hollister 2002; Agodini and Dynarski 2004). This literature is reviewed in Glazerman, Levy, and Myers (2002), who show where nonrandomized approaches appear to generate valid conclusions and where they can fail.

To improve our knowledge of what works in social development programs, more investment is needed in rigorous impact evaluations. The appropriate methods for such studies need to reflect the particular policy question being asked and the context in which the program is implemented. Random-assignment approaches have been demonstrated to be a feasible and rigorous approach to impact evaluation in many situations and should therefore be encouraged and promoted where appropriate.

Notes

1. Use of the word *council* is motivated by a consultation in Mexico City, at which Mexico's efforts to improve social program evaluation through a new National Council for Social Policy Research were presented. It is, however, only one possible term for whatever entity might be created.
2. The Center for Global Development has no interest or intention of assuming responsibility for such a council.
3. It is possible to classify and analyze evaluations in different ways. This discussion draws from Jacquet (forthcoming), but other examples from this large literature include Rossi, Freeman, and Lipsey (1999); Habicht, Victora, and Vaughan (1999); Altman and others (2001); and DAC (2002).
4. A public good is a product or service that can be used by many people without being depleted (it is nonrivalrous in consumption) and whose benefits cannot be restricted to a particular individual or group (nonexcludable). The classic example is a lighthouse. Boats that benefit from seeing a warning beacon do not, thereby, reduce its benefit to anyone else nor, ostensibly, can they be excluded from its benefits. For reasons why a lighthouse may not be a good example, however, see Coase (1974).
5. For a discussion of how self-selection can bias results in comparing schools, see Cullen, Jacob, and Levitt (2005).
6. Discussion with Stephen Quick, Director, Office of Evaluation, Inter-American Development Bank, Washington, D.C., September 2005.
7. For example, Seva Mandir in India, Internationaal Christelijk Steunfonds in Kenya, and Freedom from Hunger in Ghana and Bolivia.
8. Similarly, a review of 10 randomized control studies on the policy of grouping students by skill level showed that this approach has little or no effect on student achievement (Mosteller, Light, and Sachs 1996).
9. Discussion with Santiago Levy, former Undersecretary of Finance, Mexico, February 9, 2006.
10. MDRC, formerly known as the Manpower Demonstration Research Corporation, is a private nonprofit organization established in 1974 with support from the Ford Foundation and six US government agencies to assess welfare, training, and education programs.
11. For a complementary discussion of the obstacles to good impact evaluations, see Levine (2005).
12. Discussion with Suzanne Duryea, Senior Research Economist, Inter-American Development Bank, Washington, D.C., August 2005. For further information on Familias en Acción, see www.ifs.org.uk/edepo/wps/familias_accion.pdf.
13. Use of the word *council* is motivated by discussions in Mexico City, at which Mexico's efforts to improve social program evaluation through the new National Council for Social Policy Research were presented. It is, however, only one possible term for whatever entity might be created.

14. This section benefited particularly from ideas provided by Dan Kress, Blair Sachs, and Smita Singh.

15. The Center for Global Development has no interest or intention of assuming responsibility for such a council. For a schematic presentation of different institutional arrangements, along with their advantages and disadvantages, see Bill & Melinda Gates Foundation (2002).

16. This is true unless, or until, the council's functions reach a scale that justifies its own independent administrative capacities.

References

Agodini, R., and M. Dynarski. 2004. "Are Experiments the Only Option? A Look at Dropout Prevention Programs." *The Review of Economics and Statistics* 86 (1): 180–94.

Altman, D. G., K. F. Shulz, D. Moher, M. Egger, F. Davidoff, D. Elbourne, P. C. Gotzsche, and T. Lang. 2001. "The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration." *Annals of Internal Medicine* 134 (8): 663–94.

Angrist, J., and V. Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *Quarterly Journal of Economics* 114 (2): 533–75.

Banerjee, A. V., and R. He. 2003. "The World Bank of the Future." BREAD Working Paper 13. Harvard University, Bureau for Research and Economic Analysis of Development, Cambridge, Mass.

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden. 2003. "Remedying Education: Evidence from Two Randomized Experiments in India." Massachusetts Institute of Technology, Cambridge, Mass.

Banerjee, A. V., S. Jacob, and M. Kremer (with J. Lanjouw and P. Lanjouw). 2001. "Promoting School Participation in Rural Rajasthan: Results from Some Prospective Trials." Harvard and Massachusetts Institute of Technology, Cambridge, Mass.

Betsey, C. L., R. G. Hollister, and M. R. Papgeorgiou, eds. 1985. *Youth Employment and Training Programs: The YEDPA Years*. Washington, D.C.: National Academy Press.

Bill & Melinda Gates Foundation. 2002. "Developing Successful Global Health Alliances." Global Health, Seattle, Wash. [www.gatesfoundation.org/nr/downloads/globalhealth/].

Bratberg, E., A. Grasdal, and A. E. Risa. 2002. "Evaluating Social Policy by Experimental and Nonexperimental Methods." *Scandinavian Journal of Economics* 104 (1): 147–71.

Bryce, J., C. G. Victora, J.-P. Habicht, J. P. Vaughan, and R. E. Black. 2004. "The Multi-Country Evaluation of the Integrated Management of Childhood Illness Strategy: Lessons for the Evaluation of Public Health Interventions." *American Journal of Public Health* 94 (3): 406–15.

Buddlemeyer, H., and E. Skofias. 2003. "An Evaluation on the Performance of Regression Discontinuity Design on PROGRESA." IZA Discussion Paper 827. Institute for Study of Labor, Bonn, Germany

Campbell Collaboration. 2005a. "Methods of Synthesis Project." *Journal of Health Services Research and Policy* 10 (3).

———. 2005b. "Synthesizing Evidence for Management and Policy-making." *Journal of Health Services Research and Policy* 10 (3).

Chattopadhyay, R., and E. Duflo. 2001. *Women as Policy Makers: Evidence from a India-Wide Randomized Policy Experiment*. NBER Working Paper 8615. Cambridge, Mass.: National Bureau of Economic Research.

Coase, R. H. 1974. "The Lighthouse in Economics." *Journal of Law & Economics* 17 (2): 357–76.

Conley, T., and C. Udry. 2000. "Learning about a New Technology: Pineapple in Ghana." Discussion Paper 817. Yale University, Economic Growth Center, New Haven, Conn.

Cullen, J. B., B. Jacob, and S. Levitt. 2005. "The Impact of School Choice on Student Outcomes: An Analysis of the Chicago Public Schools." *Journal of Public Economics* 89 (5–6): 729–60.

DAC (Development Assistance Committee). 1991. "Principles for Evaluation of Development Assistance." Organisation for Economic Co-operation and Development, Paris.

———. 1998. "Review of the DAC Principles for Evaluation of Development Assistance." Organisation for Economic Co-operation and Development, Paris.

———. 2002. "Glossary of Key Terms in Evaluation and Results Based Management." Organisation for Economic Co-operation and Development, Paris.

DFID (UK Department for International Development). 2002. "DFID Public Service Agreement (PSA) 2003–2006." London.

Diaz, J. J. and S. Handa, 2004, "An Assessment of Propensity Score Matching as a Nonexperimental Impact Estimator: Evidence from a Mexican Poverty Program." University of North Carolina, Chapel Hill.

Dickersin, K., and Y. I. Min. 1993. "Publication Bias: The Problem that Won't Go Away." *Annals of the New York Academy of Sciences* 703 (1): 135–46.

Duflo, E. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813.

———. 2004. "Scaling Up and Evaluation." Paper prepared for the Annual World Bank Conference on Development Economics, May 20, Bangalore, India.

Duflo, E., and M. Kremer. 2003. "Use of Randomization in the Evaluation of Development Effectiveness." Paper presented at the World Bank Operations Evaluation Department Conference on Evaluation and Development Effectiveness, July 15, Washington, D.C.

Dugger, C. 2004. "World Bank Challenged: Are the Poor Really Helped?" *The New York Times*. July 28.

Ekman, B. 2004. "Community-Based Health Insurance in Low-Income Countries: A Systematic Review of the Evidence." *Health Policy and Planning* 19 (5): 249–70.

Fraker, T. and R. Mayard. 1984. *An Assessment of Alternative Comparison Group Methodologies for Evaluating Employment and Training Programs.* Princeton, N.J.: Mathematica Policy Research, Inc.

Friedlander, D., and P. K. Robins. 1995. "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods." *American Economic Review* 85 (4): 923–37.

Friedlander, D., D. H. Greenberg, and P. K. Robins. 1997. "Evaluating Government Training Programs for the Economically Disadvantaged." *Journal of Economic Literature* 35 (4): 1809–55.

Gertler, P. 2000. "Final Report: The Impact of PROGRESA on Health." International Food Policy Research Institute, Washington, D.C.

Glazerman, S., and D. M. Levy. 2003. "Nonexperimental Versus Experimental Estimates of Earnings Impacts." *The Annals of the American Academy of Political and Social Science* 589 (1): 63–93.

Glazerman, S., D. M. Levy, and D. Myers. 2002. "Nonexperimental Replications of Social Experiments: A Systematic Review." Interim Report/Discussion Paper 8813-300. Mathematica Policy Research, Princeton, N.J.

Glewwe, P., N. Ilias, and M. Kremer. 2003. "Teacher Incentives." National Bureau of Economic Research, Cambridge, Mass.

Glewwe, P., M. Kremer, and S. Moulin. 2001. "Textbooks and Test Scores: Evidence from a Randomized Evaluation in Kenya." World Bank, Development Research Group, Washington, D.C.

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz. 2004. "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics* 74 (1): 251–68.

Gosden, T., F. Forland, I. S. Kristiansen, M. Sutton, B. Leese, A. Giuffrida, M. Sergison, and L. Pedersen. 2004. "Capitation, Salary, Fee-for-Service and Mixed Systems of Payment: Effects on the Behaviour of Primary Care Physicians." *The Cochrane Library* Issue 3.

Grossman, J. B. 1994. "Evaluating Social Policies: Principles and U.S. Experience." *The World Bank Research Observer* 9 (2): 159–81.

Gueron, J. M. 1997. "Learning about Welfare Reform: Lessons from State-Based Evaluations." *New Directions for Evaluation* 76: 79–94.

———. 2002. "The Politics of Random Assignment: Implementing Studies and Affecting Policy." In F. Mosteller and R. Boruch, eds., *Evidence Matters: Randomized Trials in Education Research*. Washington, D.C.: Brookings Institution Press.

Gueron, J. M., and G. Hamilton. 2002. "The Role of Education and Training in Welfare Reform." Welfare Reform & Beyond Policy Brief 20. Brookings Institution, Washington, D.C.

Habicht, J.-P., C. G. Victora, and J. P. Vaughan. 1999. "Evaluation Designs for Adequacy, Plausibility and Probability of Public Health Programme Performance and Impact." *International Journal of Epidemiology* 28 (1): 10–18.

Inter-American Development Bank. 2002. "Annual Report of the Office of Evaluation and Oversight, 2001." RE-286-1. Washington, D.C.

———. 2004. "Progress Report on Management's Actions in 2003 and Future Actions to Enhance the Bank's Development Effectiveness." CA-456. Washington, D.C.

ILO (International Labour Office). 2002. "Extending Social Protection in Health through Community Based Health Organizations: Evidence and Challenges." Discussion Paper. Universitas Programme, Geneva.

Jacquet, P. 2006. "Evaluations and Aid Effectiveness." In Nancy Birdsall, ed., *Rescuing the World Bank: A CGD Working Group Report and Collected Essays.* Washington, D.C.: Center for Global Development.

Jakab, M., C. Krishnan, A. Preker, A. Gumber, A. Kelly, K. Ranson, P. Schneider, and S. Supakankunti. 2001. "The Impact of Community Financing on Health, Protection Against Impoverishment and Social Inclusion:

What Do Household Data Tell Us?" HNP Discussion Paper submitted as a background report for the Commission on Macroeconomics and Health. World Bank, Washington, D.C.

Jamison, D. T. 1978. "Radio for Formal Education and for Development Communication." *Development Communication Report* 24: 1–2.

Kremer, M. 2003. "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons." *American Economic Review* 93 (2): 102–15.

LaLonde, R. J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (4): 604–20.

Levine, D. I. 2005. "Learning to Teach (and to Inoculate, Build Roads and . . .)." University of California at Berkeley, Haas School of Business, Berkeley, Calif.

Levine, R., and the What Works Working Group with M. Kinder. 2004. *Millions Saved: Proven Successes in Global Health*. Washington, D.C.: Center for Global Development.

Liebman, J. B., L. F. Katz, and J. Kling. 2004. "Beyond Treatment Effects: Estimating the Relationship between Neighborhood Poverty and Individual Outcomes in the MTO Experiment." KSG Working Paper RWP04-036. Harvard University, John F. Kennedy School of Government, Cambridge, Mass.

Lynam, D. R., R. Milich, R. Zimmerman, S. P. Novak, T. K. Logan, C. Martin, C. Leukefeld, and R. Clayton. 1999. "Project DARE: No Effects at 10-year Follow-up." *Journal of Consulting and Clinical Psychology* 67 (4): 590–93.

Marcel, M. 2003. "Evaluación de programas sociales en el sistema de presupuesto por resultados en Chile." Presentation at Conferencia Internacional Mejores Prácticas de Política Social, May 7, Mexico City.

Miguel, E., and M. Kremer. 2001. *Worms, Education and Health Externalities in Kenya*. NBER Working Paper 8481. Cambridge, Mass.: National Bureau of Economic Research.

MkNelly, B., and C. Dunford. 1998. "Impact of *Credit with Education* on Mothers and Their Young Children's Nutrition: Lower Pra Rural Bank *Credit with Education* Program in Ghana." Research Paper 4. Freedom from Hunger, Davis, Calif. [www.ffhtechnical.org].

Morley, S., and D. Coady. 2003. *From Social Assistance to Social Development: Targeted Education Subsidies in Developing Countries*. Washington, D.C.: Center for Global Development.

Mosteller, F., and R. Boruch, eds. 2002. *Evidence Matters: Randomized Trials in Education Research*. Washington, D.C.: Brookings Institution Press.

Mosteller, F., R. J. Light, and J. A. Sachs. 1996. "Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size." *Harvard Education Review* 66 (4): 797–842.

National Institutes of Health. 2003. "NIH Program Evaluation Guide: How to Develop a Proposal for Evaluation Set-Aside Funding." National Institutes of Health, Bethesda, Md.

Newhouse, J. P. 2004. "Consumer-Directed Health Plans and the RAND Health Insurance Experiment." *Health Affairs* 23 (6): 107–13.

O'Donoghue, T., and M. Rubin. 1999. "Doing It Now or Later." *The American Economic Review* 89 (1): 103–24.

Olken, B. A. 2004. *Monitoring Corruption: Evidence from a Field Experiment in Indonesia*. NBER Working Paper 11753. Cambridge, Mass.: National Bureau of Economic Research.

Orr, L. L. 1996. "Why Experiment? The Rationale and History of Social Experiments." Part I of *Social Experimentation: Evaluating Public Programs with Experimental Methods*. Washington, D.C.: US Department of Health and Human Services. [http://aspe.os.dhhs.gov/hsp/qeval/part1.pdf].

Petrosino, A., C. Turpin-Petrosino, and J. O. Finckenauer. 2000. "Well-Meaning Programs Can Have Harmful Effects! Lessons from Experiments of Programs Such as Scared Straight." *Crime & Delinquency* 46 (3): 354–79.

Picciotto, R. 2000. "Economics and Evaluation." Paper presented at the European Evaluation Society Conference, August 28, Lausanne, Switzerland.

Preker, A., G. Carrin, D. Dror, M. Jakab, W. Hsiao, and D. Arhin-Teknorang. 2001. "A Synthesis Report on the Role of Communities in Resource Mobilization and Risk Sharing." CMH Working Paper Series WG3:4. World Health Organization, Commission on Macroeconomics and Health, Geneva, Switzerland.

Pritchett, L. 2002. "It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." *The Journal of Policy Reform* 5 (4): 251–69.

Quigley, R., S. Cavanagh, D. Harrison, and L. Taylor. 2004. "Health Development Agency 2004: Clarifying Health Impact Assessment, Integrated Assessment, and the Health Need Assessment." National Health Service in England, Health Development Agency, Chesterfield, UK.

Rosenbaum, D. P., and G. S. Hanson. 1998. "Assessing the Effects of School-Based Drug Education: A Six-Year Mutli-Level Analysis of Project D.A.R.E." *Journal of Research in Crime and Delinquency* 35 (4): 381–412.

Rossi, P. H., H. E. Freeman, and M. W. Lipsey. 1999. *Evaluation: A Systematic Approach.* Thousand Oaks, Calif.: Sage Publications.

St. Pierre, Robert G. and Jean I. Layzer. 1999. "Using Home Visits for Multiple Purposes: The Comprehensive Child Development Program," *The Future of Children* 9 (1) 134–51.

Schultz, T. P. 2000. "Final Report: The Impact of Progresa on School Enrollments." International Food Policy Research Institute, Washington, D.C.

Scott, C. 2005. "Measuring Up to the Measurement Problem: The Role of Statistics in Evidence-Based Policy-Making." PARIS21, London.

Sommer, A., I. Tarwotjo, E. Djunaedi, K. P. West, A. A. Loeden, R. Tilden, and L. Mele. 1986. "Impact of Vitamin A Supplementation on Childhood Mortality: A Randomised Controlled Community Trial." *Lancet* 1 (1): 1169–73.

Stern, N. 2002. "The Challenge of Monterrey: Scaling Up." Keynote Speech at the Annual World Bank Conference on Development Economics, December 17, Oslo.

Todd, P., and K. I. Wolpin. 2004. "Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility: Assessing the Impact of a School Subsidy Program in Mexico." University of Pennsylvania, Philadelphia.

USAID (US Agency for International Development). 2002. "USAID Child Survival and Health Programs Fund Progress Report." Washington, D.C.

Viadero, D., ed. 2004. "Ed. Dept. Issues Practical Guide to Research-Based Practice." *Education Week* 23 (16): 12.

Victora, C. G. 1995. "A Systematic Review of UNICEF-Supported Evaluations and Studies, 1992–1993." Evaluation and Research Working Paper Series 3. United Nations Children's Fund, New York.

Victora, C. G., J.-P. Habicht, and J. Bryce. 2004. "Evidence-Based Public Health: Moving Beyond Randomized Trials." *American Journal of Public Health* 94 (3): 400–05.

Wilde, E. T., and R. G. Hollister. 2002. "How Close Is Close Enough? Testing Nonexperimental Estimates of Impact against Experimental Estimates of Impact with Education Test Outcomes." 1242-02. University of Wisconsin at Madison, Institute for Research on Poverty, Madison, Wis.

Wilson, M. 1998. "How Congressional Conferees Can Improve Job Training Reform." *Backgrounder* 1203. The Heritage Foundation, Washington, D.C.

Wiseman, M., P. Szanton, E. Baum, R. Haskins, D. Greenberg, and M. Mandell. 1991. "Research and Policy: A Symposium on the Family Support Act of 1988." *Journal of Policy Analysis and Management* 10 (4): 588–666.

World Bank. 1999. *Poverty Reduction and the World Bank: Progress in Fiscal 1998*. Washington, D.C.

———. 2001. *Poverty Reduction and the World Bank: Progress in Fiscal 2000*. Washington, D.C.

———. 2004a. "Influential Evaluations: Evaluations that Improved Performance and Impacts of Development Programs." Washington, D.C.

———. 2004b. "Monitoring and Evaluation: Some Tools, Methods and Approaches." Washington, D.C.

WHO (World Health Organization), Commission on Macroeconomics and Health. 2001. *Macroeconomics and Health: Investing in Health for Economic Development*. Report of the Commission on Macroeconomics and Health. Jeffrey D. Sachs, Chair. Geneva.

WHO (World Health Organization). 1978. "Financing of Health Services." Technical Report Series 625. Geneva, Switzerland.

WHEN WILL WE EVER LEARN?

IMPROVING LIVES THROUGH IMPACT EVALUATION

ISBN 1-933286-11-3

Center for Global Development
1776 Massachusetts Avenue, N.W.
Third Floor
Washington, D.C. 20036
Tel: 202 416 0700
Web: www.cgdev.org