



**When Will We Ever Learn?
Recommendations to Improve Social Development
through Enhanced Impact Evaluation**

**Consultation Draft
September 15, 2005**

by William D. Savedoff
Ruth Levine
Nancy Birdsall

Center for Global Development, Washington, DC

<http://www.cgdev.org/section/initiatives/active/evalgap>

When Will We Ever Learn? Recommendations to Improve Social Development through Enhanced Impact Evaluation

Foreword

This document is a draft position paper drawn from analyses, discussions and other inputs from the Evaluation Gap Working Group, which was convened by the Global Health Policy Research Network, a program of the Center for Global Development. (See Appendix II for a list of Working Group members.) These inputs included a review of the literature, an inventory of current activities designed to improve evaluation of development assistance, more than 50 interviews with development experts, and a series of in-person and e-discussions among Working Group members.

While the current draft position paper represents the views of the authors (Savedoff, Levine and Birdsall) and not necessarily the views of any of the individuals who participated in the Working Group, we have attempted to accurately represent the areas of agreement and disagreement among Working Group members. Working Group members generally concurred on the conclusions presented in this paper about the importance of impact evaluation of social programs in developing countries and the reasons for limited impact evaluation within both developing country governments and international agencies. The recommendations regarding the establishment of an Impact Evaluation Club and its institutional functions and funding were subscribed to by most, but not all, of the Working Group members. As noted in the report, some members of the Working Group favored a different type of approach that would focus more on coordinating across institutions, rather than on direct funding of evaluation-related activities.

With the distribution of this draft position paper, we are soliciting comment and critique from a broad set of interested parties between September and November 2005, and will use those views to finalize the work of the group by December 2005.— Please address comments to the authors, William Savedoff (savedoff@socialinsight.org), Ruth Levine (rlevine@cgdev.org) and Nancy Birdsall (nbirdsall@cgdev.org) or visit our website at <http://www.cgdev.org/section/initiatives/active/evalgap>.

We are grateful for financial support and intellectual contributions from the Bill & Melinda Gates Foundation and the William and Flora Hewlett Foundation.

Table of Contents

Executive Summary iii
I. An Evaluation Gap is Restraining Progress 1
II. What is the Evaluation Gap? 4
III. Impact evaluation is critical for building knowledge..... 17
IV. Impact evaluations are feasible and desirable 18
V. Why are good impact evaluations so scarce? 23
VI. Solutions 25

Boxes, Figures and Additional Material

Box 1: How Would an Impact Evaluation Club Work?x
Figure 1: Basic Scheme for Monitoring and Evaluation.....5
Box 2: The Development IMPact Evaluation (DIME) Initiative.....8
Box 3. “Before and After” Studies Can Be Misleading.....10
Box 4. Advantages and Limitations of Random Assignment Studies.....12
Figure 2: Many studies lack methodological rigor.....16
Box 5: Timeliness of funds can be critical to good impact evaluations.....24
Box 6. Draft Proposal for an Impact Evaluation Club.....34
References.....39
Endnotes.....45

Executive Summary

We need to know more about social development

Social programs are at the core of an international consensus regarding global progress. Whether the debates are held in parliamentary and ministerial meetings over social and economic planning or in conferences held by the Bretton Woods Institutions, the United Nations, or the G-8, the need for investments in health, education and poverty reduction is broadly recognized. Governments and international agencies spend more than US\$30 billion each year on social development programs, while developing countries spend at least another US\$300 billion annually on such programs. Through commitments to the Millennium Development Goals and subscriptions to a variety of global initiatives, donor countries have proposed doubling or tripling their financial assistance to the world's poorest countries.

Yet after five decades, in which development agencies have disbursed billions of dollars, it is still not clear how this development assistance should be spent. While the goals may be clear – improving health, increasing educational attainment, reducing poverty – the documented record of specific social interventions that succeed is relatively sparse. Developing countries that choose to face the challenges of social development lack strong evidence to guide their social policies and budget allocations.

Despite applying resources to design social development projects, monitor their implementation and measure their outputs, little is actually done to assess whether the projects ultimately achieve their aims. Thus, we know, for example, that India spent US\$1.3 billion on primary education as part of the DPEP program, that this money was spent on schools, teacher training, and the like, and that many people think the program is a good one; but we do not actually know whether the program “worked” in terms of increasing educational attainment.

We are facing an Evaluation Gap

The missing puzzle piece in the process of learning what kinds of social interventions can succeed requires studies, known as “impact evaluations”, to document whether changes in social outcomes can be attributed to particular programs. An “Evaluation Gap” has emerged because governments, official donors and other funders do not demand or produce enough of these impact evaluations, and because those that are conducted are frequently of poor quality.

The numbers of impact evaluations actually being carried out is too low, even in light of the fact that building knowledge does not require an evaluation of each and every project. 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 are less useful for evaluating which interventions work under given conditions and at what cost.

Even when impact evaluations are commissioned, they frequently fail to provide useful information because they are methodologically weak or they are poorly implemented. A systematic review of UNICEF estimated that 15% 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). Similarly, a review of 127 studies of 258 community health financing programs found only *two* studies that were able to derive robust conclusions about the impact on access to health services.

Poor quality evaluations are misleading. No one would consider prescribing strong medications without properly evaluating their impact or potential side effects. Yet, in social development programs, where huge sums of money 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 – namely that programs of unknown impact should not be widely replicated without proper controlled evaluation – is frequently dismissed.

Why is there an Evaluation Gap?

This Evaluation Gap occurs because there are too few incentives to conduct good impact evaluations and too many obstacles. While impact evaluations generally have to be designed as part of a new program, politicians and project managers are focused in these early program phases on designing and implementing their programs. The costs of starting an impact evaluation at this early stage are real and present, while the benefits of measuring the impact will only materialize in 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.

In addition, the outcome of an impact evaluation cannot be known in advance. Hence, those in charge of social programs in developing countries may be unwilling to face the risk of proving their program “failed”; while those involved in international assistance may fear that a few “negative” studies could undermine domestic support for foreign aid more broadly.

Reasons for optimism

Yet tolerance for the Evaluation Gap is waning. Developing country governments themselves are demanding better information about the efficacy of social spending. In 2001, Mexico passed legislation requiring 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 does and doesn't work. Donor countries are increasingly concerned that international financial assistance should generate "results".

Meanwhile, good impact evaluations are becoming cheaper and easier to implement. 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 the impacts that can be attributed to a particular program or policy. From among these methods, expert panels have identified random assignment as the most reliable approach to measuring net impact. The methods of random assignment – in which pilot programs are randomly offered to potential beneficiaries – have been proven effective in a variety of settings and for a wide range of programs. Furthermore, notable examples show that good quality impact evaluations are feasible, ethical, timely and useful. Such studies have helped NGOs modify educational 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. Rigorous impact evaluations using other methods are also needed whenever random assignment is not possible or appropriate, for example, when a program has already been implemented at large scale, or when the intervention affects the whole population, as is the case for changes in legislation.

Building this knowledge requires that governments and agencies take a strategic view and conduct impact evaluations in those projects that can yield important information about what is and is not effective. 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 generalizability of any findings.

What can we do?

We found that a wide range of actors – governments, public agencies, intergovernmental commissions, non-governmental networks, research centers and foundations – are already addressing the need for building knowledge about social development programs and should be further strengthened. Such activities include:

- Monitoring programs and conducting operational research
- Increasing access to existing information through reviews, searchable databases, and policy pamphlets and newsletters.
- Improving regular data collection by developing country governments
- Promoting specific impact evaluations with grants and other kinds of funding
- Conducting research and demonstrating good evaluation practices

However, the value of impact evaluations in building knowledge about “what works” in social development policies is so large relative to the existing investments, and the incentives for undertaking these studies are so weak, that a special initiative is needed. In particular, we seek feedback on this consultation draft’s principal recommendation: that governments, international agencies, and private foundations should establish an Impact Evaluation Club — a funding facility with its own governance, dues, and standards in which membership would be voluntary.¹ The particular design of such an Impact Evaluation Club would ultimately be determined by its initial members. The following ideas are presented to illustrate a potential framework for addressing the incentive problems, the public good issues, and the need for independence and technical excellence.

Main features of the Impact Evaluation Club. The Impact Evaluation Club would have several important features:

First, membership would be voluntary and would include as members the governments of several developing country governments, as well as agencies that provide development assistance. This would help to ensure that the Club serves the needs of developing country decision makers, rather than solely those whose interest might be limited to the performance of donor-funded projects. The diverse membership would also permit development agencies to contribute to the improvement in the use of evaluation within developing country social policy, which is essential if the new modalities of aid delivery, including expanded use of budget support and/or multi-sectoral poverty reduction credits, are to live up to their potential.

Second, the Club would aim to improve international learning from impact evaluations through such activities as setting or adopting quality standards; exchanging information among members and/or building a registry for impact evaluations; investing in development of improved evaluation methods; and other related activities.

Third, such a Club would help developing countries and international agencies cluster studies around common themes and questions to increase their usefulness in learning what interventions are effective for achieving particular types of development objectives, and under what circumstances.

In the process of developing recommendations for an Impact Evaluation Club that would improve the supply and quality of impact evaluations, the Working Group faced the question of whether such an entity would have to involve itself in financing impact evaluations. Under one possible scenario, a Club would not have the capacity or mandate to mobilize and provide funding for independent impact evaluations. This would be left up to the individual developing country governments and their development partners, and the Club’s role would be consultation about key evaluation questions, coordination,

¹ This is not a proposal for an undertaking by the Center for Global Development, whose mission does not include conducting or managing major impact evaluations. Rather, it is a proposal for consideration by the broad international community of developing country governments, donor agencies, international technical agencies and coordinating bodies, and their constituencies.

quality assurance and communication. To some extent, these functions are currently being undertaken by existing bodies, and so such an approach is relatively undemanding. Under an alternative scenario, which includes the additional features listed below, the Club would mobilize and strategically allocate funds for the design and conduct of impact evaluations. The principal reasons for favoring this approach are:

- The value of independence: An independently funded evaluation is far more likely to be seen as credible by the range of policymakers and stakeholders who use evaluation results to decide which social programs to support and in what form.
- The benefits of single-mindedness: An institution or group that has the core mission of promoting and generating impact evaluations is more likely to be able to achieve this aim than an institution with many other roles. In donor agencies, for example, whose main business is delivering funding for projects, operational demands often dominate.
- The potential for shared resources: If every organization funds impact evaluations out of its own resources, inevitably some agencies and governments will have more evaluation funding than others. This greatly limits the ability of smaller and/or less well-resourced governments and agencies to engage in policy debates about what works and what doesn't, to demonstrate the effectiveness of their own programs, and to learn. Pooled resources can level this playing field.

Thus, in the view of most Working Group members, several additional features would be required for the Club to make a qualitative improvement in the base of knowledge about social development programs.

As a fourth feature, the Club would mobilize and distribute funds for impact evaluation. Members would contribute funds; part of their contribution could be provided in the form of a defined portion of their internal budgets to impact evaluation. A member's contribution would bear some relation to the scale of its development assistance, national budget, or endowment. Although international agencies and donor governments would benefit substantially from the creation of this international fund, the real beneficiaries would be citizens in developing countries to the extent that knowledge from good impact evaluations would be used to accelerate social development.

Different "windows" would be created to issue requests for proposals around pre-selected themes; review spontaneous grant proposals; and finance short-term grants oriented toward exploring the feasibility of evaluating a program or collecting baseline data.

Fifth, because studies with randomized assignment face the largest obstacles relative to their promise in knowledge building, more than half of the Club's funds should be earmarked to support studies with randomized designs. Other evaluation methods that can be effectively employed when random assignment designs may be undesirable or

infeasible would also be supported, while maintaining an emphasis on *impact* evaluation rather than routine monitoring or assessment of process-related or other intermediate endpoints.

Sixth, the Club would finance impact evaluations on a competitive basis. Any studies chosen for support by the Club would have to fulfill the following conditions:

- the research design is independently reviewed under rigorous quality standards,
- basic fiduciary responsibilities are fulfilled,
- the final study is submitted for independent review under rigorous quality standards, and
- once approved on methodological grounds, the study must be in the public domain regardless of its findings; and the data must be made available to other researchers.

Benefits of the Club. High-level decision-makers would have an incentive to enroll their institutions as Club members to leverage additional funds, participate in selecting study subjects, and comply with mandates from their own stakeholders to be more results oriented.

At the program level, managers, agency staff and policymakers who have an interest in learning from impact evaluations would find that the existence of the Impact Evaluation Club would lower the costs and barriers they face since the Club would provide dedicated funds with established guidelines for technical support, design standards, and dissemination.

Developing country governments would be empowered by the Club in several ways. First, they could participate fully as members in advising the Club's administrative team regarding priority areas for research. Second, as members, they would dedicate some domestic resources for impact evaluations that would accelerate their own learning about which of domestic social programs are effective. Third, they could leverage additional funds to supplement domestic resources for these studies. Fourth, they could set the priorities for which internal programs should be evaluated by choosing which proposals to submit for funding. Fifth, they would get access to expert advice on impact evaluation design. Finally, the studies commissioned under the Club's guidelines would have international credibility, gaining legitimacy from the review process and transparency created by this independent institution.

Resource requirements. Completing good impact evaluations requires a substantial commitment of funding and time. A more detailed study of financing requirements should be undertaken as part of negotiating the final design of the Impact Evaluation Club; however, prospective members should recognize that some studies might cost as much as US\$10 or US\$20 million over a 7 to 10 year period. A preliminary estimate suggests that within five years, the Club could be operating with an annual budget of US\$30 million, of which 7% would be dedicated to costs of administration and professional networking. Most of this funding would be additional to current spending on impact evaluations in member organizations. In addition, funds that are currently

dedicated to evaluation in donor-funded social development projects could be used more effectively with the addition of resources at the margin from the Impact Evaluation Club.

Ultimately, the design of the Club will be decided by its initial members. This proposal merely aims to show how a Club *could* be constituted on a *voluntary* basis with a clear mandate. It would *not* impose agendas or centralize management of impact evaluation; rather it would serve as an external source of earmarked funds and technical support for impact evaluations and provide a forum for exchanging information among members on their evaluation work and agendas. Since the initial members will determine the shape of the Club, it is essential that some minimum threshold be established for participation, including enough developing country governments, donor country governments, international agencies, foundations and NGOs to assure that the design can serve its purpose and still be of value to this wide range of important stakeholders.

Hope for the future

Governments, public agencies and private foundations are making progress toward a world with better health, more education and less poverty. But we can reach those goals faster and more effectively by systematically building knowledge about what kinds of social development interventions do and do not work. The recommendation we propose here will not single-handedly achieve this goal, but it can contribute an important, and missing, element to that effort – a way to find out what works.

Box 1: How Would an Impact Evaluation Club (ImEC) Work?**Box 1: How Would an Impact Evaluation Club (ImEC) Work?***Hypothetical Story #1: An African Education Program:*

The Education Minister of a low-income African country has a new program that hires local residents to serve as teacher aides in rural areas. The program aims to reduce school closures when the teacher is absent, increase attendance, and raise student achievement. The program is loosely modeled on a similar program that the Minister visited in India, but she doesn't know if it will work in her country. The Minister has only obtained enough funding from the national budget to begin the program in 20 districts, along with a loan from a multilateral development bank (MDB). The Minister wants an evaluation so that, if it is successful, she will have a strong argument for obtaining funding from the Finance Ministry to expand the program nationwide.

The Minister's advisor is negotiating a loan with the MDB but finds that the evaluation funds in the loan will only cover the requirements for monitoring implementation, and he doesn't want to take funds from the program activities for an impact evaluation. After consulting with his counterpart at the MDB and at the Impact Evaluation Club (ImEC), the advisor prepares a proposal to the Impact Evaluation Club's rapid disbursement project preparation window. The initial grant makes it possible to contract a team of experts from a university to suggest a research strategy. The strategy involves rolling out the program in randomly selected districts in three phases, with regular feedback to policymakers on the program's progress. The strategy forms the basis of an application to the Impact Evaluation Club for US\$5 million to collect data and conduct the impact evaluation over a 7-year period. The initial data collection is financed with a US\$100,000 budget item, earmarked for impact evaluation as part of the country's membership in the Impact Evaluation Club. The Minister and Advisor are pleased when their proposal is awarded the grant by the ImEC after competing with other proposals in an international peer-review process.

Hypothetical Story #2: An NGO Responds to a Request for Proposals

An NGO has a microcredit program in a few municipalities of a large Asian country. The program appears to be successful and a program officer is charged with expanding the program to 24 new municipalities. The NGO is a member of the ImEC, and it receives a request for proposals to evaluate the impact of microcredit programs on family poverty. The program officer discusses the possibility of conducting an evaluation with her counterparts in the Asian country and they apply for funding to the ImEC.

Box 1: How Would an Impact Evaluation Club (ImEC) Work? (continued)*Hypothetical Story #3: A Bilateral Agency Needs Evidence*

Parliament tells a bilateral agency to close down its nutrition program in Central America, claiming that the program is wasteful. A staff member in the research department finds two impact evaluations on the ImEC website that show similar projects have had some success in an Asian country. The Agency's director uses those reports to convince Parliament to continue funding with the stipulation that an impact evaluation will be conducted to ascertain whether the Agency's programs are equally effective in the Central American context.

A researcher is given responsibility to work with an operations officer and the local government to see whether an impact evaluation would be feasible and determines that it would be difficult to do a rigorous evaluation on the existing program. However, through discussions with staff at the ImEC, they discover that a neighboring country is planning to implement a similar program, phased in over several years, with the support of another bilateral agency. The researcher and program officer arrange a meeting in the neighboring country and offer to cosponsor an impact evaluation. The ImEC provides some seed money to pay for experts in research design to participate in that meeting. The bilateral agencies and the Central American country agree to co-finance the impact evaluation; however, their budgetary allocations are only guaranteed for the evaluation's first two years. Consequently, they apply for a grant from the ImEC to assure funding in the program's third and fourth years – necessary to complete the evaluation.

When Will We Ever Learn? Recommendations to Improve Social Development through Enhanced Impact Evaluation

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

OECD/DAC “Principles for Evaluation of Development Assistance”

The Millennium Development Goals call, among other things, for increasing primary school completion, reducing child mortality rates and the incidence of malaria, and making the benefits of new technologies, especially information and communication technologies, available in developing countries. However, we know surprisingly little about what are the most effective ways to reach these goals. For example, when the Center for Global Development convened a group of health experts to nominate successful public health projects, resulting in the book *Millions Saved*, it discovered that less than two dozen programs could substantiate their claims of success with good evidence.

This paper demonstrates that there is a serious gap in our knowledge about what are the most effective social development programs in developing countries. This in turn reflects a lack of high quality impact evaluations, that is, evaluations that are designed to measure the impact directly attributable to a specific program or policy as distinct from other potential explanatory factors. The paper argues that the cost of not having this information – in terms of misallocating resources and even in some cases causing harm – is high. It analyzes the obstacles to more systematic production of good quality impact evaluations and demonstrates that confronting these obstacles requires collective action by international agencies and governments. It is hoped that the recommendations in this paper will serve as the basis for international discussion and agreements that will improve the quality and focus of impact evaluations so that 5 or 10 years from now, we will have answers to important questions about what kinds of social development programs are most likely to succeed.

I. An Evaluation Gap is Restraining Progress

Knowledge is a global public goodⁱ which has the potential to improve millions of lives without being depleted. For example, the discovery in the late 19th Century that cholera was transmitted through contaminated water has saved lives around the world and over many years. In this regard, the knowledge gained from evaluating public programs is the most systematic approach available for improving public policy and spending public money well. Thus, it is an invaluable complement to the political processes that establish public policy. Fortunately, the capacity to conduct good evaluations and learn from them has increased dramatically in the last few decades as a consequence of methodological

advances, expanding research institutions, a growing number of educated and skilled evaluators, and more effective evaluation offices in public agencies around the world.

Nevertheless, the evidence base for designing new programs and providing financial assistance remains quite weak. A substantial amount of resources are applied to designing projects, monitoring their implementation and even to measuring their outputs, but very little is done to assess whether the projects are ultimately successful and necessary for achieving positive impacts. This underinvestment in good quality “impact evaluations” – studies designed to measure the net impact directly attributable to a program or policy – means that opportunities for expanding good projects are lost and that funds continue to be wasted on bad ones.

We urge the MDBs 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

G7 Finance Ministers, Nova Scotia, June 20, 2002

This gap is puzzling because building this knowledge does not require an impact evaluation for every project. It requires only that agencies take a strategic view and conduct impact evaluations in those projects that can yield important information about what is and is not effective. For example, it would suggest paying particular attention to projects that are widespread but for which effectiveness has not been demonstrated, or to new approaches that have not yet been tested. Given the volume of money devoted to social programs and the emphasis on “results,” it is puzzling that the design of careful impact evaluations for these kinds of projects – with clear strategies for drawing valid inferences based on plausible counterfactual scenarios – are so rare.

This gap in knowing what projects and programs are effective is made worse by the poor implementation of impact evaluations in those cases where they have been commissioned. Too often, impact evaluations are neglected by development agencies and governments who, quite reasonably, are focused on implementation of the current project but then lose important opportunities to improve future policies. As a result, while they may generate studies that are important for improving processes, operations and implementation, they frequently neglect studies of impact. And when impact evaluations are conducted, all too often, their designs are so weak that they cannot reliably measure the net impact of the programmed activities.

As one example, 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). Since that time, millions of dollars have been spent on such programs in dozens of countries and reviews of the literature evaluating such programs give the impression that we know a great deal about these programs and that they are beneficial (e.g. Commission on Macroeconomics and Health 2001 and 2002). However, reviews that explicitly

discount studies that are methodologically weak, find that there is very little evidence actually available on whether these strategies are effective. The ILO's Universitas Programme reviewed 127 studies, studying 258 community health schemes, and found that only 2 of these cases had "internal validity", that is, only 2 of these studies 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. As they state:

...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 alia (sic), 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)

Other reviews on community health schemes confirm the methodological weakness of the literature identified in the ILO Report. (Ekman 2004, Jakab *et al* 2001).

This is not to say that impact evaluation has been barren in all fields or at all times. Good evaluations do happen, and when they are disseminated, they stand out in their field. Decades later, 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). More recently, 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).

In 2004, we reviewed existing initiatives to address this apparent underinvestment in good impact evaluations. We found that several organizations are working to improve impact evaluation in various ways: through advocacy, publishing guidelines, training programs, literature reviews, and promoting or conducting specific evaluations. However, no organization was asking why

international agencies and developing countries continually face the same dilemma over the course of decades. International agencies make it clear that they want to finance proven programs and governments insist on value for money, yet the frequent absence of evidence does not spur adequate investment in the good quality impact evaluations that are necessary to fill these gaps in our knowledge.

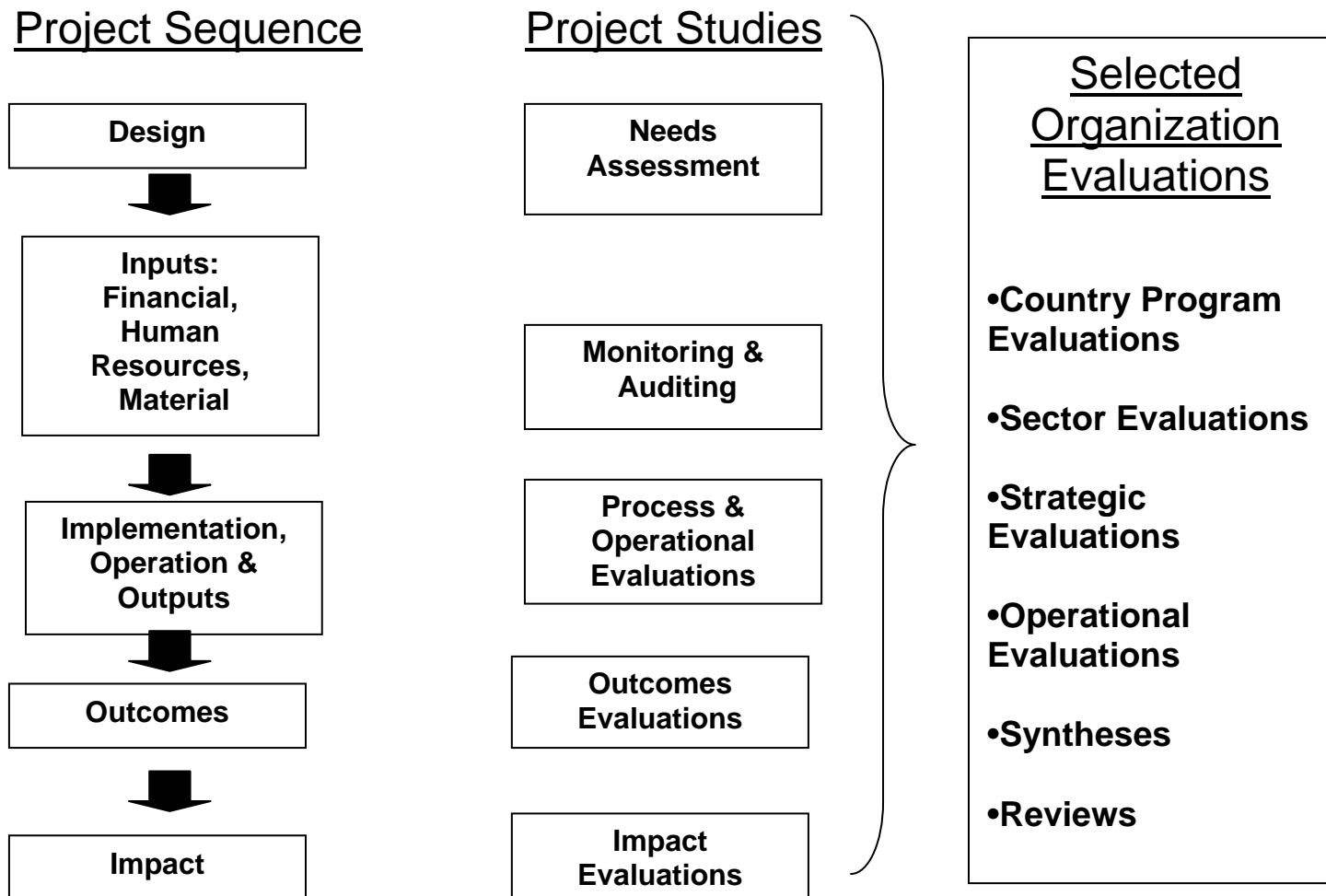
In particular, there is a need for the multilateral and bilateral financial and development institutions to intensify efforts to . . . [i]mprove ODA targeting to the poor, coordination of aid and measurement of results.

Monterrey Consensus

II. What is the Evaluation Gap?

Social development programs are complex. To promote social development wisely, decision-makers need many different kinds of information from different kinds of studies and reports. The basic scheme for monitoring and evaluation of social programs recognizes these different requirements (See Figure 1). Studies that report on the application of financial, human and material resources are essential to monitoring implementation, as are studies of program activities and outputs. Process Evaluations and Operational Evaluations – providing information on these inputs, activities and outputs – are important for improving program implementation, and critical to understand why a program was or wasn't effective. Without implementation, no program can achieve its goals. But without measuring the impact of a program, improved implementation has no purpose – in fact, it may be wasteful or harmful. Thus, another kind of study – impact evaluation – is necessary to find out whether or not a program is having the desired effects.

Figure 1 Basic Scheme for Monitoring and Evaluation



Evaluations can be generally classified according to their focus and objectives. Many evaluations are concerned with reviewing an entire portfolio of projects in a country or a particular sector and require evidence on the individual and joint effects of the projects. Other evaluations are concerned with internal performance, such as adherence to strategies or efficiency of administration. Evaluations of specific social development projects and programs tend to fall into the following categories:ⁱⁱ

- **Needs Assessment** – answers questions about the social conditions a program is intended to address and the need for the program.
- **Monitoring Implementation** – verifies the use of inputs in program activities according to the program design or approved adjustments, including audits.
- **Process & Operational Evaluation** – investigates the degree to which the program was implemented as designed and was completed on time and within budget, with the aim of improving implementation
- **Outcome evaluation** – documents the status of people after participating in a program (i.e. measuring gross changes).
- **Impact Evaluation** – documents the extent to which changes in the wellbeing of the target population can be attributed to the particular program (measuring net changes).

Of these five categories, our investigation has found substantial resources are spent on the first four of these categories of evaluations, but insufficient effort is going toward producing good quality impact evaluations. In developing countries, resources for social development programs are often overstretched such that information gathering activities are neglected. In countries with effective information gathering, most of these resources are directed toward monitoring the use of funds, the deployment and management of personnel, and the production of outputs and services. By contrast, relatively little is spent to rigorously assess whether programs are having the desired impact above and beyond what would occur without them.

Among bilateral and multilateral development assistance agencies, a share of program funds are generally dedicated to monitoring implementation and disbursement, but studies that look to these data sources for learning about program effectiveness are regularly disappointed. The kinds of data collected do not lend themselves to measuring net impact of programs.

In addition to evaluations of specific social development programs, governments and agencies regularly evaluate their own performance through independent studies of all their programs in a given country, region or sector, or of their policies, operations and strategies (See “Selected Organizational Evaluations”, Figure 1). While evaluating specific social programs is generally the responsibility of operational and research departments, independent Evaluation Departments in these agencies are primarily responsible for what might be termed “organizational evaluations” – analyzing processes and operations, program implementation, and policies and strategies. Much progress has been made to make these Evaluation Departments more independent, effective and useful to their agencies through internal reforms and through coordinated efforts among bilateral

agencies, such as the Development Assistance Committee's Evaluation Network, and including the multilateral development banks through the DAC/MDB Joint Venture on Managing for Development Results.

Despite this progress, these Evaluation Departments are constrained in areas where good impact evaluations are lacking. Without selective impact evaluations, Evaluation Departments have difficulty drawing conclusions about aid effectiveness. And yet, the independence of these departments bars them from collaborating with staff in the project preparation and implementation phasesⁱⁱⁱ – a necessary requirement for designing effective impact evaluations. The core work of these Evaluation Departments – providing independent assessments of the performance of their organizations and developing recommendations – would benefit if more good quality impact evaluations were undertaken so that they could draw inferences and conclusions about their agencies' resource allocation decisions, strategies, priorities, effectiveness and efficiency. In this way, more and better impact evaluations would complement and strengthen accountability among bilateral agencies and multilateral development banks.

The need for good impact evaluations is increasing as a result of current trends in international development assistance away from funding tied to specific projects and toward “budget support”, sector-wide approaches, “Poverty Reduction Strategies”, “Indefinite Quantity Contracts” at USAID and similar initiatives at other bilateral agencies, and a wide range of other efforts aimed at reinforcing developing countries as “owners” of their domestic social development agendas. Only through selective use of impact evaluation will it be possible for these broad programs to determine whether resources are flowing toward effective uses. Without this attention to measuring the impact of social development programs, the essential learning of what kinds of programs do and do not work will not take place.

Very few good impact evaluations are carried out

Given that many projects can learn from a single well-done evaluation, it is not necessary for all programs to include such studies. However, given the knowledge gaps identified above, the numbers of impact evaluations actually being carried out is clearly too low. For example,

- A systematic review of UNICEF reports found that 44 out of 456 were impact evaluations. The review estimated that 15% 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).
- Out of 139 studies conducted by Chile's Budget Department between 1997 and 2002, only 6 were impact evaluations (Marcel 2002).

An informal survey of international development organizations also demonstrates that few impact evaluations are undertaken. At the Inter-American Development Bank, out of 593 projects that were active as of July 2004, only 97 reported they had collected data on beneficiaries and, of these, only 18 had data on non-participants – information that is

necessary, though not sufficient, for evaluating impact.^{iv} Similar results can be found at most other regional development banks and bilateral agencies.

The World Bank may be doing better of late. In 1998, it reported that only 5% of its projects had associated impact evaluations, while in 2000, this share had increased to 10%. (World Bank 1999, World Bank 2001). The World Bank's Research Department is currently engaged in an initiative to make impact evaluation a more systematic endeavor within the bank, focused around six thematic areas (See Box 2). Impact evaluations that the World Bank has already completed or reviewed have already informed policy toward social funds, conditional cash transfers and educational decentralization. Much less is apparently known about other questions confronting the World Bank in nutrition, health sector reform, and attention to indigenous groups. Studies are also geographically concentrated in certain regions, most notably in Latin America and to a lesser extent in Sub-Saharan Africa.^v

Box 2: The Development IMPact Evaluation (DIME) Initiative

Box 2: The Development IMPact Evaluation (DIME) Initiative

The World Bank identified several bottlenecks that limit its ability to conduct impact evaluations at the necessary scale and with the needed continuity: insufficient resources, inadequate incentives, and, in some cases, lack of knowledge and understanding. To address these bottlenecks, the Development IMPact Evaluation (DIME) Initiative is a Bank-wide collaborative effort under the leadership of the Bank's Chief Economist that is oriented at: (1) increasing the number of Bank projects with impact evaluation components, particularly in strategic areas and themes; (2) increasing the ability of staff to design and carry out such evaluations, and (3) building a process of systematic learning on effective development interventions based on lessons learned from completed evaluations.

The Bank identified five thematic areas to concentrate its current efforts at impact evaluation – school based management and community participation in education; information for accountability in education; Teacher contracting; conditional cash transfer programs to improve education outcomes; and slum upgrading programs. Additional themes are under consideration. DIME envisions working in partnership with its member countries to identify opportunities for learning from impact evaluations that would be given technical support. It aims to improve internal incentives to undertake more systematic development impact evaluations by explicating recognizing these studies as a valued product in their own right.

DIME is a prominent illustration of the ways that some international agencies and governmental ministries are trying to build a "Culture of Evaluation" – making the conduct of selective and strategic impact evaluations and the use of impact evaluation findings an integral part of management and decision-making.

Source: World Bank. 2005. "The Development IMPact Evaluation (DIME) Initiative: Coordinating Impact Evaluation Work At The World Bank", Draft Report, World Bank: Washington, DC. And interviews with Francois Bourguignon, Ariel Fiszbein, Paul Gertler, and Carolie Gevers.

A further indication that too few impact evaluations are being conducted in sufficient numbers, despite being asked for and financed, comes from the shortcomings listed in evaluation reports themselves. The following selection from such studies does not represent a systematic survey, but is quite recognizable for anyone familiar with reading evaluation reports (See the appendix for further examples).

- “... this review revealed that, with the exception of Jalan and Glinskaya, none of the studies could qualify as true impact evaluations.” [An evaluation of a US\$1.3 billion primary education program in India with support from the World Bank, the European Commission, DFID, UNICEF and the Dutch government].
- “... there 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 NGO program to use computer technology in literacy training for adults in Zambia and India].

Many impact evaluations have flawed designs

Even when impact evaluations are carried out, they are often too flawed to provide reliable information. Some of the most common ways to estimate the impact of a program are to:

- compare outcomes before and after a program is implemented and
- compare outcomes in areas that receive a program with those that do not receive a program.

Unfortunately, both of these approaches can be seriously misleading. First, showing an improvement by comparing outcomes before and after a program has been introduced may tell us very little about the program’s impact (See Box 3). Many things change at the same time that a project is implemented, so without further information, it is not valid to assume that observed outcomes are due to the project.

Second, a comparison of changes in communities (or people) who receive program benefits with those who do not will be misleading whenever systematic differences between the two groups are ignored. In fact, many studies reach erroneous conclusions because they either ignore such systematic differences or cannot properly account for their confounding effects. Such systematic differences are common in social programs because in many cases the programs are implemented in communities where they are expected to have the best chances of success, while in other cases, program beneficiaries “self-select” into the programs.

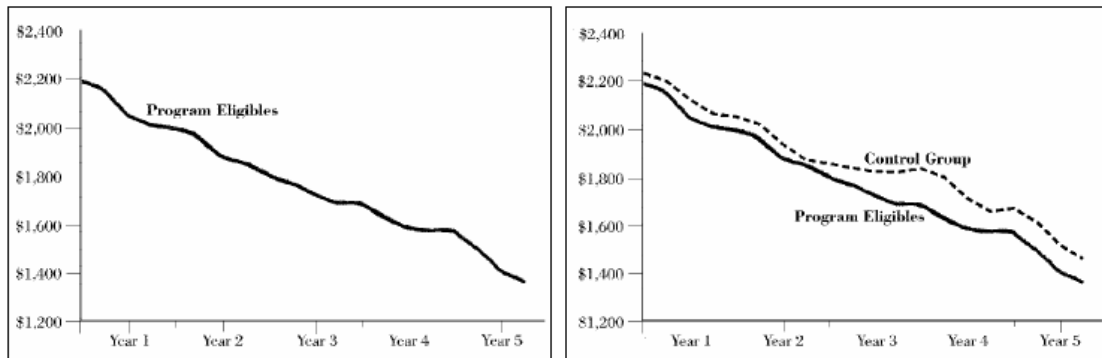
Box 3. “Before and After” Studies Can Be Misleading

Box 3. “Before and After” Studies Can Be Misleading

“Exhibit 1 shows how failure to take into account the temporary nature of the need for assistance can create a misleading impression of program effectiveness. The left hand panel of the exhibit shows the time path of quarterly public assistance benefits to a sample of AFDC [Aid for Families with Dependent Children] recipients who were eligible for a program designed to help them become employed and leave the welfare rolls. The steep downward trend in assistance payments would appear to indicate that the program was quite effective. This is in fact the type of “evidence” that is frequently used to demonstrate program effectiveness in the legislative process and in the popular media. As shown in the panel on the right, however, the decline in benefits was nearly as steep for a control group of program eligibles who were *excluded* from the program, as part of a social experiment. This comparison demonstrates that most of the decline in benefits experienced by the participant group was the result of normal turnover of the welfare rolls, as recipients’ circumstances improved and they were able to leave welfare. Attributing the effects of this turnover to the program greatly overstates its effectiveness.”

Average Quarterly Assistance Payments,
New York Child Assistance Program

EXHIBIT 1



Source: Orr, Larry L. 1996. “Why Experiment? The Rationale and History of Social Experiments”, Part I of *Social Experimentation: Evaluating Public Programs With Experimental Methods*. U.S. Department of Health and Human Services. Washington, DC. <http://aspe.os.dhhs.gov/hsp/qeval/part1.pdf>

A prominent example of the first kind of systematic bias is a large World Bank-financed initiative on education in India (DPEP) that was placed in areas where the potential for improvement was judged to be greatest. As a result, any comparison of target and non-target areas would be unable to tell whether improvements in primary schooling were due to the program or due to the better institutional conditions and implementation capacities of the beneficiary communities.

The second kind of bias, “self-selection”, occurs whenever the people who choose to participate in a program systematically differ from those who decide not 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 whether it can be attributed to the intervention.

For example, it is common to find studies that show students in private schools perform better than those in public schools. However, the students in private schools are there as a result of choices made by families. Even after controlling for socioeconomic factors, children who attend private schools are likely to come from families who are more committed to education and more motivated. To disentangle the effect of the schools from the effects of these unobservable characteristics is exceedingly difficult unless random assignment is used.^{vi}

When the selection of beneficiaries is influenced by any of these patterns, 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 and numerous studies have shown how prevalent such bias can be (Glazerman and Levy 2003, Lalonde 1986).

To avoid these problems it is usually necessary to build an evaluation into the design of a program from the start so that appropriate control groups can be identified. The most straightforward and the most reliable way to generate appropriate control groups is to use random assignment – that is, randomly choosing which individuals, families or communities will be offered a program and which ones will not. Where this method can be applied, it has been shown to be the most effective way to assure that impact measurements are not confounded by systematic differences between beneficiary and control groups (See Box 4).

BOX 4

Box 4. Advantages and Limitations of Random Assignment Studies

The key advantage of random assignment studies is that they can effectively reduce unobserved bias and give 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 only account 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.

To see why this is the case, consider job training programs that were financed by the United States in the 1970s and 1980s. Several studies compared the earnings of job trainees to individuals in the general population with similar characteristics who did not partake in the job training program. These studies 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 these 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 (See Figure A).

A prospective study that randomly assigned which applicants could participate in the job training program, however, yielded opposite results – finding that job training did lead to improved earnings. The random assignment study was able to reach the correct conclusion because it only compared individuals who would be eligible for the program, thereby eliminating the unobserved differences that skewed the other studies (See Figure B).

BOX 4 (continued)

Figure A. Impact Evaluation Using Comparisons Based on Observed Characteristics

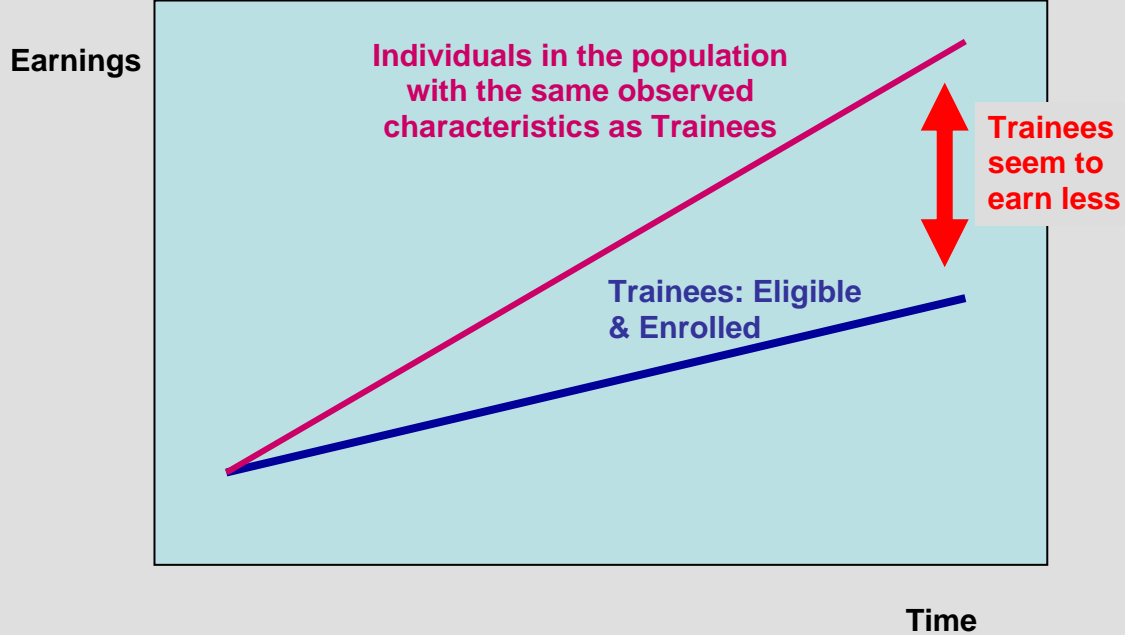
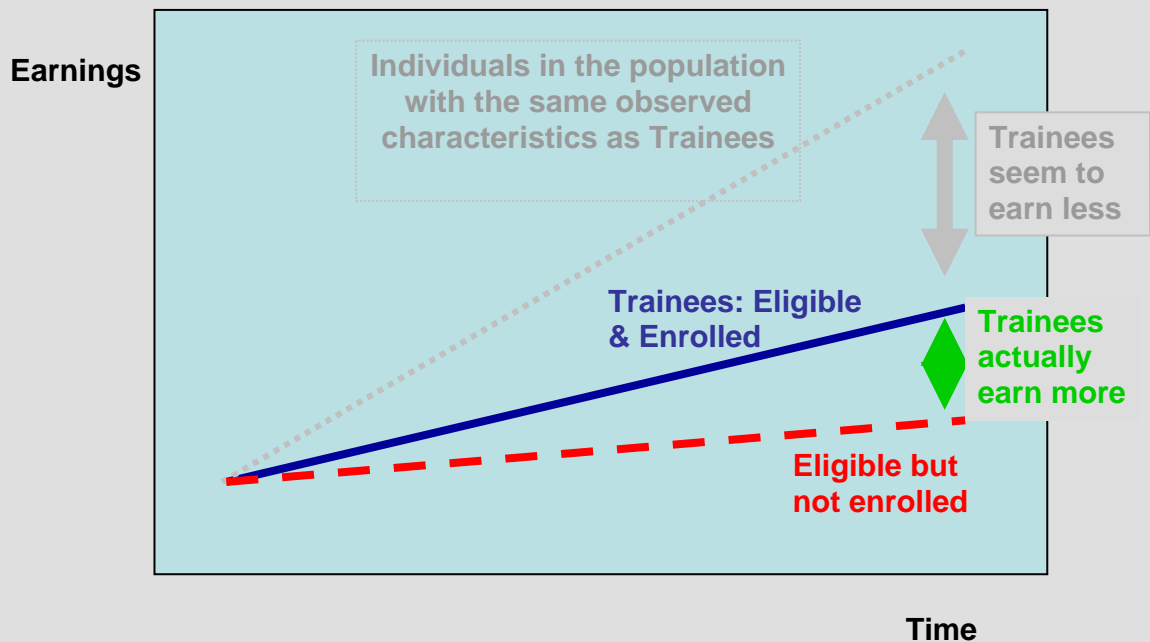


Figure B. Impact Evaluation Using Comparisons Based on Random Assignment



The literature that demonstrates how misleading studies can be is extensive and growing. It includes comparisons between studies with and without randomized assignment on topics such as: the impact of neighborhood poverty on individuals (Liebman, Katz and Kling 2004); the effectiveness of training programs (Mathematica 1984; Lalonde 1986; Friedlander and Robins 1995; and Friedlander, Greenberg and Robins 1997); social welfare policy in Sweden (Bratberg

BOX 4 (continued)

et al 2002); welfare, job training, and employment services programs (Glazerman and Levy 2003); conditional cash transfer programs (Diaz and Hanshu 2004); and improving test scores and reducing dropout rates (Wilde and Hollister 2002; Agodini and Dynarski 2001). This literature is reviewed more fully in Glazerman, Levy and Myers, 2002.

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 et al 1985, p. 18).

This does not mean that random assignment studies are a panacea, nor that they should completely replace other studies. Random assignment studies should only be used 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. Sometimes random assignment studies sacrifice generalizability in order to assure that they can properly attribute impacts to an intervention, but this should then be compensated by replicating such studies in different contexts and accumulating a broader body of knowledge about the intervention. Finally, the use of random assignment studies to analyze complex social policies is still developing and researchers need to pay careful attention to the underlying mechanisms and models of behavior that are being tested so as to be sure that the method is applied where it is appropriate.

Well-done random assignment studies are an essential tool for building knowledge on “what works” in social policy and greater investment in such studies is needed.

END BOX 4

As demonstrated below, such random assignment studies are feasible. However, they must be applied where they are appropriate. Random assignment studies are most effective at improving the “internal validity” of a study, that is, assuring that the measured changes among beneficiaries can be attributed to the program and not to other potential contributing factors. They are good at answering questions regarding whether or not a program in a particular context had an impact. They are less effective at “external validity”, that is, at proving that the program would have the same impact in other places or times. Nevertheless, by replicating random assignment studies in different contexts and by carefully documenting the conditions under which programs are started and implemented, a body of knowledge regarding external validity can be established to complement the findings from particular projects.

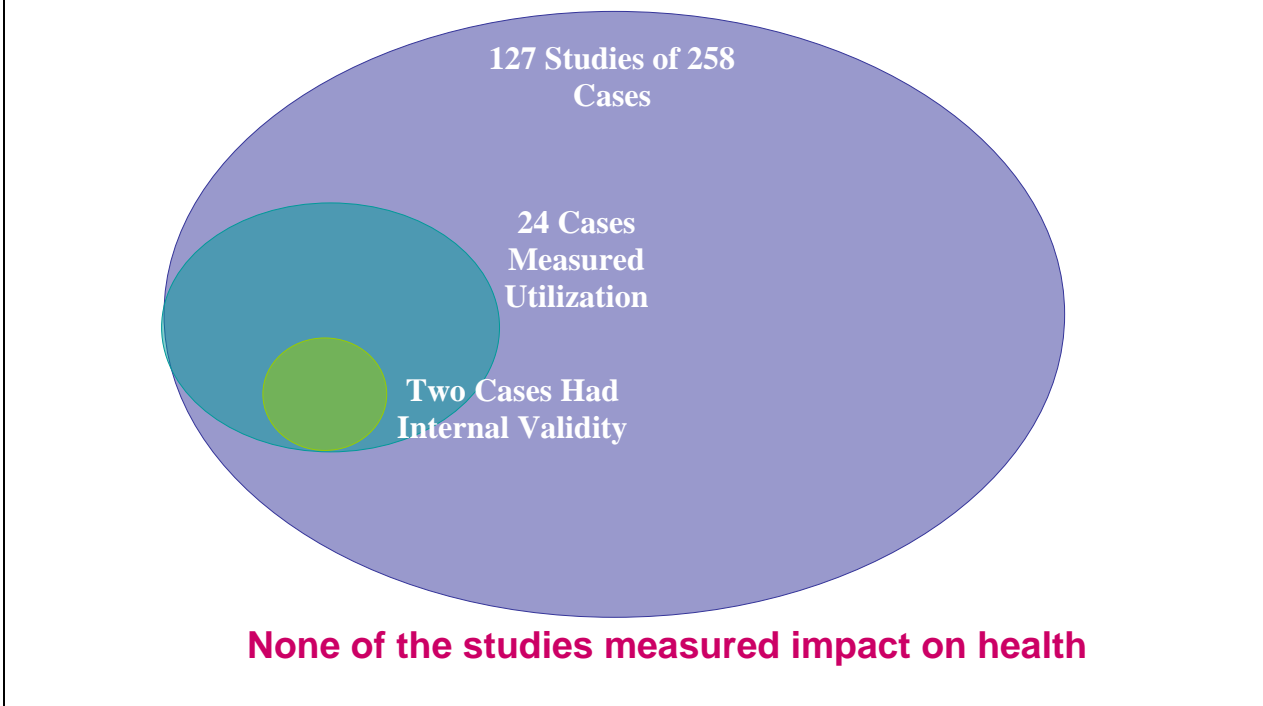
Random assignment studies also require advance planning and careful attention to the process of random assignment and attrition, as well as time and money (See, for example, Victora *et al* 2004 and Altman *et al* 2001). As a result, it is quite easy to neglect impact evaluations or do poor ones, and the number of well-done studies remains limited. Part of the challenge for addressing the Evaluation Gap is to do better studies, while another part is to improve our methods for addressing these limitations.

Too few impact evaluations meet the quality test

Demand for the knowledge that good quality impact evaluations provide is apparent every time a policy advisor or agency staff member commissions a literature review. However, many of these reviews refer to the conclusions of reports and studies without regard to their methodological rigor. When quality is taken into account, such reviews frequently discover that most studies are flawed and that only a few of them actually provide reliable information on impact.

This is the case with studies of community health insurance cited earlier. One review of 127 studies found that only 2 cases that studied the impact of community health insurance on health service utilization had internal validity (See Figure 2). That is, only 2 cases could measure the net impact of the program after accounting for potentially competing explanations. Another review on this topic found that only 5 out of 43 studies that considered the impact of community health insurance on mobilizing funds and improving financial protection for members were able to use statistical controls to support their findings (Ekman 2004).

Figure 2. Many studies lack methodological rigor: Example of Community Health Insurance Programs



Source: Based on information in ILO/Universitas, 2002.

Additional examples include:

- The impact of payment mechanisms on health care 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 *et al* 2004).
- The “What Works” Working Group 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 *et al* 2004).

The very small number of good quality impact evaluations being undertaken, in contrast to the large demand, demonstrates that there is serious underinvestment in these studies. This is not to imply that impact evaluation should be done to the exclusion or detriment of process, operational and other kinds of evaluation that answer different kinds of questions. These other kinds of studies are extremely important for tracking inputs, improving implementation, and making programs more efficient at delivering outputs (World Bank 2004, Scott 2005). But those studies do not ultimately answer the question of whether these better implemented programs are having a positive impact. This is why

more systematic production of good quality impact evaluations on questions of enduring importance are needed in order to inform policy design and resource allocation.

III. Impact evaluation is critical for building knowledge

Unfortunately, evaluations that are done improperly are misleading. They purport to reach conclusions that are actually unsubstantiated. This means the risk of wasting public resources or even harming participants is real. It is for this very reason that clinical trials of medications

have become a standard and integral part of medical care. No one would consider prescribing strong medications without properly evaluating their impact or potential side effects. Yet, in social development programs, where huge sums of money 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 – namely that programs of unknown impact should not be widely replicated without proper controlled evaluation – is frequently dismissed.

This element of the compact will lead us to focus on results and outcomes in a partnership manner.

It's not icing on the cake.

Jean-Claude Faure

Good studies avoid costly mistakes

Findings from impact evaluations can help avoid costly mistakes. For example, an Indian NGO (Seva Mandir) decided to hire a second teacher for their non-formal education centers in the hopes that it would increase attendance and attainment. Twenty-one of the NGO's 42 centers were randomly selected to receive a second teacher. Intermediate indicators – such as the number of days the school was in session – did improve; however, 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 (See Duflo 2003).

The value of impact evaluations in avoiding costly mistakes takes on particularly urgency for programs that are scaled up to a national level. For example, in the US, a program entitled Drug Abuse Resistance Education (DARE) had been adopted in 75% of US school districts because it was believed to be effective; however, evaluations with random assignment demonstrated that the program was ineffective, thereby wasting financial resources and school time (Lynam *et al* 1999, Rosenbaum and Hanson 1998). Similarly, a review of 10 randomized control studies on the policy of “tracking” students, i.e. grouping them by skill levels, showed this approach has little or no effect on student achievement (Mosteller *et al* 1996). Despite the lack of evidence, skill grouping continues to be the most common basis for organizing classes in US middle and high schools.

Good studies can distinguish successes even under adverse circumstances

For those convinced of the efficacy of their programs, money spent on demonstrating impact through comparisons of participants and non-participants often 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 other appropriately controlled groups, they will look like failures.

This ability to distinguish 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, non-participants 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 identify real successes

Unfortunately, it is all too common to find poor evaluations that claim to show a program's positive impact when, in fact, the positive results are due to something other than the program. For example, retrospective studies in Kenya can demonstrate that providing audiovisual aids improved student test scores. But this finding is apparently the result of some unobserved factors because more rigorous randomized assignment studies demonstrate that there is little or no effect (Glewwe et al 2004). Similarly, a US program aimed at assisting poor families through social service visits demonstrated that those receiving the program experienced improvements in family welfare – but so did the families who 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 more promising alternatives.

IV. Impact evaluations are feasible and desirable

Though the value of the information from impact evaluations may be very great, the suggestion that more such evaluations are required is often met with skepticism – if not outright rejection. The main objections are that good impact evaluations:

- are not appropriate for answering questions of importance to decision-makers;
- cannot be ethically implemented;
- are too costly;
- produce results too late to be of use to decision makers;
- do not provide important information about how programs operate; and
- are too complex and do not influence policymaking.

Such objections are not supported by the facts, as we now demonstrate.

Current methods can answer questions that are important to decision-makers

It is difficult to design high quality impact evaluations than can answer policy questions such 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 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 non-participants – who have been randomly assigned to a “treatment” and “control” group – because they relate to how individuals respond to specific interventions:

- Do Vitamin A supplements reduce infant mortality? (Sommer *et al* 1986)
- Do textbooks increase students learning? (Glewwe *et al* 2001)
- Do microfinance programs improve child nutrition? (MkNelly and Dunford 1998)

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 non-formal schools improve attendance and performance? (Banerjee *et al* 2003)
- What are the most effective ways to reduce provider absenteeism in schools and government clinics? (Banerjee and Duflo 2005)
- 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 *et al* 2003 and Kremer 2003)
- Does community monitoring of development programs reduce corruption? Is it more or less effective than audits? (Olken 2004)

It is still feasible, though more resource-intensive, to use randomized assignment 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 affected through the diffusion of their learning to their neighbors? (Conley & Udry 2001)

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 utilization of preventive health care services improve children's health and schooling? (Gertler 2000, Schultz 2000)

Even questions that might be considered quite difficult to answer – such as the impact of gender on political decision-making – can be rigorously studied:

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

In short, the only real limitation on this type of impact evaluation is for addressing questions for which no credible counterfactual can be constructed. But even in these cases, there is usually a series of underlying questions that need to be answered through impact evaluation to begin to find answers. For example, it would be difficult to develop a random assignment study to assess “budget support” (i.e. recent strategies of development assistance that replace project-specific funding with direct financial support to developing country governments and linked to broader policy targets). However, if we had a good evidence base on what were the most effective interventions in particular circumstances (built up from impact evaluations), we would be in a position to indirectly judge the impact of moving from project assistance to direct budget support. This is because we would have a measure of the effectiveness of the programs implemented by the government under the budgetary support scheme. By contrast, in the absence of information about which programs are effective, it is impossible to judge whether a government has made a “better” or “worse” allocation of its budget. More importantly, without good impact evaluations, governments who receive additional budgetary support do not have information on which to base their decisions about how to spend the money effectively.

Impact evaluations can be ethical

Some argue that impact evaluations that rely on collecting data from control groups are *unethical* because they exclude people from program benefits. But this criticism only applies when resources are available for serving everyone as soon as the program starts. In fact, whenever funds are scarce 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 (e.g. school choice in the United States, voucher programs in Colombia), or for transparency and fairness (e.g. random rotation of local governments seats to be set aside for women in the Indian elections).

Furthermore, whenever we have a reasonable doubt of a program's efficacy or concerns with unforeseen negative effects, it is not only ethical but actually an imperative responsibility to adequately monitor and evaluate the impact. This is exactly the premise

underlying the routine use and regulation of medical trials; and it applies to many social programs as much as to medicine.

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

Ignorance is more expensive than impact evaluations

It is also argued that impact evaluations are too costly or difficult. This argument is often made by comparing the cost of an evaluation to the program that is its subject. But the appropriate comparator is not the program cost but *the value of the knowledge it would produce*. For example, evaluations of demonstration training programs in the US and Latin America have sometimes exceeded a third of the initial program costs, but the evaluation results affected decisions regarding 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 actually quite small. When projects are results-oriented and require baseline data, an intelligent design for data-gathering can determine whether or not an impact evaluation will be feasible – sometimes without any additional cost of data collection. Costs are also likely to 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 a random assignment study is the cost of data collection; and the cost of collecting data for a bad study is just as expensive as collecting data for a good one. For example, a large primary education program in India (DPEP) spent millions of dollars collecting data on all the districts in which the program was implemented. But, as noted above, this kind of data collection does not lend itself to a proper impact evaluation. A proper data collection strategy (e.g. randomly choosing which districts would be offered the program and then conducting surveys in a sample of districts that were offered and those that were not offered the program) might have cost less and, most importantly, would have provided useful information about the program's impact (Duflo and Kremer 2003, Duflo 2004).

At a minimum, critics must recognize that the cost or difficulty of good impact evaluations is not a universal fact, but rather one that has to be judged for particular questions and contexts.

Impact evaluations can provide timely information

A fourth critique of the need for impact evaluation studies is that the results are not useful for decisions because they take *too long to produce results*. However, the time taken to produce results depends a lot 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 Pratham (an NGO in India) in time for it to expand a program of community-based teachers who had been shown to be at least as effective as and less costly than new teachers.

It is also possible to design impact evaluations that generate useful feedback during implementation. For example, a multi-year study of the impact of HIV education in Kenya was designed to monitor the long-term impact but also to assess intermediate outcomes, such as the accuracy of knowledge about HIV transmission. The transfer of knowledge is only a necessary, not 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.

More fundamentally, however, it is more important to have accurate information about what programs work even if it takes some years to acquire than to have inaccurate information generated quickly.

Impact evaluations complement other studies

Critics sometimes claim that impact evaluations only tell us whether or not something has an impact without telling us why and how. But a good impact evaluation can provide reliable evidence about the mechanism by 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 it is equally true that the knowledge gained from impact evaluations is a necessary complement to these other kinds of analyses.

Findings from impact evaluation can be simple and transparent

The final critique is that such studies are too complex to be understood by policymakers and do not influence policymaking. In fact, good impact evaluations and randomized evaluations in particular, are relatively easy to present to policy makers with a little work. MDRC, formerly known as the Manpower Demonstration Research Corporation, conducted randomized control trials of numerous state welfare programs in the United States.^{vii} Because the findings were readily conveyed to policymakers, MDRC's studies

had a significant impact on U.S. welfare reform legislation in the mid-1980s (Wiseman *et al* 1991, Gueron 1997, and Gueron 2002).

Other cases where impact evaluation affected subsequent policy can be found in Latin America. 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-90s is widely credited with preserving that social program in the transition to an opposition administration (the program was retained but 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 don't start now, then when will we ever learn?

Most important of all, 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 for AIDS, TB and Malaria, the replenishment of World Bank 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 in the future. 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 today.

V. Why are good impact evaluations so scarce?

If impact evaluations are feasible and useful, why are so few conducted? One important reason is that good impact evaluations have to be designed as part of a new program; yet, in these early program phases, politicians and project managers are primarily focused on designing and implementing their programs. The costs of starting an impact evaluation at this early stage are real and present, while the benefits of measuring the impact will only materialize in 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 required to design and implement studies that could benefit others in the future.

In addition, the outcome of an impact evaluation cannot be known in advance. Hence, those in charge of social programs in developing countries may be unwilling to face the risk of proving their program “failed”; while those involved in international assistance may fear that a few “negative” studies could undermine domestic support for foreign aid more broadly.

Consequently, those who are best positioned to conduct impact evaluations have very few positive incentives to do so, and face many costs and obstacles when they try.

Demand, Money and Incentives^{viii}

The main obstacles or barriers to conducting impact evaluations appear to be: diffuse demand for the knowledge, funding that is not available in a timely fashion, and weak incentives for actors to seek knowledge of program impact.

Demand for the knowledge produced by impact evaluations tends to be spread out across many actors and across time. It happens every time someone in a government, multilateral development bank or bilateral agency asks the question: “What programs are effective at ... ?” This sometimes occurs at the beginning of designing a new program. Other times it arises when producing a document to lobby for additional funding. It also emerges when doing internal reviews of institutional performance.

But it is only at the moment of designing a new program that anything can be effectively done to start an impact evaluation. At that exact moment, program designers want the benefit of prior research, yet have few *incentives* to invest in starting a new study. Ironically, if they do not invest in a new study, the same program designers will find themselves in the exact same position four or five years later because the opportunity to learn whether or not the intervention has an impact was missed (O’Donoghue and Rubin 1999). Since information from impact evaluations is a public good, 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. (See Box 5). Despite the 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. If funding were readily available, it might make the difference between doing or not doing a rigorous study.

Box 5: Timeliness of funds can be critical to good impact evaluations

Box 5: Timeliness of funds can be critical to good impact evaluations

“A ‘rapid-response’ fund to support program preparation is a much needed initiative. The lack of baseline data is critical. For large projects, often the process of preparing the project for approval requires so much time that there is an unwillingness to delay for baseline data collection once the official approval is announced. For example, in the “Familias en Acción” project in Colombia, the project began implementing in some of the communities before the baseline information was collected. A ‘rapid-response’ trust fund could allow data activities to advanced independent of project approval.”

Source: Comments by Suzanne Duryea, Senior Research Economist, Inter-American Development Bank.

Note: for further information on “Familias en Accion” see http://www.ifs.org.uk/edepo./wps/familias_accion.pdf

Other incentives exist at the institutional level to discourage conducting 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 both ways – demonstrating positive or negative impact – any 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 alter, soften or modify unfavorable studies, or to simply repress the results – despite the fact that knowledge of what *doesn't* work is as useful as learning what *does*.

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

Reasons for optimism

Despite these barriers, a number of factors favor conducting impact evaluations. The first is personal commitment from people who recognize the value of impact evaluations, many of whom are employed by governments and international agencies today. A second factor is the growing capacity to collect data and do research around the world. More people 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. Third, there appears to be growing recognition among different agencies of the need for measuring results and that this cannot be done well without complementary studies that get at the issue of attribution. Finally, skepticism about the use of funds also puts pressure on agencies to measure impact. When the overriding risk is closure of a program, then the downside risk of negative findings is less problematic and both managers and project designers see greater benefit in measuring the impact of their programs in the hopes that they can demonstrate that the programs should continue.

VI. Solutions

Existing Initiatives

Concern about the Evaluation Gap is widespread as demonstrated by the many ways that public agencies, intergovernmental commissions, non-governmental networks, research centers, and foundations are addressing it. In particular, initiatives are underway 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

The following discussion of initiatives is by no means comprehensive. Rather, it is presented as a demonstration of the range of existing efforts.

Access to data and information

Numerous organizations are trying to make existing information and data more readily accessible. The OECD's Development Assistance Committee has a searchable Evaluation Inventory of studies done by its member bilateral assistance agencies. IDS, with support from DFID, has a database of studies called "ID-21" (www.id21.org) with an associated strategy for outreach and dissemination via an e-newsletter. Other initiatives aimed at increasing the exchange of existing information include the Development Gateway and the Global Development Network, as well as official channels such as the United Nations Evaluation Forum and the ECG Network (comprising multilateral development banks).

Some initiatives aim to provide access to knowledge by synthesizing the results of many studies on the same question. 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 the medical field but has paid only limited attention to health system policy. The Robert Wood Johnson Foundation also has a "Synthesis Project" oriented toward health policy. The Canadian Health Services Research Foundation is currently analyzing methods for synthesis of social policy studies.

Better data collection

A range of initiatives aim to improve data collection in developing countries through conducting surveys or building local capacity to establish ongoing data collection efforts. Some examples include the Demographic and Health Surveys sponsored by USAID, the Living Standard Measurement Surveys sponsored by the World Bank, the MECOVI program sponsored by the Inter-American Development Bank to support improvement of government statistical offices, and a recent initiative to improve data collection by Paris 21 (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 Operations and Evaluation Department, Canadian Health Services Research Foundation, and many bilateral agencies.

In addition, international efforts are aiming to standardize and systematize the collection and interpretation of indicators, such as the Health Metrics Network, the Child Survival Partnership, and the 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 other. Some agencies and private foundations have also established grant programs that are open to unsolicited proposals (e.g. Canadian Health Services Research, 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. Different agencies in Chile, Kenya, and India 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.

A wide range of research centers, such as the Institute for Fiscal Studies (London), the Instituto Nacional Salud Publico (Mexico), GRADE (Peru), and the International Food Policy Research Institute (Washington) 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 thereby contribute to building a supply of researchers and to promoting an appreciation within public policy debates for evaluation findings.

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 evaluation and that several initiatives are underway to improve the number and quality of such studies. The World Bank's 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 (See Box 2).

Recommendation

Of the initiatives above, the only ones that address the fundamental incentive problems are those that involve internal reform of public agencies – and for most organizations these efforts are infrequent and fragile without some form of sustained external support. Something far bolder, with greater collective support and engagement, is required.

We believe that a desirable solution would have the following characteristics:

- Focus specifically on the public good aspect of impact evaluation and directly alter the incentives for producing good impact evaluations;
- Be a collective response to the problem;
- Mobilize additional funds only to the extent that they are necessary to leverage existing funding;
- Establish a process for assuring and distinguishing high quality impact evaluations from those of poor quality;
- Direct impact evaluation work toward questions of enduring importance and high value to decision-making; and
- Support intelligent and strategic selection of a relatively small number of subject programs from which the most can be learned.

The kinds of impact evaluation studies that should be promoted would:

- Address questions of enduring importance;
- Measure the net impact of a program or policy by establishing appropriate unbiased controls ex ante and utilize rigorous methods;
- Earmark substantial resources for random assignment studies;
- Involve collaboration among program designers, researchers, and implementers from start to finish; and
- Engage policymakers in defining questions and in discussing findings.

A wide range of solutions were considered. The least demanding suggestions involved working through existing institutions and advocating changes in existing practices. Other suggestions required inter-agency and inter-governmental accords, commitments to earmark funding, or the establishment of new institutions.^x One specific idea, for example, was the creation of an entity that would coordinate evaluations across development agencies, principally by establishing common priority questions, and establishing quality standards. The funding of the evaluations themselves would be the responsibility of development agencies. Another idea, one that is elaborated in more detail below, would be to create a facility that would be able to mobilize and distribute resources for independent impact evaluation.

The approach of consultation, coordination, quality assurance and communication has some clear advantages. To some extent, these functions are currently being undertaken by existing bodies, and so such an approach would be relatively straightforward to fund and implement. Under an alternative scenario, which includes the additional features listed below, the Club would mobilize and strategically allocate funds for the design and conduct of impact evaluations. Most, but not all, Working Group members favored this approach, for the following reasons:

- The value of independence: An independently funded evaluation is far more likely to be seen as credible by the range of policymakers and stakeholders who use evaluation results to decide which social programs to support and in what form.

- The benefits of single-mindedness: An institution or group that has the core mission of promoting and generating impact evaluations is more likely to be able to achieve this aim than an institution with many other roles. In donor agencies, for example, whose main business is delivering funding for projects, the operational demands often dominate.
- The potential for shared resources: If every organization funds impact evaluations out of its own resources, inevitably some agencies and governments will have more evaluation funding than others. This greatly limits the ability of smaller and/or less well-resourced governments and agencies to engage in policy debates about what works and what doesn't, to demonstrate the effectiveness of their own programs, and to learn. Pooled resources can level this playing field.

Based on our understanding of the need to fundamentally alter incentives for the conduct of impact evaluations, we therefore propose the creation of a new external and independent institution that would have a variety of coordination and standard-setting functions, but would also directly fund evaluations.

Such an institution would not be imposed on governments and agencies nor should it control their evaluation priorities or studies. Rather, it should be voluntary to assure that (1) it only continues so long as it is demonstrating value to its members; (2) it does not create unnecessary additional burdens for its members; and (3) it does not interfere in internal budget allocations or priority setting. The institution's influence on member organizations should be through information, demonstration, and persuasion.

This is not a proposal for an undertaking by the Center for Global Development, whose mission does not include conducting or managing major impact evaluations. Rather, it is a proposal for consideration by the broad international community of developing country governments, donor agencies, international technical agencies and coordinating bodies, and their constituencies.

An International Impact Evaluation Club

This consultation draft recommends that governments, international agencies, and private foundations establish an Impact Evaluation Club with its own governance, dues, standards and staff.

The particular design of such an Impact Evaluation Club will be determined by its initial members. The following ideas illustrate a potential framework for addressing the incentive problems, the public good issues, and the need for independence and technical excellence. Additional details for how such an institution could be constructed appear in Box 6.

Membership in this “Club” should be voluntary, and the “Club” should have a clear mandate to promote and finance impact evaluations that:

- address questions of enduring importance;
- would not otherwise be conducted;
- are hard to finance in other ways;
- provide models of good practice for emulation; and
- promote stronger evaluation standards.

In promoting good impact evaluations, the Impact Evaluation Club could help developing countries and international agencies to cluster studies around common themes and questions so as to increase their usefulness in learning what interventions are effective and under what circumstances.

Studies with randomized assignment face the largest obstacles relative to their promise in knowledge building, so more than half of the Club’s funds should be earmarked to support studies with randomized designs. The rest of the Club’s funds should be used to support good quality impact evaluations that use other methods.²

The incentive problems are specifically tackled by separating two different decisions: (1) a high level decision by governments, agencies and foundations to dedicate funds for good quality and independent impact evaluations, and (2) a program level decision to conduct an evaluation.

The incentives for high-level decision-makers to seek Club membership would be:

- to leverage funds – because the institution could potentially access more funds than it contributed if its impact evaluations proposals are accepted;
- to participate in a committee that would select “enduring questions” to guide requests for proposals;
- to participate in a committee that would identify potential subject programs for studies based on expected learning; and
- to comply with mandates from their stakeholders requiring implementation of results-oriented management.

Those public program managers, agency staff and policymakers who have an interest in learning from impact evaluations would find that the existence of the Impact Evaluation Club would lower the costs and barriers they face because the Club would:

- provide short-term grants for exploring the feasibility of evaluating a program or collecting baseline data;
- be a well known source for longer term funds dedicated to impact evaluation;
- provide models for good impact evaluation design and implementation;

² The extent to which the “Club” should favor random assignment studies was discussed extensively by the members of the Evaluation Gap Working Group. The consensus was to recognize and support random assignment studies, however the group did not reach a specific conclusion about how much of the club’s funding should be earmarked for random assignment studies.

- give external credibility, legitimacy and continuity to impact evaluation studies;
- act as a link to experts and technical review processes.

The Impact Evaluation Club could improve international learning from impact evaluations in several ways: by setting quality standards, exchanging information among members, and providing long term and substantial funding for impact evaluations.

Quality Standards and Clearing House: The Impact Evaluation Club could develop, or identify existing standards, for good quality impact evaluations. It would then endorse an existing process or establish a process for registering impact evaluation studies and providing independent reviews that would certify to what degree the study met established standards. The registration of studies would serve to inform others of the existing range of studies and to begin to address publication bias.^{xi}

Thematic Selection. Contributing entities would be encouraged to have their representatives participate in periodic deliberations regarding the questions they would like to see addressed by impact evaluations. This non-binding guidance would provide valuable information to the “Club” in soliciting proposals and developing its program of activities. The information exchanged among members and disseminated by the “Club” would also help institutions cluster impact evaluation work around common themes to further the collective interest in learning about common questions and about the generalizability of specific interventions.

Funding for Impact Evaluations. The Impact Evaluation Club would require members to contribute funds and/or to dedicate a portion of their internal budgets to impact evaluation. A member’s contribution would bear some relation to the scale of its development assistance, national budget, or endowment. Although international agencies and donor governments would benefit substantially from the creation of this international fund, the real beneficiaries would be citizens in developing countries to the extent that knowledge from good impact evaluations would be used by policymakers in their countries to accelerate social development.

Developing country governments would be empowered by the Club in several ways. First, they could participate fully as members in advising the Club’s administrative team regarding priority areas for research. Second, as members, they would dedicate some domestic resources for impact evaluations that would accelerate their own learning about which of domestic social programs are effective. Third, they could leverage additional funds to supplement domestic resources for these studies. Fourth, they could set the priorities for which internal programs should be evaluated by choosing which proposals to submit for funding. Fifth, they would get access to expert advice on impact evaluation design. Finally, the studies commissioned under the Club’s guidelines would have international credibility, gaining legitimacy from the review process and transparency created by this independent institution.

Completing good impact evaluations requires a substantial commitment of funding and time. A more detailed study of financing requirements should be undertaken as part of

negotiating the final design of the Impact Evaluation Club; however, prospective members should recognize that some studies might cost as much as US\$10 or US\$20 million over a 7 to 10 year period. An initial estimate suggests that within five years, the Club could be operating with an annual budget of \$30 million, of which 7% would be dedicated to costs of administration and professional networking. Most of this funding would be additional to current spending on impact evaluations in member organizations. In addition, funds that are currently dedicated to evaluation in donor-funded social development projects could be used more effectively with the addition of resources at the margin from the Impact Evaluation Club.

The Club would finance impact evaluations on a competitive basis. Any studies chosen for support by the Club would have to fulfill the following conditions:

- the research design is independently reviewed under rigorous quality standards,
- basic fiduciary responsibilities are fulfilled,
- the final study is submitted for independent review under rigorous quality standards, and
- once approved on methodological grounds, the study must be in the public domain regardless of its findings; and the data must be made available to other researchers.

Strengthen existing initiatives:

As discussed earlier, a variety of initiatives currently address particular aspects of the Evaluation Gap. Many of these are proceeding within existing institutions. However, the creation of an Impact Evaluation Club would complement and strengthen existing initiatives in several ways. By becoming members of the Club, organizations will be able to strengthen their internal efforts at improving impact evaluation by linking them to an external and independent source of standards and credibility. Organizations that are not members would still have access to the registries, guidelines, and studies that the Club would finance. Furthermore, the Club would interact with groups that are trying to improve the quality of impact evaluations and synthesize their results and provide a channel for dissemination and communication.

Specifically, the Club could encourage its members to:

- Fund, promote, and disseminate more synthesis studies via existing channels (e.g. Campbell Collaboration, CHSRF, Robert Woods Johnson). These initiatives address the Evaluation Gap in two ways. They show the value of impact evaluation findings where they exist, and demonstrate the magnitude of the Evaluation Gap where the evidence base is weak.
- Distinguish impact evaluations on the basis of quality. Existing initiatives could be encouraged to propose and disseminate criteria for distinguishing evaluation studies on the basis of the quality of their design, data, execution, methodology and analysis. This would make it easier for policymakers and their advisors to assess the reliability of different studies and would create an incentive for evaluation designers to pay attention to quality standards.

- Create “rapid-response” trust funds to support incorporating impact evaluations during program preparation. The program preparation phase is the ideal moment for program designers, researchers, implementers and policymakers to identify appropriate questions and methods for a proper and useful impact evaluation. Small amounts of funds at this stage can have a large impact on subsequent expenditure and the effectiveness of any future spending on evaluation. Specific funds can be established within existing development agencies or made available externally to provide such rapid-response funding to contract evaluation advisors and collect baseline data.
- Lobby for legislation in middle- and low-income countries to promote more and better impact evaluations of social development programs; and in high-income countries to promote more and better impact evaluation of programs funded by bilateral and multilateral development agencies.
- Support initiatives that aim to increase the capacity for conducting good impact evaluations in low- and middle-income countries.

Hope for the future

Governments, public agencies and private foundations are making progress toward a world with better health, more education and less poverty. But we can reach those goals faster and more effectively by systematically building knowledge about what kinds of social development interventions do and do not work. The recommendations proposed here will not single-handedly achieve this goal, but it can contribute an important, and missing, element to that effort – a way to find out what works.

START BOX 6

Box 6. Illustration of how an Impact Evaluation Club could be configured

The following discussion of how a “Club” could be configured is intended purely as an illustration and as a focus for comments and debate.

Mandate

The Impact Evaluation Club (ImEC) is a non-profit institution whose mandate is to increase learning about effective social programs in low- and middle-income countries through encouraging high quality impact evaluations of social development interventions, particularly through promoting random assignment studies.

To achieve this, most of its funds will be earmarked for random assignment studies; it will create or adopt high quality standards; it will engage with the policy community in all phases of the project cycle and program activities; and it will promote transparency by disseminating the results of all studies along with associated data.

Initial objective: Within 5 years of its inception, the Club will have directly financed 10 good random assignment studies on important questions and supported another 10 good random assignment studies through technical assistance or project preparation funds.

Structure

The ImEC would have the following structure:

A **Board** of 7 members, at least two of whom must be selected on the basis of their technical expertise and knowledge of impact evaluations and methods (including random assignment). The remainder would be appointed or elected by the Club’s Member Institutions. The Board Members would not represent specific members or their interests. Rather they would be charged solely with carrying out the Club’s mandate and holding the administration of the Club accountable for fulfilling the mandate.

A larger **Advisory Committee**, composed of dues-paying Members’ representatives, would be convened to share information about impact evaluation opportunities, to debate the leading questions of the day, and to provide non-binding advice to the Club’s Administration regarding themes and priorities. The Advisory Committee would also assist members in exchanging information about planned and ongoing impact evaluations and encourage clustering of research around questions of common interest.

____ BOX 6 (continued) _____

An *Administration* staffed for the following responsibilities: manage peer reviews of evaluation proposals; make decisions regarding which proposals will be funded; monitor funded studies to assure compliance with the Club's standards and mandate; issue requests for proposals on previously identified themes; raise funds from and provide services to members; collect and disseminate impact evaluations; manage and disburse funds; and submit semi-annual reports to the Board Members on all activities.

Funding Windows

Activities would be financed through at least three "windows":

- (1) rapid disbursement funds for responding to opportunities to begin a random assignment study during a project's design phase
- (2) financing for unsolicited impact evaluation proposals, and
- (3) requests for proposals around enduring questions.

Conditions for Funding Recipients

All studies financed by the ImEC:

- will be registered in a database of studies;
- will be subject to external review of their research design and final reports;
- will be judged on their methodological rigor, the importance of the question they are addressing and the degree of stakeholder engagement in the design and implementation of the study;
- will be publicly available once they have been approved through an external review process, even if the results are "negative";
- will make their data available for re-analysis and replication, with appropriate controls to protect confidentiality.
- will provide sufficient information about context to assist people in assessing generalizability.

Funding

Members will pay dues to join the ImEC. The minimum required dues will vary in proportion to an organization's size and income (e.g. different rates for multilateral development banks, large bilateral agencies, developing country governments, private foundations, and NGOs). Members will be able to contribute as much as they like, above and beyond their dues, as unrestricted funds in support of the ImECs program or as "matching grants" to encourage greater financial participation by others. Innovative mechanisms may be designed – for example, developing country governments might be asked "to contribute" in the form of earmarking internal resources for impact evaluations.

_____ BOX 6 (continued) _____

Benefits of Membership

Priority for funding: Once they have satisfied all the conditions and have passed the review process, research proposals from member organizations would receive priority over proposals from non-members.

Input on “enduring questions” to be addressed through window 3 funding: Members would be able to provide non-binding guidance regarding which questions should be the subject of requests for proposals through participation in the Advisory Committee.

Coordination: Members would be able to exchange information on planned and ongoing impact evaluations and reach agreements on clustering research around particular themes or areas of common interest.

Technical review: Members would have access to technical review services provided in the course of the Club administration’s work.

In addition, Members would benefit from being associated with an organization that has the explicit aim of generating knowledge for better results – a symbol of “good governance.”

Budget Estimates

Administration of the Club will require three operational departments: (1) a financial and operational management office, (2) a grant review office, and (3) a technical services and information office.

The initial budget would require between US\$1 million and US\$3 million per year to cover administrative expenses and to begin funding preparatory work for impact evaluations. *After five years*, however, the Club should be running at its full capacity. At that time, the administrative budget will cover approximately 15 staff members and funding to contract peer reviewers and external technical advisors, along with a full program of impact evaluations.

ILLUSTRATION: Estimates for Administrative Expenses	
Staff	\$1,200,000
Overhead (incl. travel)	\$200,000
Technical Advisors & Peer Reviewers	\$600,000
Total annual budget	\$2,000,000

BOX 6 (continued)

ILLUSTRATION: Estimates for the three financing windows:

Window #1 – Rapid disbursement for preparing studies		
2 “large” studies per year requiring quick disbursement	\$1,000,000	
10 small studies preparation grants (\$100,000 average)	\$1,000,000	
Window #1 Total		\$2,000,000

Window #2 – Funding for studies (steady state annual budget)		
3 new “large” studies per year – ave. 5 years @ \$3mn / year	\$9,000,000	
10 new smaller studies per year – ave. 5 years @ 200,000/year	\$2,000,000	
Window # 2 Total		\$11,000,000

Window # 3 – Requests for Proposals		
5 new studies per year on selected theme - \$3mn/year	\$15,000,000	
Window # 3 Total		\$15,000,000

Total annual grant program **\$28,000,000**

Note: More than half of the total grant program will be earmarked for random assignment studies.

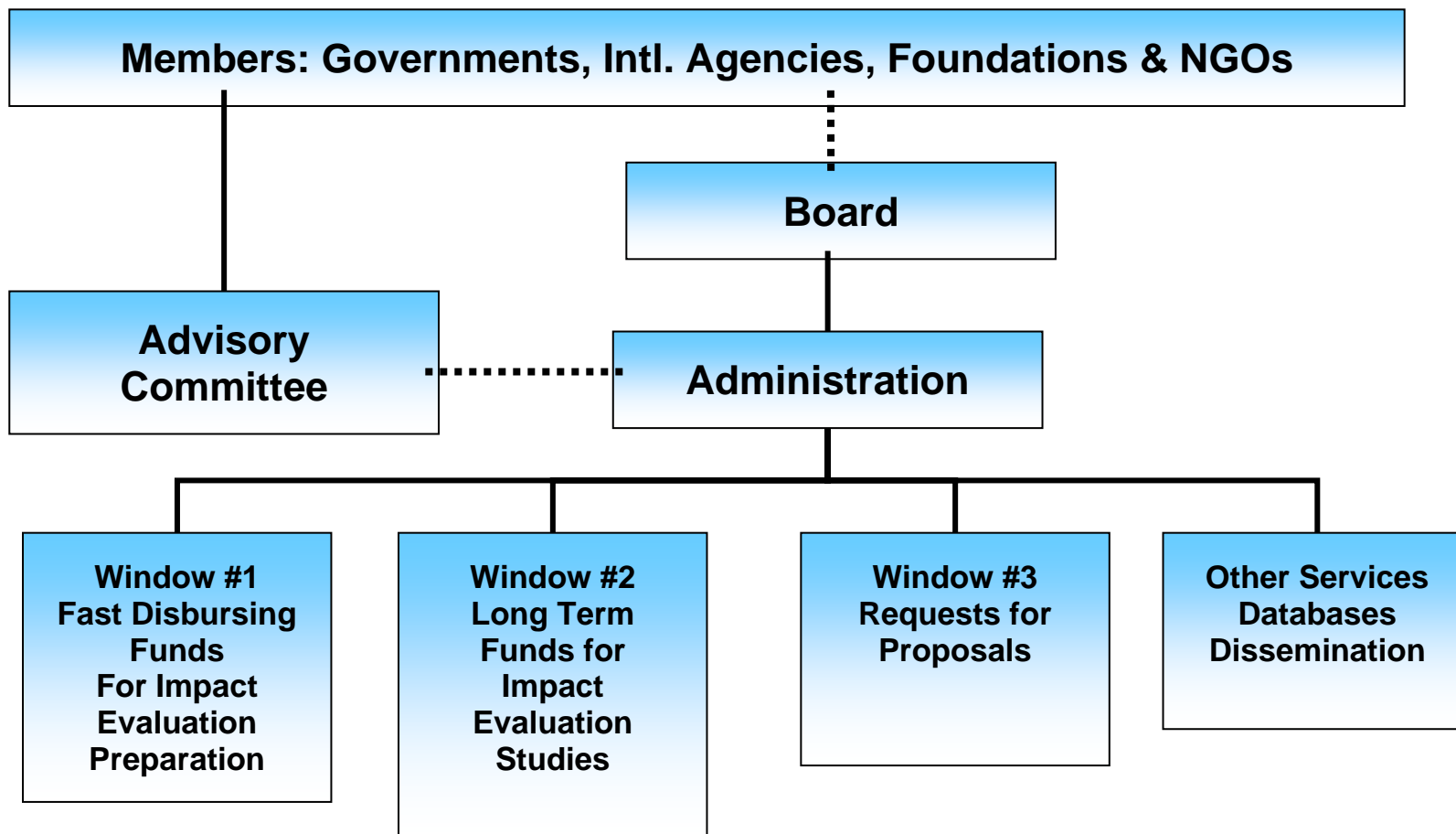
ILLUSTRATION: Annual Funding Sources for a Club

Members paying dues in internally earmarked funds	
Developing Country Governments	\$5,000,000
International Agencies or NGOs	\$1,000,000
Members paying dues in cash contributions	
Donor Countries or bilateral agencies	\$6,000,000
Intl. Development Banks or Agencies	\$4,000,000
Private Foundations	\$4,000,000
Matching Grant Contributions from Private Foundations	\$10,000,000
Total Dues and Contributions	\$30,000,000

ILLUSTRATION: Estimated Income & Expense Summary

Income		Expenses	
Earmarked dues	\$6,000,000	Administration	\$2,000,000
Paid in dues	\$14,000,000	Window #1	\$2,000,000
Contributions	\$10,000,000	Window #2	\$11,000,000
		Window #3	\$15,000,000
Total Income	\$30,000,000	Total Expenses	\$30,000,000

Draft Organization Chart for an Impact Evaluation Club



END BOX 6

References

- Agodini, Roberto and Mark Dynarski. 2004. "Are Experiments the Only Option? A Look at Dropout Prevention Programs." *The Review of Economics and Statistics*, February.
- Altman, D.G., Shulz, K.F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., Gotzsche, P.C., Lang, T., 2001. The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration. *Annals of Internal Medicine* 134, 663-694.
- Banerjee, A.V., He, R. The World Bank of the Future. [013]. 2003. *Bread Working Paper*.
- Banerjee, A.V., Cole, S., Duflo, E., Linden, L. Remedying Education: Evidence from Two Randomized Experiments in India. 9-15-2003. Cambridge, MA, MIT. Mimeo.
- Betsey CL, Hollister RG, Papegeorgiou MR. eds. 1985. *Youth Employment and Training Programs: The YEDPA Years*. Washington, DC: National Academy Press.
- Bryce, J., Victora, C.G., Habicht, J.-P., Vaughan, J.P., Black, R.E., 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, 406-415.
- Chattopadhyay, R., Duflo, E. Women as Policy Makers: Evidence form a India-Wide Randomized Policy Experiment. No. 8615. 2001. Cambridge, MA, NBER. Working Paper.
- Christensen, J. Asking the Do-Gooders to Prove They Do Good. *The New York Times* . 1-3-2004.
- Coase, R H, 1974. "The Lighthouse in Economics," *Journal of Law & Economics*, University of Chicago Press, vol. 17(2), pp. 357-76.
- Commission on Macroeconomics and Health. *Macroeconomics and Health: Investing in Health for Economic Development*. Sachs, Jeffrey D. Report of the Commission on Macroeconomics and Health. 2001. Geneva, World Health Organization.

- Conley, T., Udry, C. "Learning About a New Technology: Pineapple in Ghana". 817. 2000. New Haven, CT, Yale University. Economic Growth Center Discussion Paper.
- Cullen, J.B., B., Jacob and S. Levitt (forthcoming). "The Impact of School Choice on Student Outcomes: An Analysis of the Chicago Public Schools". *Journal of Public Economics*.
- Development Assistance Committee. "Principles for Evaluation of Development Assistance". Organisation for Economic Co-operation and Development. 1991. Paris, OECD. 2004.
- Development Assistance Committee. "Review of the DAC Principles for Evaluation of Development Assistance". Organisation for Economic Co-operation and Development. 1-120. 1998. Paris, OECD. 2004.
- Development Assistance Committee. "Glossary of Key Terms in Evaluation and Results Based Management". 6, 1-37. 2002. Paris, OECD. Evaluation and Aid Effectiveness.
- DFID. DFID Public Service Agreement (PSA) 2003-2006. 2002. London and Glasgow, DFID.
- Duflo, E., Kremer, M. "Use of Randomization in the Evaluation of Development Effectiveness". The World Bank Operations Evaluation Department Conference on Evaluation and Development Effectiveness. 7-15-2003. Washington, DC.
- Duflo, E. "Scaling Up and Evaluation". Paper Prepared for the ABCDE in Bangalore. 1-39. 5-20-2004.
- Dugger, C. "World Bank Challenged: Are the Poor Really Helped?" *The New York Times*. 7-28-2004.
- Ekman, B., 2004. "Community-based health insurance in low-income countries: a systematic review of the evidence". *Health Policy and Planning* 19, 249-270.
- France, M.d.l.d.F.e.d.l. "Partners in Development Evaluation: Learning and Accountability". 3-25-2003. Paris.
- Friedlander D, Robins PK. Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods. *American Economic Review* 1995;85(4):923-37.
- Friedlander D, Greenberg DH, Robins PK. Evaluating Government Training Programs for the Economically Disadvantaged. *Journal of Economic Literature* 1997;35(4):1809-55.

- Gertler, P. 2000. "Final Report: The Impact of Progesa on Health". Washington, DC, International Food Policy Research Institute.
- Glazerman, S., Levy, D.M., 2003. "Nonexperimental Versus Experimental Estimates of Earnings Impacts". *The Annals of the American Academy of Political and Social Science* 589, 63-93.
- Glazerman, Steven, Dan M. Levy and David Myers, 2002, "Nonexperimental Replications of Social Experiments: A Systematic Review. Interim Report/Discussion Paper, Mathematica. MPR Reference No.: 8813-300, September.
- Glewwe, Paul; Kremer, Michael; Moulin, Sylvie; Zitzewitz, Eric. 2004. Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya. *Journal of Development Economics*, 74:1, June, pp. 251-268.
- Glewwe, P., Kremer, M., Moulin, S. 2001 "Textbooks and Test Scores: Evidence from a Randomized Evaluation in Kenya". Washington, DC, Development Research Group. The World Bank.
- Glewwe, P., Ilias, N., Kremer, M. 2003. "Teacher incentives". National Bureau of Economic Research, Cambridge, Mass.
- Gosden, T., Forland, F., Kristiansen I.S., Sutton, M., Leese, B., Giuffrida, A., Sergison, M., Pedersen, L., 2004. "Capitation, salary, fee-for-service and mixed systems of payment: effects on the behaviour of primary care physicians". The Cochrane Library.
- Grossman, J.B., 1994. "Evaluating Social Policies: Principles and U.S. Experience". *The World Bank Research Observer* 9, 159-181.
- Gueron, J.M., 1997. "Learning About Welfare Reform: Lessons from State-Based Evaluations". *New Directions for Evaluation* 76.
- Gueron, J.M. 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*. Brookings Institution Press, Washington, DC.
- Gueron, J.M., Hamilton, G. 2002. "The Role of Education and Training in Welfare Reform". Policy Brief No. 20.. *Welfare Reform & Beyond*. Washington, DC, The Brookings Institution.
- Habicht, J.-P., Victora, C.G., Vaughan, J.P., 1999. "Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact". *International Journal of Epidemiology* 28, 10-18.

- Inter-American Development Bank. 2004. "Progress Report on Management's Actions in 2003 and future actions to enhance the Bank's development effectiveness". CA-456, 1-7. Washington, DC, IADB.
- Inter-American Development Bank. 2002. "Annual Report of the Office of Evaluation and Oversight, 2001". RE-286, 1-30. Washington, DC, IADB.
- International Labour Office. 2002. "Extending Social Protection in Health Through Community Based Health Organizations: Evidence and Challenges". Geneva, ILO. Discussion Paper Universitas Programme.
- Jakab, M., Krishnan, C., Preker, A., Gumber, A., Kelly, A., Ranson, K., Schneider, P., Supakankunti, S. The Impact of Community Financing on Health, Protection Against Impoverishment and Social Inclusion: What do Household Data Tell Us? 2001. Washington, DC, World Bank. HNP Discussion Paper submitted as a Background Report for the Commission on Macro-Economics and Health.
- Jamison, D.T., 1978. "Radio for formal education and for development communication". *Dev Commun Rep* 1-2.
- Kremer, M., 2003. "Randomized Evaluations of Educational Programs in Developing Countries: Some Lessons". *American Economic Review Papers and Proceedings* 93, 102-115.
- Lalonde, R.J., 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review* 76, 604-620.
- Levine, D.I. 2005. "Learning to Teach (and to Inoculate, Build Roads and ...). mimeo, Haas School of Business, University of California, Berkeley.
- Levine, R., What Works Working Group, 2004. *Millions Saved: Proven Successes in Global Health*. Center for Global Development, Washington, DC.
- 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 No. RWP04-036, Harvard University, Cambridge, MA. August.
- Lynam, D.R., Milich, R., Zimmerman, R., Novak, S.P., Logan, T.K., Martin, C., Leukefeld, C., Clayton, R., 1999. "Project DARE: no effects at 10-year follow-up". *J Consult Clin Psychol* 67, 590-593.
- 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*. México.

- Mathematica Policy Research. An Assessment of Alternative Comparison Group Methodologies for Evaluating Employment and Training Programs. Princeton, NJ: Mathematica Policy Research; 1986.
- 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 Programs in Ghana". Freedom From Hunger Research Paper No. 4, Freedom from Hunger, Davis, CA. (www.ffhttechnical.org).
- Miguel, E., Kremer, M. 2001. "Worms, education and health externalities in Kenya". National Bureau of Economic Research, Cambridge, MA.
- Morley, S., Coady, D., 2003. *From Social Assistance to Social Development: Targeted Education Subsidies in Developing Countries*. Center for Global Development and International Food Policy Research Unit, Washington, DC.
- Mosteller, F. and Boruch, R. eds. 2002. *Evidence Matters: Randomized Trials in Education Research*. Brookings Institution Press, Washington, DC.
- Mosteller, F., Light, R.J., Sachs, J.A., 1996. "Sustained Inquiry in Education: Lessons from Skill Grouping and Class Size". *Harvard Education Review* 66, 797-842.
- National Institutes of Health. 2003. "NIH Program Evaluation Guide: How to Develop a Proposal for Evaluation Set-Aside Funding". Bethesda, MD, NIH. 1-57.
- Newhouse, J.P., 2004. "Consumer-Directed Health Plans And The RAND Health Insurance Experiment". *Health Affairs* 23.
- O'Donoghue, T., Rubin, M., 1999. "Doing It Now or Later". *The American Economic Review* 89, 103-124.
- Olken, B.A. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia". 2004. Cambridge, MA, NBER.
- Orr, Larry L. 1996. "Why Experiment? The Rationale and History of Social Experiments", Part I of *Social Experimentation: Evaluating Public Programs With Experimental Methods*. U.S. Department of Health and Human Services. Washington, DC. <http://aspe.os.dhhs.gov/hsp/qeval/part1.pdf>
- Picciotto, R. Economics and Evaluation. European Evaluation Society Conference. 2000. Lausanne.
- Preker, A., Carrin, G., Dror, D., Jakab, M., Hsiao, W., Arhin-Teknorang, D. 2001. "A Synthesis Report on the Role of Communities in Resource Mobilization and Risk Sharing". CMH Working Paper Series. Paper No. WG3:4. Geneva, WHO.
- Pritchett, L., 2002. It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation. *The Journal of Policy Reform* 5, 251-269.

- Quigley, R., Cavanagh, S., Harrison, D., Taylor, L. 2004. "Health Development Agency 2004: Clarifying Health Impact Assessment, Integrated Assessment, and the Health Need Assessment". National Health Service: Health Development Agency.
- Rosenbaum, D.P., Hanson, G.S., 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, 381-412.
- Rossi, P.H., Freeman, H.E., Lipsey, M.W., 1999. *Evaluation: A Systematic Approach*. Sage Publications, Thousand Oaks, London and New Delhi.
- Schultz, T.P. 2000. "Final Report: The Impact of Progesa on School Enrollments". Washington, DC, International Food Policy Research Institute.
- Scott, C. 2005. "Measuring Up to the Measurement Problem: The Role of Statistics in Evidence-Based Policy-Making". London, Paris 21.
- Sommer, A., Tarwotjo, I., Djunaedi, E., West, K.P., Loeden, A.A., Tilden, R., Mele, L., 1986. "Impact of vitamin A supplementation on childhood mortality. A randomised controlled community trial". *Lancet* 1, 1169-1173.
- Stern, N. "The Challenge of Monterrey: Scaling Up". ABCDE Oslo Keynote Speech. 12-7-2002. Washington, DC, World Bank.
- United States Agency for International Development. USAID Child Survival and Health Programs Fund Progress Report. 1-84. 2002. Washington, DC, USAID.
- Viadero, D. Ed. Dept. Issues Practical Guide to Research-Based Practice. Education Week 23[16], 12. 1-7-2004.
- Victora, C.G. 1995. "A Systematic Review of UNICEF-Supported Evaluations and Studies, 1992-1993". No. 3. New York, UNICEF. Evaluation & Research Working paper Series.
- Victora, C.G., Habicht, J.-P., Bryce, J., 2004. "Evidence-Based Public Health: Moving Beyond Randomized Trials". *American Journal of Public Health* 94, 400-405.
- Wilde ET, Hollister RG, 2002. "How close is close enough? testing nonexperimental estimates of impact against experimental estimates of impact with education test outcomes." No. 1242-02 Madison, WI: Institute for Research on Poverty, University of Wisconsin, Madison.
- Wilson, M. 1998. "How Congressional Conferees Can Improve Job Training Reform". *Backgrounder* No. 1203. Washington, DC, The Heritage Foundation.
- Wiseman, M., Szanton, P., Baum, E., Haskins, R., Greenberg, D., Mandell, M., 1991. "Research and policy: A symposium on the Family Support Act of 1988". *Journal of Policy Analysis and Management* 10, 588-666.

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

World Bank. 2001. *Poverty Reduction and the World Bank: Progress in Fiscal 2000*. Washington, DC, The World Bank.

World Bank. 2004. “Influential Evaluations: Evaluations that Improved Performance and Impacts of Development Programs”. Washington, DC, World Bank.

World Bank. 2004. “Monitoring & Evaluation: Some Tools, Methods & Approaches”. Washington, DC.

World Health Organization. 1978. “Financing of Health Services”. Technical Report Series No. 625. Geneva, WHO.

Endnotes

ⁱ The technical definition of a “public good” is a product or service that can be used by many people without being depleted (i.e. “non-rival in consumption”) and whose benefits cannot be restricted to a particular individual or group (i.e. “non-excludable”). The classic example is a Light House – 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 the Light House may not be a good example, see Coase 1974.

ⁱⁱ This scheme draws upon the large literature on this topic, including, inter alia, Rossi 1999, Habicht *et al* 1999, Altman *et al* 2001 and Development Assistance Committee 2002.

ⁱⁱⁱ This point was made explicitly in interviews at the World Bank and with members of the Development Assessment Committee’s Evaluation Network.

^{iv} Interview with S. Quick, Inter-American Development Bank, Washington, DC, September, 2005.

^v Interview with A. Fiszbein and C. Gevers, World Bank, Washington, DC, January 5, 2005.

^{vi} One randomized study in Chicago found that students who were randomly given the option of choosing their school performed better than those who remained in their assigned schools; however, the entire effect was due to the higher level of motivation among the children (and their families) who decided to change schools. A study based on non-randomized methods might have erroneously attributed the difference in performance to the quality of the schools that were chosen (Cullen, Jacob and Levitt 2002).

^{vii} MDRC is a private non-profit organization established in 1974 with support from the Ford Foundation and six US government agencies to assess welfare, training, and education programs. It

^{viii} For a complementary discussion of the obstacles to good impact evaluations, see Levine, D.I. 2005.

^{ix} There is a wide literature on publication bias. With regard to publication bias in clinical trials, see 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):135-146.

^x Working Group members discussed alternative solutions at length. Many favored the solution proposed here while others were not persuaded of the need for a new institution with its own ability to finance impact evaluations. This issue is still under discussion and will be addressed by the Working Group in the future once feedback to this consultation draft has been received from stakeholders.

^{xi} Publication bias results from the greater likelihood that a study will be published if its results are positive (i.e. showing the impact of a program) than if its results are negative. By registering all studies at inception, conclusions from published works can be appropriately qualified.

**When Will We Ever Learn?
Recommendations to Improve Social Development
through Enhanced Impact Evaluation**

Consultation Draft – September 15, 2005

Appendices

Appendix I. Selected Examples of Program Evaluations and Literature Reviews..... 2
Appendix II. Summary of Evaluation Group Deliberations 9

Appendix I. Selected Examples of Program Evaluations and Literature Reviews

Examples of program evaluations that were unable to achieve their goals due to insufficient availability of impact evaluation studies or appropriate information on non-program participants:

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

This project spent approximately Tk 652 million between 1996 and 2001 from international funds (World Bank, UNICEF, Canadian International Development Agency, the Dutch Government, and others) and the Government of Bangladesh to reach 16% of the rural population (approximately 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, underlines (sic) 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, The World Bank: Washington, DC. September 1.

The World Bank committed US\$1.3 billion to the two phases of this program, aimed at improving primary education in disadvantaged communities. The European Commission, DFID, UNICEF and the Dutch government also committed large sums to this program over a seven year period. Ultimately, the program came to serve over 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. “

DANIDA. 2004. “Evaluation, Nepal, Joint Government – Donor Evaluation of Basic and Primary Education Programme II”. Ministry of Foreign Affairs: Denmark.

This study evaluated a nation-wide primary education program in Nepal that involved more than US\$150 million primarily from Denmark, Norway, Finland, the World Bank (WB), the European Union and the Nepalese Government, with smaller contributions from JICA, UNICEF, and the Asian Development Bank.

“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.” (pp 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.” Ministry of Foreign Affairs: Denmark. May.

“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.

[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*. Commonwealth of Learning: Vancouver.

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.”

However, the report itself 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." (p. 73-74).

Appendix I (continued) Reviews of Evaluation Studies

UNICEF Evaluations

Victora, C.G. 1995. "A Systematic Review of UNICEF-Supported Evaluations and Studies, 1992-1993," Evaluation & Research Working Paper Series, No. 3. UNICEF: 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.

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.

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 March 1993, 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.

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).

Agricultural Information Management

Bellamy, Margot. 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 Number 8021, Technical Centre for Agricultural and Rural Cooperation (CTA): Wageningen, The Netherlands. November.

"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." (p. 16).

"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."

Girls' Education

Bernard, Anne. 2002 "Lessons and Implications from Girls' Education Activities: A Synthesis from Evaluations," Evaluation Office, UNICEF, New York. September.

"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). [italics from original]

Appropriate Skill Mix in Health Care

Buchan J. and Dal Poz MR (2002). "Skill mix in the health care workforce: reviewing the evidence", *Bulletin of the World Health Organization*, 80(7):575-80.

... there are extreme limitations to deriving general conclusions and lessons from the available literature in this area. There are four main reasons for this. Firstly, many published 'studies' are, in practice, descriptive accounts, which add little to the evidence base in terms of use of methods or interpretation of results. Secondly, where studies do move beyond description, their utility is often

constrained by methodological weaknesses, or the lack of appropriate evaluations of quality/outcome and cost, or the use of small sample sizes (or all three). Thirdly, with few exceptions, the published analytical studies are derived from the USA, and therefore the findings may not be relevant to other systems and countries. Finally, public bias has to be considered.

The end result is that the end results of some evaluative studies may be suspect, and the results of many other studies are difficult to compare or generalise. ...

Appendix II. Summary of Evaluation Group Deliberations

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

Nancy Birdsall, President, Center for Global Development; Francois Bourguignon, Chief Economist & Sr. Vice President, World Bank; Esther Duflo, Professor of Economics, MIT; Paul Gertler, Professor of Economics, Haas School of Business; Judith Gueron, Visiting Scholar, Russell Sage Foundation; Indrani Gupta, Reader, Institute of Economic Growth; Jean-Pierre Habicht, Professor, Cornell University; Dean Jamison, Senior Fellow, National Institutes of Health; Patience Kuruneri, Senior Policy Analyst, African Development Bank; Ruth Levine, Senior Fellow, Center for Global Development; Richard Manning, Chair, Development Assistance Committee; Stephen Quick, Director, Inter-American Development Bank; William D. Savedoff, Senior Partner, Social Insight; Raj Shah, Senior Policy Officer & Senior Economist, Bill & Melinda Gates Foundation; Smita Singh, Special Advisor for Global Affairs, William & Flora Hewlett Foundation; Miguel Szekely, Undersecretary for Planning and Evaluation, Ministry of Social Development of Mexico; Cesar Victora, Professor, Universidade Federal de Pelotas.

The major areas of agreement among the Working Group members were:

- Too few good quality impact studies are being conducted and made available
- Too much money is being wasted on poorly done studies
- Impact evaluation is best done in collaboration between researchers with knowledge of evaluation methods, executing agencies, and project designers and interacting with policymakers and stakeholders.
- Credibility of studies does not depend so much on being produced by “independent” institutions, rather credibility depends on the quality of the work, its interpretation and its presentation
- The knowledge produced by impact evaluations is a “public good” (in the technical sense)
- Creating incentives to demand impact evaluations would be more effective than improving the supply of impact evaluations
- Existing interest in impact evaluations could be leveraged if funds and technical support were more readily available
- Experimental methods should be used where possible, but they are not appropriate for all questions or contexts
- No single initiative will fully address the Evaluation Gap
- Preference for non-bureaucratic solutions
- Preference for engagement and commitment of major international agencies and governments

A major area of debate was whether the EGWG should indicate a strong preference for randomized control studies. The final deliberations concluded that if the Working Group's mandate were to address issues related to all kinds of evaluation, focused attention on randomized control studies would not be warranted. However, the Working Group's mandate was to focus on impact evaluations and within this category of studies it is apparent that randomized control studies hold significant promise for advancing knowledge, that they are not currently being conducted in sufficient numbers, and that other initiatives are not adequately addressing the need to promote such studies.

Once this was resolved, conclusions were quickly reached on other areas of disagreement. Given the focus on improving impact evaluation by promoting some additional, well-done, impact evaluations – preferably through supporting randomized control studies, it became apparent that earmarking funds and managing them through a new and independent agency would be the best way to proceed. Hence, the decision to recommend creation of the International Impact Evaluation Consortium.

The following individuals were interviewed or provided comments to the EGWG:

- Raj Shah (BMGF)
- Ruth Levine (CGD)
- Catherine Cameron, Consultant, DFID & Agulhas, Inc.
- Max Pulgar-Vidal, Special Advisor, Office of Development Effectiveness, IDB
- Mayra Buvinic, Division Chief, Social Program Division, IDB
- Inder Ruprah, Senior Economist, Office of Evaluation, IDB
- Eduardo Lora, Senior Economist, Research Department, IDB
- Charles Griffin, World Bank
- Gregory Ingram, Director, Office of Evaluation & Development, World Bank
- Charles Sherman, National Institute of Health (NIH)
- Carol Peasely, Counsellor, USAID
- Carol Lancaster, Prof. GWU (formerly with USAID)
- Patrick Kelley, Director, IOM Board on Global Health
- Eduardo Gonzalez Pier, Ministry of Health, Mexico
- Paul Gertler, Professor, Berkeley
- Jeremy Hurst, OECD
- Julio Frenk, Minister of Health, Mexico
- Stephano Bertozzi, Berkeley & Institute of Public Health (Mexico)
- Ricardo Hausman, Professor, Kennedy School
- Tom Bossert, Professor, HSPH
- Rachel Glennerster, Director, Poverty Action Lab, MIT
- Abhijit Banerjee, Professor, Poverty Action Lab, MIT
- Bernhard Schwartlander, GFATM
- Richard Feachem, GFATM
- Elizabeth Docteur, OECD
- Daniel Klagerman, Ministry of Economy, France
- Richard Manning, DAC
- Hans Lundgren, DAC

- Paul Delay, UNAIDS
- Ties Boerma, WHO (and Health Metrics)
- Francois Bourguignon, Chief Economist, World Bank
- Stephen Quick, Manager, Office of Evaluation, IADB
- Jim Heiby & Karen Cavanaugh, USAID
- Calestous Juma, Harvard
- Rob D. van den Berg, Global Environment Facility
- Michael Schroll, WHO
- Binh Nguyen, Asian Development Bank
- Ariel Fiszbein, World Bank
- Coralie Gevers, World Bank
- Owen Barder, Center for Global Development
- Neils Dabelstein, DANIDA

The EGWG process and findings were discussed at the following meetings:

Health Metrics Network. Meeting at Johns Hopkins University, Baltimore, MD. May, 2004.

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

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

Global Fund for Aids TB and Malaria. Nov. 8, 2004. Geneva, Switzerland.

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

2ème conférence AFD / EUDN. « Aide au développement: Pourquoi et Comment. Quelles stratégies pour quelle efficacité? » Nov. 25, 2004.