The Evaluation Agenda

by Ruth Levine and William D. Savedoff

The Bank's Success Depends on Knowledge

We will start with an obvious point: To succeed as an institution, the World Bank must succeed in its main business. Its main business is financing projects and programs that lead to better economic and social conditions than would have occurred without those projects or programs. A higher—and technically superior—definition would require that the returns for these projects and programs be at least a little better than their true economic costs. And a still more demanding standard might ask that the projects represent the *best* (most cost-effective) of all possible ways to achieve the same ends. But let's not be fussy here; let's just stick to the basic message that the Bank succeeds when poor people's lives improve because of the funding, technical expertise, accountability requirements or other dimensions of the Bank's lending and other instruments.

In contrast, the World Bank's success cannot be measured on the basis of whether the institution remains solvent, gets along well with NGOs, keeps employees happy, or fights corruption in-house and abroad. These are all probably necessary, but they're not sufficient. The Bank's success rests on whether it can make the lives of those who are sometimes referred to as the "ultimate beneficiaries" better off, in a meaningful way.

Ruth Levine, Center for Global Development

Ruth Levine is the director of programs and a senior fellow at the Center for Global Development. She is a health economist with 16 years of experience working on health and family planning financing issues in Latin America, Eastern Africa, the Middle East, and South Asia. At CGD, she manages the Global Health Policy Research Network. Before joining the Center, Ruth designed, supervised, and evaluated health sector loans at the World Bank and the Inter-American Development Bank.

William D. Savedoff, Social Insight

Bill Savedoff is a senior partner at Social Insight. In addition to preparing, coordinating, and advising development projects in Latin America, Africa and Asia, he has published books and articles on labor markets, health, education, water, and housing. He currently serves on the editorial board of *Health Policy & Planning* and as an editorial advisor to Transparency International.

Whether the Bank's projects and programs help borrowing governments to achieve good results for their citizenry depends in part on whether the programs are designed and implemented well. Take, for example, the case of a Bank-financed project aimed at improving enrollment, retention and learning outcomes of primary school students, which directs financing toward school construction, curriculum development, teacher training and new information systems in the Ministry of Education. Whether the project will achieve the desired results depends on whether the various project activities, from contracting for civil works to developing manuals for computer users, are conducted in a timely, cost-effective manner that takes into consideration local conditions. It also depends on whether the problem of low school attendance and performance can be solved with buildings, teacher skills, textbooks and computerized enrollment records. One would imagine that guidance on both of those questions, if not definitive answers, are within reach. Given the Bank's base of institutional experience—more than 50 years, across more than 100 countries and every sector, with billions of dollars of investments—a ready reserve of knowledge about what works should be available to inform critical design and implementation decisions. Indeed, it is just this type of asset that inspired the notion that the Bank could be a "knowledge bank."

The reality is quite distant from this idealized expectation, as any candid Bank employee will attest. The Bank appears to be structurally and perennially unable to learn.

The Bank Creates but Does Not Use Operational Knowledge

For operational questions, the Bank has shown itself to be reasonably good at generating "lessons"—but a mediocre student when it comes to applying them. The World Bank's Implementation Completion Reports (ICR) (prepared by staff or consultants at the conclusion of every project) and the broader sector studies generated by the Independent Evaluation Group (IEG), previously called the Operations Evaluation Department (OED), are replete with hard-won operational lessons, which are conveyed to the Bank's Board of Executive Directors in confidential documents. Classic and oft-repeated ICR conclusions include the inadvisability of establishing project management units that are isolated from line ministries; the importance of political commitment, managerial continuity, and timely follow-through when problems are detected; the need for operational research to inform mid-course corrections; the benefits of focused rather than multi-component "Christmas tree" investment programs; and the importance of developing a realistic financing strategy for the recurrent costs associated with the program. Though the same mistakes may be repeated from project to project, there's no doubt they are documented in detail each time.

The fact that these conclusions are oft-repeated is testimony to the limited impact that their documentation has on Bank practices and procedures, although they are ritualistically invoked at particular moments. It is striking, in fact, that one can often find essentially the same "lessons" in both the design document justifying the funding for a program and in the report after the funds have been spent. The design document may say that a "lesson learned" from similar operations is that project activities should be clustered so that the newly trained teachers are working in the rehabilitated schools that have the additional textbooks. Then, the ICR for the same project, seven years later, may say that results were disappointing because the project had to disperse investments widely, to maintain political support.

The reasons for this lack of learning about even the operational basics are many, and include everything from the extreme time pressure on staff, to the limited funding for disseminating the ICR results in a meaningful way, to the underlying incentives that result in oversized, unwieldy, unrealistically ambitious projects. Essentially no management attention is given to the sharing and application of this knowledge; and ICRs tend to be seen as bureaucratic by-products that yield no benefits to line managers or those who design and supervise the implementation of Bank projects.

The Bank Rarely Creates New Knowledge about What Works

While the Bank at least documents the operational lessons, it seldom generates the right kind of technical knowledge, or knowledge about what really works to achieve the desired impact. Technical lessons would come from analyses of how well similar projects achieved their aims in the past and would answer very basic questions that are at the core of project designs: What are the most effective (and cost-effective) ways to get girls to complete secondary school in rural Africa? What AIDS prevention strategies work to reduce the incidence of infection among mobile populations? Under what conditions do road-building projects reduce rural poverty?

On these sorts of questions, and the generation of knowledge about what works, the Bank's track record has been as wanting as virtually all other development institutions. It has systematically failed to even attempt to learn from one project or program how to get more and better results the next time around. Moreover, it has rarely undertaken and shared the type of data collection and analytic work that would contribute much needed light to the darkness of development assistance more generally.

The type of knowledge needed comes from impact evaluations, defined as evaluations that measure the results of an intervention in terms of changes in key variables (e.g., mortality, health status, school achievement and labor force status) that can be credited to the to the program itself, as distinguished from changes that are due to other factors. That is, they are evaluations that permit attribution of program-specific effects. At the Bank, as in the field of development more broadly, much emphasis has been placed on monitoring project performance and comparing before- and after-project conditions, while insufficient investments have been made in conducting rigorous impact evaluations that are necessary to tell us which interventions or approaches do and do not work.

This underinvestment in impact evaluation (and consequent undersupply of evidence about the relationship between specific types of investments and their effects) has a major, if painfully obvious, result: If we don't learn whether a program works in changing the well-being of beneficiaries, how do we know it's worth putting the money and effort into similar programs? If we don't bother to measure the results that are direct consequences of the specific program, how can we make a credible case for this, or any other type of expenditure of public funds?

This is the current scenario, lacking accumulated knowledge from impact evaluations. The typical World Bank social sector project is designed with a narrow range of inputs, sometimes generated by World Bank staff and consultants, and sometimes by the government receiving the loan or credit: a very detailed description and analysis of current (bad) conditions; guesses about the reasons for those bad conditions; a theory of change, or argument, that says "if you make these particular sorts of investments and institutional changes, the world will be a better place in these specific ways: fewer children will get sick and die, more children will go to primary school and learn something that will permit them to make a living in the future, and so forth. And not only will more of these good things happen because of the program's investments and institutional changes, but those good things would not have happened—or would not have happened so quickly—in the absence of this program."

Importantly, the dependence on this sort of argument is central to even broad "country-driven" programs that look much like budget support— for example, the Poverty Reduction Strategy Credits (PRSCs) are based on the notion that if you give the equivalent of block grants, or credits, to countries, they will allocate the resources in ways that reduce poverty. So while the Bank may not micromanage or "projectize" the spending, the successes of PRSCs or other forms of budget support is contingent, eventually, on the success of government decisions about how to spend those resources on public health, education and many other types of programs intended to reduce poverty and improve the life chances of the poor.

What is missing as an input into the design of most programs is a genuine evidence base to systematically support (or refute) that theory of change. Will those particular investments and institutional changes really make a positive difference, or do they just sound good? Have those investments resulted in the desired change before, in the same country or region, or elsewhere? We simply have no systematic information about this outside of a very narrow set of experiences that, primarily because of historical accident, have been well evaluated (e.g., conditional cash transfers in Mexico and Central America).

Why So Little Impact Evaluation?

Good impact evaluations are not a core part of the practice of international development. There are lots of very good reasons for this:

First, good impact evaluations require a degree of technical sophistication that is often lacking in the field of "applied development," where practitioners are accustomed to dealing with poor data and unfamiliar contexts. While many studies compare conditions before and after a project, such comparisons can be quite misleading without attention to other factors that might have also contributed to observed changes. For example, improvements in population health status might come from the introduction of new health care services, but they might also have been induced by rapid economic growth, migration, personal hygiene, climate change, or investments in infrastructure. Only by comparing observed changes among those who benefited from a project to some other control group is it possible to begin to disentangle how much of the effects can be attributed to the project or program itself.

Separating out the changes due to projects from changes due to other things is a complicated business, and to date the development community has been satisfied with weak alternatives, viewing more rigorous methods as inappropriate to the context of developing countries. Fortunately, advances in research methods and increasing capacity around the world to conduct such impact evaluations is beginning to surmount these technical difficulties.

Second, demand for the knowledge produced by impact evaluations tends to be spread out across many actors and across time. 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 do want the benefit of prior research, yet have few incentives to invest in starting a new study. Paradoxically, if they do not invest in a new study, the same program designers will find themselves in the same position four or five years later because the opportunity to learn whether or not the intervention has an impact was missed.[1] Because 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.

Third, 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. Because 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.[2] 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. Similar disincentives to finding out "bad news" about program performance exist within institutions like the World Bank. For task managers, in fact, attempting to communicate negative results up the managerial "chain of command" can be one of the least career-savvy moves one can make.

Fourth, evaluation simply is not the central business of the Bank, and when material and human resources are stretched—as they typically are, even in the comparatively well-endowed environment of the Bank—short-term operational demands will override the longer-term, more strategic imperative of evaluation and learning. As one indication, resources spent to design and implement impact evaluations were not even recognized as a separate item in the World Bank's budgeting system until this year.

All of these reasons contribute to the situation observed today. For most types of programs, a body of scientific evidence about effectiveness is lacking. Bank task managers designing projects are left to their own devices. The general strategy is to observe that many other projects are based on the same theory of change, and on a plethora of anecdotes, "best practice"-style documentation to support a given program design, and reference to the writings of those who are regarded as particularly brilliant thinkers. This is "eminence-based decision making" rather than "evidence-based decision making."

A Smarter Future

Fortunately, many have recognized this problem, care about solving it, and are trying hard to find a way to do so.[3] Within the Bank, the IEG advocates for more resources for good evaluation, and makes heroic efforts to squeeze knowledge out of the experiences of projects that are conducted without baseline data, without comparison groups, sometimes without any impact indicators at all. In the past couple of years, the World Bank has created an initiative called the Development IMpact Evaluation (DIME) Initiative to: increase the number of Bank projects with impact evaluation components, particularly in strategic areas and themes; to increase the ability of staff to design and carry out such evaluations; and to build a process of systematic learning on effective development interventions based on lessons learned from those evaluations.[4]

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. The Bank is also aiming to improve internal incentives to undertake more systematic development impact evaluations by explicitly recognizing these studies as a valued product in their own right.

This represents an important shift in the Bank's recognition of the value of impact evaluation—and particularly in the leadership of key individuals who have taken on this topic as a personal mission within the institution. The work of both IEG and the DIME deserve political and financial support.

But the chances are that this will not be enough. Even the best intentioned efforts, such as the DIME, will find it difficult over time to sustain their resources and maintain enthusiasm and rigor—particularly when some evaluations will inevitably show that many programs have been unsuccessful. In most institutions, internal offices that generate negative reports are subject to pressures to paint results in a positive light or, over time, find themselves increasingly isolated and with fewer resources.

A broader and bolder solution to the problem is required. Three central elements are required for a lasting and genuine solution to the problem of lack of knowledge about what works.

First, we need to use good evaluation methods to get answers to important questions. This means identifying the enduring questions, a process that would be done best if it were done in true partnership between developing countries and the range of institutions that provide development finance. The World Bank has made a start by identifying five thematic areas within its impact evaluation initiative. But the benefits of concentrating such studies around enduring questions across agencies and countries would be even greater. Surely there is an immense opportunity to learn by collaboration with different organizations that address similar health, education and other social problems in profoundly different ways.

Second, we need to use evaluation methods that yield answers. This means increasing the number of impact evaluations that use rigorous methods—such as random assignment and regression discontinuity—and applying them to a small number of programs from which the most can be learned. This does not obviate the need to continue process-oriented evaluation work, which can be tremendously informative to answer operational questions, but it does mean there is a new and large agenda for impact evaluation.

Third, while the overall set of important questions should be developed by the “interested parties” in development agencies and developing countries, the impact evaluations themselves need to be done independently of the major international agencies and borrowing country governments. Independent evaluations would be more credible in the public eye, and less subject to inappropriate pressures to modify results, interpretations or presentation. It is still important to work on changing the culture of the Bank in its entirety and all the myriad internal incentives to get projects done and implemented so that evaluation and learning become a regular part of all the Bank’s activities. But the existence of an independent source of impact evaluation results—geared to a longer time frame and toward learning—will avoid many of the inevitable pressures to restrict the communication of bad news to higher levels of management.

A Proposed Solution

If leaders with vision in a few development agencies and a few developing country governments put their minds to it, a major improvement is within grasp. An international initiative could be established to promote more and better independent impact evaluations, undertaking, for example:

- Development of a shared agenda of "enduring questions" for selective evaluations around which multi-country/multi-agency evaluation could be done.
- Creation and dissemination of standards for methodological quality of impact evaluation.
- Provision of financial resources for design of impact evaluations.
- Provision of complementary resources for implementation of impact evaluations.
- Creation of a registry of impact evaluations.
- Dissemination of impact evaluation results.
- Development of a data clearinghouse to facilitate reanalysis.
- Support for the development of new and improved methods.

Appropriately, this would be a collective response to ensure supply of knowledge, a global public good in the truest sense. The ideal financing arrangements would be one based on sharing costs across those who benefit, or at least those agencies that choose to participate. Those resources should be additional to the current evaluation budgets, which have been pared down to subsistence level. Foundations and other private sector actors who see the long-term benefits and wish to facilitate generation of and access to knowledge might also be willing to provide start-up resources.

The question remains whether the Bank, which uses so little of its internally produced knowledge from ICRs and products of the IEG, would develop mechanisms to apply technical knowledge generated with the input of an independent facility. The answer to this may matter relatively little. If such knowledge were part of an international effort that disseminated findings broadly, those with whom the Bank works—counterparts who co-design projects with Bank staff—might find themselves well equipped with much more evidence about what works than they have today, and be able to shape the Bank's actions in a positive way.

The World Bank's participation in such an initiative would be truly win-win.The international community would benefit from the institution's expertise and access to knowledge from the Bank's tremendous portfolio of projects.The Bank would benefit from enhanced credibility and influence, as well as access to knowledge from other agencies' projects. Participating actively in the global process of learning what works is a natural role for the Bank to take on. Genuine success—making lives better—depends on it.

Notes

1. Ted O'Donoghue and Matthew Rabin "Doing It Now or Later," *The American Economic Review 89* (1999): 103–124.

2. Lant Pritchett, "It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation," *The Journal of Policy Reform 5* (2002): 251–269.

3. This section and those that follow are derived from discussions of the Evaluation Gap Working Group, and its final report, "When Will We Ever Learn? Improving Lives through Improved Impact Evaluation" (Center for Global Development, 2006). The ideas included here, however, are those of the essay authors, and not necessarily those of all working group members. For more information, see http://www.cgdev.org/section/initatives/_active/evalgap.

4. Information on DIME from personal communications with Francois Bourguinon, Paul Gertler and Ariel Fiszbein.