Impact Evaluation 2.0

Chris Blattman[*]

Presentation to the Department for International Development (DFID)

London, UK

14 February 2008

We tend to associate randomized trials with new drugs for blood cholesterol and baldness, not with school textbooks, mosquito bed nets, and police patrols. But a growing number of development agencies are beginning to change our minds; they are assessing the impact of humanitarian aid and development programs with the randomized control trial.

I am going to assume that this fact is already a familiar one. For convenience, this talk will actually presume a great deal: that you appreciate the case for evidence-based programming; that you understand the nature of a randomized trial; and that you believe the ethical and political concerns with randomization are important but in many cases can be overcome.[†]

My goal is not to make the case for randomized evaluation. Rather, I want to challenge the conventional wisdom. In particular, I want us to be more ambitious, more aware of the needs of implementers, and more forward-thinking in our approach to evaluation. Much of what is being done in evaluation today is very good, but I will argue we can do better. The hard work and ingenuity of researchers and implementers the world round have pushed us to a point where we are ready for a revision of evaluation practice: an impact evaluation 2.0.

The majority of my comments will concern randomized control trials of fairly specific development interventions. In other words: experimental impact evaluations. This scope is deliberately narrow, and I will fail to satisfy those interested in the big picture—say, the effectiveness of foreign aid overall. I will also not say much about qualitative evaluation methods, however valuable and instrumental their role.

My comments are applicable, however, to a broader range of evaluation methods that use non-randomized comparison groups—that is, non-experimental evaluation. I will return to this point

[*] Yale University & Center for Global Development, cblattman@cgdev.org.

[†] See the Center for Global Development publication When Will We Ever Learn? for the case for randomized evaluation and evidence-based aid.

towards the end, when I argue that reports of the death of non-experimental evaluation methods are greatly exaggerated.

All impact evaluations ask a simple question: what would the individual (or village, or region) have looked like in the absence of the intervention? This alternate reality is, of course, impossible to observe. Evaluators thus take the counterfactual approach; a relevant comparison group is observed and compared against program beneficiaries.

An impact evaluation is only as good as the comparison group, however, and for that reason random assignment to beneficiary and control groups has become the method of choice. Only random assignment ensures that the control group is, on average, indistinguishable from the beneficiaries.

There has been an enormous increase in excitement for randomized control trials. Economists like Abhijit Banerjee and Esther Duflo of MIT have called for all possible development aid to be evaluated experimentally.

Other economists like Angus Deaton of Princeton have cautioned that randomized control trials have all the characteristics of a fad, and are bound to disappoint. Program impacts are so contextual (goes one such argument) that the results from one trial may be of little assistance to decision-makers in other times and places.

One could call this the battle of the Randomistas versus the Relativistas.

I will argue today that the truth is not only between these two poles, but also beyond them. The kind of evaluation on which they focus, what I call impact evaluation version 1.0, delivers exactly all of the gains promised, and suffers all of the feared drawbacks. We can do better with a version 2.0.

But first, what do I mean by impact evaluation 1.0?

In essence, impact evaluation 1.0 provides a cost-benefit analysis—what I will call return on investment (ROI) figures–for alternative interventions. Such ROI figures improve upward accountability to donors, enhance strategic decision making, and can even boost fundraising.

The most straightforward impact evaluations calculate an ROI by compare program to no program. One example is the study of de-worming medicine on Kenyan primary school students by economists Ted Miguel and Michael Kremer.

Another productive approach is been to focus on *relative* ROIs: the effect of program 1 versus program 2 versus program 3 versus no program. In a series of experiments in western Kenya, Michael Kremer and various co-authors have assessed the impacts of various primary school interventions, looking at the size of impacts as well as their equity.

A final, and supremely important, approach is the evaluation of program 1 versus no program in context A, context B, context C. Such multi-country studies not only validate relationships from a single place or pilot, but also enable us to calculate an *average* ROI—an average we might generalize with more confidence. A good example of this approach is the World Bank's recent push to evaluate human development programs—such as youth vocational training—across nearly a dozen countries. Another recent example is the set of studies in Kenya, Uganda, and South Africa that demonstrated the dramatic effect that male circumcision has on reducing HIV transmission.

All these findings have been unquestionably valuable. Nevertheless, version 1.0 has several potential limitations.

First, consider that the average ROI will not be helpful if the variability in returns is greater than the average return itself.

Fast forward, if you will, to 2015, when there will be dozens upon dozens of education and health impact evaluations giving us average ROI figures for interventions from textbooks to scholarships to vocational training. It is extremely possible that in some contexts we will see a particular intervention yield 100 percent improvements, in some 30 or 40 or 50 percent improvements, and in some much less than that.

If so, we may find ourselves in an uncomfortable position: the average ROI of textbook provision may be statistically indistinguishable from the ROI of another program, like school meals, simply because of the variability of the impacts.

I fear the current ream of impact evaluations will yield one overwhelming result: how and where we implement is more important than what we implement. Performance is essentially conditional on context and processes.

The backlash against expensive experimental impact evaluations, the disillusionment and bitterness, could be significant.

There is a second potential limit on the hopes of impact evaluation 1.0: the average ROI that we identify for particular programs may also be overestimated by the non-reporting of negative or zero-results.

I am not the first to note that there is a bias in publishing positive results. Papers showing insignificant or noisy results often do not get published, or simply go unwritten.

Moreover, programs may be stopped midway when the lack of progress is evident. Or follow-up surveys are not conducted when the conclusion is already foregone.

If so, these small, zero, or negative results will not enter into our ROI average. The fabled decision making tool will be highly variable and biased.

A third potential problem impact evaluation 1.0—and this is probably the most important critique I want to make—is that it is not a system of performance management or process learning, and thus of too-limited use to implementers.

The problem is simple: even when we exactly identify relative ROIs, we are often in no position to say specifically who or what is responsible for a positive or negative impact.

The basis of true accountability, learning, and improvement is knowing in a timely fashion what people, processes, and program factors were effective in achieving the result. Likewise, knowledge of bottlenecks and barriers is also crucial.

Impact evaluation 1.0 provides only one form of accountability: upward accountability from implementers to donors. As a tool for operational decision-making, ROI is severely limited.

The simple fact is that a well-run program with mediocre average returns could have more consistently positive development impact than a program with high returns but high variability.

The question is how to reveal the causal mechanism? How to learn where to fine tune a process?

A fourth but smaller problem is that the results from impact evaluation 1.0 are generally slow in arrival. To the extent that timeliness is crucial to improving performance management and decision-making, we ought to focus on what evaluation methods deliver faster feedback.

Finally, version 1.0 evaluations typically focus on evaluation programs or program components. These program ROIs are useful, but they represent only a small range of the objectives that are important to implementers and activities that can be evaluated. I'll say more about these alternatives in a moment.

What, however, is impact evaluation 2.0?

Version 2 goes beyond the simple evaluation of programs and program components. One of the least explored frontiers in impact evaluation is program targeting: how to reach the poorest and most vulnerable? Another unexplored frontier is program scaling. Both are crucial concerns to implementers, especially anti-poverty agencies and emergency and humanitarian organizations.

More importantly, however, impact evaluation 2.0 brings us closer to the goal of performance management rather than simply upward accountability.

Version 2 evaluations try to understand why a program works, and what it reveals about the process of development. That is, they try to understand the causal mechanism.

A simple way to do so is to look at impact heterogeneity. What do I mean by this term? Well, a program affects different individuals in different ways. For instance, a micro-finance program may be more beneficial to women with social networks, to women with access to markets, or to women with education. By taking care to measure such initial characteristics we are able to partially unpack the determinants of success and the reasons programs work or fail. A DFID- funded microfinance evaluation by BRAC in Bangladesh is essentially doing just this.

A more formal way of unpacking the causal mechanism, however, is to vary the program components themselves experimentally. For instance, many micro-finance programs are a mixture of skills training, access to capital, and social network creation. These components can be mixed and matched, or provided in different combinations, to understand the marginal contribution and productivity of each factor.

For instance, I am planning a women's economic empowerment program in post-conflict northern Uganda. The objective is to help women increase their incomes, their social status, and their community integration and participation by building small businesses. The NGO has been providing women with small grants and training, and anecdotally have seen much success. The most successful women, however, appear to be those with strong business and social networks.

To assess this belief, and to see if NGOs can effectively build social capital and business networks, half of the beneficiaries in the program will be formed into groups with more experienced women, and will be encouraged and enabled to meet and exchange information and experiences regularly. In this way we hope to learn more about developing enterprises, developing social networks, and their impacts on empowerment more generally.

Another related feature of version 2 evaluations is that they recognize that program impacts are *conditional*. It is likely that impacts are conditional above all on processes, management and governance. Hence it is these processes, management, and governance that we should seek to evaluate further.

Let me give an example. In another program in northern Uganda, youth groups are receiving cash from the government to undertake vocational training and buy tools and start-up materials. As far as we can tell, however, the most important determinant of their success is appropriate decision-making, and the most important determinant of that is effective community facilitation and extension services.

As a result, the variability in performance is great in proportion to the average performance. What we want to evaluate is not simply provision of the program, but alternative incentive, funding and control systems that lead local government to provide quality extension services to the youth groups. This is what we are randomizing within the group of program beneficiaries.

Evaluation 2.0 is less about whether vocational training is a superior investment to textbooks, or textbooks a superior investment than bed nets. It is about how to do vocation training better. How to do textbooks better. How to do bed nets better.

This approach to evaluating alternative components and processes has a number of potential advantages.

First, process and governance evaluations help us reduce the variability of program performance. Reduced variability will in turn allow us to evaluate the relative ROIs of alternative interventions with more confidence, and with fewer data points. This implies fewer costly evaluations of like programs.

Second, process and governance evaluations can be less politically difficult: they do not require denial of treatment; they can enhance the government or organization's reputation for retrospection; the treatment variations are usually not apparent (or are not of concern) to the public; and senior management and bureaucrats are usually eager to learn how to be a more effective organization.

Third, process and governance evaluations can provide more timely feedback. Intermediate measures of the quality of project governance may be available long before the final outcomes of the beneficiaries.

Fourth, the generalizabilty of process evaluation results may be greater than the generalizabilty validity of program ROIs—especially in the case with governance and management and decentralization.

This, in brief, is Evaluation 2.0. It goes beyond the mere measure of programs and also seeks to evaluate targeting, scaling-up, program components, processes and governance. It seeks to unpack the causal mechanism experimentally, non-experimentally, or even qualitatively. It seeks to provide real accountability by identifying the people or processes responsible for success and failure, and to provide this feedback in a more timely way. It reinforces our search for relative ROI, and the optimal use of development funds, by reducing variability. It is also likely to be politically and ethically feasible even when evaluation 1.0 (which requires some denial or delay of treatment) is not.

Finally, before finishing, three additional points.

First, several types of programs have received little attention by rigorous evaluators. Economic, educational, and health interventions are the most heavily evaluated types of programs so far. The programs least frequently evaluated are in peace-building, crime-reduction, and governance. By governance I mean many different things, including public management, decentralization and accountability, governance, process improvement.

By our own criteria, however, such programs are keenly in need of evaluation. Their effectiveness is almost completely unknown. The possibility of unintended negative consequences is high. They are often complex interventions, with little accountability.

The most likely reason why these programs have not been heavily evaluated is that economists have led the evaluation charge, and conflict, crime and governance are outside the focus of most economists. If so, it is a problem resolved easily enough through funds and attention.

My second point is that I firmly believe that qualitative and quantitative methods can combine for much more powerful evidence than either one alone. Quantitative researchers often dismiss qualitative evidence as anecdotal at best, and 'just stories' at worst. Yet in combination with experimental evidence of treatment effects and process improvements, the exploration of treatment effects and causal mechanisms using interviews and keen observation is a powerful research tool. It not only generates new evidence, but also otherwise unexpected findings, relationships, and hypotheses to be tested using data.

My third and final point is that the non-experimental approach is wounded but far from dead.

By non-experimental approaches, I mean the use of a control or comparison group that was not randomly assigned. The most common of these approaches is the matching method, whereby a counterfactual is constructed using data on others in the population that look demographically similar to the beneficiaries. A tiny handful of studies compare the performance experimental to non-experimental results. They are almost all uniformly bad news for matching techniques.

These studies, however, are few in number, and almost all US based. Moreover, these few studies typically match many observations based on a very small number of variables on each individual. To my knowledge none have sought to match individuals based on a rich baseline, one that includes extensive health, family, economic, and educational characteristics. Further, computer algorithms for identifying the optimal matching method (such as 'genetic matching' methods) are being developed which hold some hope for more consistent and accurate comparisons. The usefulness of these methods await data for testing, however.

The hasty rejection of non-experimental techniques smacks of a double standard. Those who condemn matching and other non-experimental methods have not applied their own standards of evaluation to evaluation.

Regardless of whether we speak of experimental or non-experimental evaluations, however, all of the limitations of impact evaluation 1.0, and all of the potential benfits of a version 2.0, still apply.

Now, to sum up.

Evaluation 1.0 still has an important role to play. Evaluation 1.0 is most useful when impacts are unconditional, unknown, unintended or unconventional

Unconditional: When program impacts are unlikely to be context or condition specific, and hence have low variability (e.g. some medical interventions)

Unknown: When programs are new, impacts are unknown, and ROIs are potentially low or even negative (e.g. peace-building through media)

Unintended: When there is potential for negative unintended consequences (e.g. gender based violence interventions)

Unconventional: When one needs to challenge the conventional wisdom

Evaluation 2.0 is going to be essential when program impacts are complex or conditional, or when we want to understand causality and accountability on a deeper level.

But virtually every 1.0 evaluation can be upgraded to a 2.0 evaluation. Conventional economic program evaluations can be adapted to also evaluate processes and governance. Moreover, moving from Evaluation 1.0 to 2.0 increases cost and time negligibly. Only the additional layer of complexity must be managed.

This complexity is optional, however. When randomization of treatment is practically or politically or ethically infeasible, process evaluations can be performed without denying treatment to any group. To stretch my analogy nearly to its breaking point, you might say that version 2.0 does not necessarily require version 1.0 to be installed.

In the end, the single biggest shift in moving from version 1 to 2 is the shift from thinking about relative ROI to performance management and process learning. It requires recognizing that to be successful impact evaluation must first and foremost meet the needs of implementers. Luckily, the goals of performance management and process learning are closely aligned with the interests of the academics who often design and implement evaluations. Academics care, or ought to care, about causal mechanisms. They also tend to be interested in the study of incentives, institutional organization, and other human behavior. The common ground between academics and implementers ought to expand under version 2.0, not contract.