Evaluation

Cyrus Samii
Columbia University
December 20, 2010

Contribution to
The Management Handbook: A Practical Guide for Managers in UN Field Missions,
International Peace Institute

1. Overview

The word "evaluation" is used in different ways in the management context. To minimize confusion, we distinguish between the following:

(i) Beneficiary assessment,
(ii) Performance evaluation,
(iii) Impact evaluation.

Each of these will be described below. We spend a bit more time on impact evaluation, because this is the most complex. It is the subject of high interest but also much confusion. Two short case studies of impact evaluations will be presented. We look at some lessons learned about difficulties that managers are likely to face in conducting or commissioning an evaluation. The chapter includes a checklist for self-evaluation and diagnosis, a list of relevant UN documents, and a bibliography.

2. Importance

Evaluation of a program serves two primary functions: *accountability* and *learning*. By a "program," we mean a medium- to long-term task that has a clearly demarcated assessment period. These include vocational training and job assistance programs, water and sanitation programs, road rehabilitation, peacekeeping deployment cycles, IDP return programs, quick-impact projects, ex-combatant reintegration programs, community reconciliation programs, and so on. When an evaluation is commissioned, the assessment period needs to be clear. We distinguish between evaluations and *investigations*. Investigations are commissioned to learn about a specific event, usually some dramatic failure. Evaluations on the other hand look into general performance of a program over an assessment period.

Accountability is directed toward the program's *principals* and *beneficiary population*. By "principals," we mean the actors who will ultimately have the authority to determine whether a program will be continued or used elsewhere. The relevant principals may be UN agency heads, donor government offices, or at the broadest level, tax payers in donor governments. The "beneficiary population" includes those who are supposed to be helped, and not harmed, by the program either directly or indirectly. This may include the individuals that receive program benefits, as well as citizens and officials in the country or region whose well-being is potentially affected by the program. In defining a program's aims, it is necessary that the interests of both principals and the beneficiary population are considered in depth. Evaluators can help program managers to consult principals and members of the beneficiary population before, during, and after implementation. The resulting information flow helps to ensure that programs target valid needs in an effective, efficient, and sustainable manner.

In fulfilling the accountability function, evaluation tasks help to answer the following questions:

1. Where deliverables in fact delivered? Was the budget spent as intended?
2. Where deliverables delivered in an appropriate manner? Were beneficiaries selected on the basis of appropriate criteria?
3. Did beneficiaries participate and engage in the program in the expected manner? If not, was it because of blockages to access or lack of interest?
4. Whose actions were responsible for the successes or failures in the program?
5. Were the intended benefits realized? Were the benefits sufficient to justify the cost? Was anyone harmed?
6. Did the program appropriately target needs, or was it misdirected? Were the assumptions on which the program was based valid?

Learning serves to improve the *efficiency* and *effectiveness* of programs. Effectiveness is judged by the extent to which benefits were realized over and above any negative consequences of the program. Efficiency is judged by the net benefits relative to the cost of program inputs---or, in the words of the OECD DAC evaluation guidelines, "the extent to which aid uses the least costly resources possible in order to achieve the desired results." Evaluation can be used to derive "lessons learned." These lessons can provide guidance on organizing a program by answering questions such as the following:

1. What kind of management structure provides the right incentives? What kind of management structure helps to ensure good problem solving and decision-making, rather than hindering it?
2. How can we design the program to increase positive impacts on beneficiaries? For example, should the program's activities be sequenced in a certain way? Are certain benefits packages more effective than others? Are there synergies between different types of programs that we should exploit?
3. How can program activities be designed so that beneficial changes are sustainable, and not tied to incentives provided by program activities?

These lessons can be used to modify an ongoing program or design a future program. We want to avoid using the term "best practices," because we are always learning how to do things better. The idea of a "best" practice is unrealistic. The learning function of evaluation is intended for identifying areas of *improvement* and getting a sense of *what works*.

3. Key terms and concepts

A useful definition for evaluation is provided by the OECD DAC *Quality Standards for Development Evaluation*:

> Evaluation is the systematic and objective assessment of an on-going or completed project, programme or policy, its design, implementation and results. The aim is to determine the relevance and fulfilment of objectives, development efficiency, effectiveness, impact and sustainability. An evaluation should provide information that is credible and useful, enabling the incorporation of lessons learned into the decision-making process of both recipients and donors. Evaluation also refers to the process of determining the worth or significance of an activity, policy or program.

Often times the term "monitoring and evaluation" (M&E) is used to describe these and related tasks. It is important for program managers to separate out *monitoring* from evaluation. Monitoring

addresses the question, *are intended program deliverables in fact being delivered?* Monitoring information on the delivery of benefits is typically entered, in real time, into a *management information system* (MIS). Typically, a program will employ a database specialist to collect M&E information on some regular basis (e.g. quarterly or yearly) and enter it into the MIS. The MIS may keep tallies on numbers of beneficiaries served, numbers and types of benefits delivered, and numbers of activities completed. Monitoring is primarily an accounting mechanism, providing information on "actual" delivery of program benefits to be compared to "intended" delivery. It can also be used to determine whether the program met unexpected difficulties. Regular monitoring reports are necessary for assessing whether the program is meeting its *output* goals. But they tell you little about whether the program is achieving desired *outcomes*, which should be measured in terms of changes brought about for beneficiaries. The latter is what evaluation does. Clearly monitoring can support evaluation, in that it provides information on the mechanisms by which outcomes may be affected. But it is not a substitute.

It is useful to distinguish three evaluation tasks: beneficiary assessment, performance evaluation, and impact evaluation. When a program manager is either commissioning an evaluation or responding to a request to do one, it is crucial to clarify the task.

3.1 Beneficiary assessment

A beneficiary assessment measures conditions among those receiving benefits, directly or indirectly, from the program. It addresses the following questions,

- Are the program outputs reaching the intended beneficiaries?
- Are beneficiaries satisfied with the benefits that they are receiving?
- Is the program generating any ill will among other community members?
- How are the conditions of beneficiaries changing over time?
- Is the program providing benefits that are relevant to the target community's needs?
- Do changes brought about by the program last beyond the period of program activities? That is, has the program brought about *sustainable* change, or only temporary change?

As an example, consider a beneficiary assessment for an ex-combatant reintegration program. Such an assessment would keep an updated tally of the following:

- Information about the beneficiaries' satisfaction with the program benefits.
- Information on whether beneficiaries are finding employment.
- Information on whether members of the community surrounding the beneficiaries are satisfied with the program.

Beneficiary assessments may use either *quantitative* (that is, statistical) information or *qualitative* (that is, narrative) information. This quantitative information may be stored in the MIS along with the output information. The information may be gathered using formal questionnaires administered to beneficiaries, focus group discussions, or interviews with key informants. These tasks are typically carried by an evaluation consultant and an evaluation team. Beneficiary assessment can help program managers to understand whether the assumptions behind the design of the program were good ones, whether the program goals are valid, and whether immediate needs are being addressed appropriately. A beneficiary assessment cannot measure the "impact" of a program directly, because there is no comparison group. This is described in more detail below, in the section on impact evaluation.

3.2 Performance evaluation

Performance evaluation studies the performance of individuals and organizations implementing a program. It addresses the following questions:

- In implementing the program, did the organization perform as efficiently as possible, and if not, why?
- Did decision-making protocols help to reduce mistakes? Did they help to ensure that all stakeholders' interests were taken into account? Or, did they unduly obstruct timely and effective problem-solving?

Performance evaluation is usually undertaken after the program is finished, but it may be done at a few points in time during the programming period. Performance evaluation is based on interviews with program staff, and they may also use a combination of quantitative and qualitative information. The interviews may use a standard questionnaire or they may be open-ended. Performance evaluation reports contain vignettes that describe specific successes or failures that occurred during the program. These vignettes are used to suggest lessons for management practice. Much of what UN-DPKO's Best Practices Unit does is performance evaluation, for example.

Performance evaluation is especially useful for programs that involve cooperation between agencies. In these cases, performance evaluation can be used to determine whether efficiency was helped or hindered by the program's protocols for authorizing program activities and approving expenditures. In other words, performance evaluation is used to determine what kinds of coordination mechanisms are effective.

3.3 Impact evaluation

Impact evaluation is the most ambitious and technically challenging evaluation task. It tries to answer the following questions:

- For the people targeted by the program, would their well-being have been worse, better, or pretty much the same had the program never taken place?
- Were there any indirect or unintended effects, whether good or bad, due to the program?
- Were the benefits of the program sufficient to justify the costs?
- What kinds of people benefited most from the program?
- What strategies are effective for making the program more beneficial?

As an example, an impact evaluation for an ex-combatant reintegration program could study how well different kinds of reintegration benefit packages (e.g. different types of job training programs) helped beneficiaries to obtain productive livelihoods. An impact evaluation as part of a peacebuilding mission might study the effectiveness of community reconciliation programs in overcoming inter-ethnic mistrust. Impact evaluations may use quantitative information to measure impacts, and qualitative information to describe the processes through which impacts occurred.

Impact evaluation generally requires a *comparison group*. That is what sets it apart from beneficiary assessment. Think of an impact evaluation approximating for your program what a *clinical trial* does for an experimental medication. Consider the example of a new headache medication. A person comes into a doctor's office with a headache. The doctor offers the person a dose of a new medication. The person takes the medication, and then after a few hours, the headache is gone. What can we say about the drug's effectiveness from this episode? In fact, there is very little we can say. Think about some of the problems:

- Do we really know whether the headache would have gone away had the person *not* taken the drug?
- Do we really know whether the headache would have gone away *sooner* had the person not taken the drug?

- Do we have any reason to believe that this single experience speaks to what we should expect more generally?

The problems listed here are what one faces when one wants to *attribute* an outcome (headache status) to some treatment (drug). Clinical trials are designed to overcome these problems. The key feature is a *randomized control trial* on *a large group* of people. In randomized control trials, individuals are selected at random to be either in the "treatment" group, in which case they will receive the drug, or in the "control" group, in which case they will receive either a placebo or nothing. Randomization over a large group ensures that the "treatment" group does not differ from the "control" groups in systematic ways. It ensures that measures of the effect of the treatment are not sensitive to abnormal responses. Thus we obtain good measures of "average" responses to the treatment. There are other subtleties to randomized control trials, such as "blinding," that are beyond the scope of our discussion but also contribute to the validity of a trial.

The same principles for *causal attribution* apply to impact evaluation for development, post-conflict, or humanitarian programs. The "treatment" is your program. Like a clinical trial, an impact evaluation tries to measure the *impact*, which is another name for a "causal effect." Here is a definition of impact, referring to development interventions, from the OECD DAC guidelines:

> The positive and negative changes produced by a development intervention, directly or indirectly, intended or unintended. This involves the main impacts and effects resulting from the activity on the local social, economic, environmental and other development indicators. The examination should be concerned with both intended and unintended results and must also include the positive and negative impact of external factors, such as changes in terms of trade and financial conditions.

And here is a definition for impact evaluation from the International Initiative for Impact Evaluation,

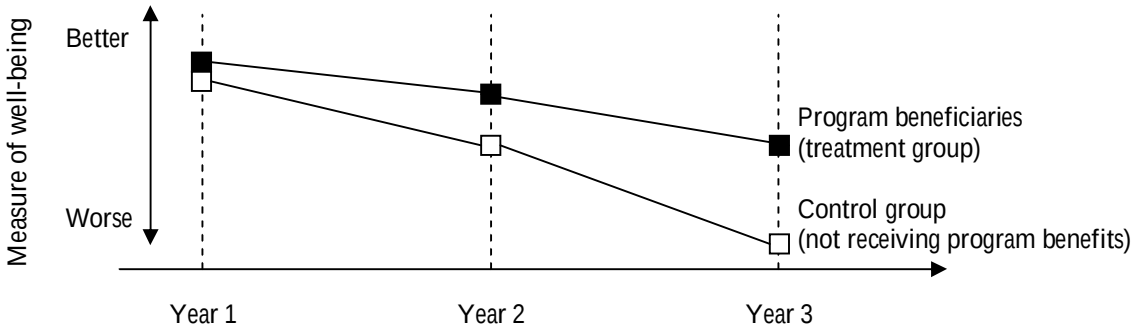
> *High quality impact evaluations* measure the net change in outcomes that can be attributed to a specific program. Impact studies help inform policy as to what works, what does not, and why.

Thus, the *impact* of a program is the *difference* between beneficiaries' well-being after the program *and* some estimate of beneficiaries' well-being *had there been no program*. The key is to construct the counterfactual of what would have happened had there been no program.

Impact is a subtle concept and it is often misunderstood. People sometimes define impact as the "difference between beneficiaries' well-being before and after a program." This is generally incorrect. We can call it the "before-after fallacy." It is a fallacy because *many things* affect how beneficiaries' well-being changes over time. There is no reason to *attribute* such change (whether positive or negative) to the program. We need to have a comparison group. Figure 1 illustrates this. It shows a case where the "before-after fallacy" would result in an unfair judgment about the program. The well-being of beneficiaries (the "treatment group") goes down over time. A naïve interpretation would be that the program caused harm. We avoid this fallacious conclusion by looking at a "control" group. What we see is that both groups experienced a decline in their well-being. However, the decline is less severe for the program beneficiaries than for the control group. Therefore, the program had a *positive* impact.

The way that we estimate what would have happened *had there been no program* is with a control group.

Figure 1: Illustrating the Before-After Fallacy



Is the randomized control trial model valid for development, post-conflict, and humanitarian programs? Sometimes it is, and sometimes it isn't. For example, post-conflict and humanitarian contexts may not permit randomization, either because of the need for sensitive application of interventions or because of the need to act quickly. That does not change the fact that randomization, in theory, is the *most reliable method* for impact evaluation. The further we stray from this standard, the less we can know about whether a program is effective or not. Sometimes it makes sense to accept such limitations, in which case alternative methods may be used.

The key to impact evaluation is constructing a comparison group. There are two types of strategies. The first type involves *designing the program itself* in such a way as to allow for a rigorous impact evaluation. The second type *exploits almost-random variation* and uses *controlled comparisons*. We call the latter *quasi-random strategies*.

Designing programs to allow for rigorous impact evaluation

1. Full randomization. This is the strongest method. A program can first narrow down potential beneficiaries to a pool of worthy individuals, communities, or organizations. The number of potential beneficiaries may be much greater than the number of people that the program can serve. A lottery system is used to determine who receives benefits and who does not. To improve the strength of the study, would-be beneficiaries can even be paired on the basis of important characteristics, and then a coin flip can be used to determine which member of the pair is receives the program benefits. The impact evaluation tracks outcomes among those selected *as well as* those not selected to receive benefits. It uses differences between these two groups to determine impact. Positive evidence can be used to apply for funding to extend program coverage. A famous example is the Mexican government's *Progresa-Oportunidades* program, which randomly selected communities to be eligible for conditional cash transfer programs.

Full randomization is routine in evaluating education, development, and public health interventions. It can be used to assess the impact of an *entire program* or it can be used to assess the relative effectiveness of different strategies *within* a program. For example, randomized impact evaluation has been used to assess the relative merits of two beneficiary selection methods for targeted income assistance programs. The first method was "means testing," which is very precise but very expensive, and the second was "community selection", where community leaders are assembled to name beneficiaries. The latter is potentially subject to "capture" by local elites. Randomized impact evaluations found that in most cases, community selection processes were fair.

Full randomization requires considerable advanced planning. It requires that those involved in the program agree on the need to learn about *whether* a program is effective. Case study 1 provides an example of a randomized impact evaluation, studying the impact of a community-directed reconstruction program in eastern Liberia.

2. Randomized roll-out. Sometimes a program cannot serve all beneficiaries at once, and so beneficiaries are served in phases. A randomized roll-out randomly assigns beneficiaries to receive benefits during either an earlier or later phase. During the early phase, we have a randomly determined group receiving benefits and a randomly determined group that is not. The latter form a control group until they start receiving their benefits. This is similar to full randomization. But it has drawbacks. It can assess only short-run impacts. Members of the early phase "control" group may be aware that they will eventually receive benefits. This may taint the analysis. This strategy is commonly used to determine whether a program should be modified before it is set to operate at full scale.

3. Beneficiary selection with a numerical index. Full randomization or randomized roll-out may not be politically feasible. The program may be required to target "the most needy." A rigorous impact evaluation may still be conducted, so long as a transparent rule is used for determining who is "most needy." The most transparent mechanism is to use a numerical index of neediness. Then, one uses a strict cut-off on this index to determine who is eligible for assistance and who is not. For example, a recent World Bank community-directed post-conflict reconstruction program in Aceh, Indonesia, used such a strategy. The program constructed a vulnerability index from data on communities' conflict affectedness and their poverty levels. They chose the communities with the highest vulnerability scores, moving down the list to include as many communities as their budget allowed. This permits a rigorous impact evaluation. The impact evaluation exploits the fact that those individuals or communities just above and just below the cut-off will be very similar. In that way, comparisons between those just above and just below the cut-off resemble a randomized experiment. The drawback is that the impacts that one measures may be specific to people or communities near the cut-off.

Quasi-random strategies

1. *Matched comparisons*. The idea behind this strategy is simple: one creates a "pseudo"-control group by matching each beneficiary with someone/something that resembles the beneficiary in important ways but is not a beneficiary. In the case of program that operates at the household or community level, one would match beneficiaries with non-beneficiaries. In the case of a region- or national-level program, one could create a so-called "synthetic match" by using information from comparable regions or countries. Then, one tracks the outcomes of the beneficiaries and the matched counterparts. Impact is estimated as the difference in outcomes between these matched units. At first glance, it may seem that this is the ideal way to construct a control group. We are ensuring that we are making comparisons between units that are comparable in important ways. Isn't this better than randomization? From a purely statistical perspective, the answer is no. Matching can only be based on information that we have measured. But, there may be many *unmeasured* things that taint the analysis. Only randomization ensures that treated and control units are similar on average in terms of *both measured and unmeasured* characteristics. Nonetheless, practicalities or ethical

considerations may be such that matched comparisons are the best available option. Case study 2 shows an example of a matched comparison impact evaluation, studying community-level impacts of security provision by UNMIL peacekeepers in Liberia.

2. *Natural experiments and fortuitous accidents*. Sometimes, we can use variation due to nature or unanticipated occurrences to obtain a situation that resembles an experiment. For example, a recent impact evaluation of ex-combatant reintegration in Burundi took advantage of the fact that a bureaucratic dispute between one of the implementing partners and the government resulted in nearly a third of the program's beneficiaries from having their benefits withheld for about a year. The impact evaluation conducted a survey during this period of disruption, using those whose benefits were being withheld as a pseudo-control group to compare to those whose benefits suffered no such disruption. Of course, this strategy is limited in its applicability, because fortuitous accidents cannot, by nature, be planned.

Impact evaluation without a comparison group

In some cases, it would seem impossible to construct a comparison group. A prominent example would be in cases where one has to measure the "impact" of a program that is administered at the national level. There are strategies for these situations, but generally they cannot reach the level of rigor of the strategies outlined above. Nonetheless, they may be the best available option. One strategy is to try to identify changes that *could only plausibly occur* as a result of the program, and then to track outcomes to see if they are manifest. Such a strategy can benefit a great deal from qualitative information from beneficiaries, who can provide details about how, exactly, the program has affected their well-being. The weakness relative to using a comparison group is that this approach relies on more assumptions; there is no way to say for sure whether an outcome is exclusively attributable to the program. Another approach is to use pre-program information and the opinions of experts and informants to project what would have happened with no program, and then to track outcomes against this benchmark. These strategies are what national governments sometimes do in assessing the impact, e.g., of tax laws, in which case economic models are used to predict what would have happened had there been no policy change. The weakness is that there is no way to validate these projections.

3.4 Importance of Prospective Evaluation

Public institutions are under increasing pressure to demonstrate *effectiveness* and do *evidence-based policy-making*. Therefore, rigorous beneficiary assessment, performance evaluation, and impact evaluation have become subjects of major interest in governments and international organizations. The best evaluations are *prospective*, meaning that they are initiated along with the start of a program, rather than being *retrospective*, meaning that they are initiated only after the program is finished. Prospective evaluations ensure that those doing the evaluation know what is happening in the program. A problem with retrospective evaluations is that the evaluators have only a foggy picture of what happened. This is a common complaint about retrospective evaluations—that the evaluators "had no idea what really happened."

3.5 Theory of change

Evaluations are, ultimately, scientific endeavors. It is therefore useful to propose a *theory of change* to provide structure for the evaluation. A theory of change puts into writing expectations about how

program outputs will translate into meaningful outcomes among beneficiaries. A beneficiary assessment for a community directed reconstruction program might propose that most members of beneficiary communities should actively participate in meetings to determine spending priorities. The theory of change states that such participation by community members will result in more equitable allocation of resources. Here, the outputs are the numbers of meetings held. The outcome is the equitableness of resource allocation. The theory of change links the two via participation. The evaluation should include measures of actual participation in meetings, along with information on meetings held and equitability. An impact evaluation for a community reconciliation intervention might propose that inter-ethnic community dialogue meetings allow community members to learn new things about their common interests, and thus help to overcome mistrust across ethnic lines. The outputs are, again, the number of meetings held. The outcome of interest is level of mistrust. The theory of change links the two via learning. The evaluation should measure such learning, along with the number of meetings held and the levels of trust. Writing down a theory of change helps to focus the evaluation and assess whether the program is working as expected. Clearly, an evaluation needs to spell out how the *outcomes of interest* are to be measured. Using our community dialogue example, it is not obvious how to measure "inter-ethnic trust." A good evaluation requires the input of social scientists to devise measures and assist with the overall design. The *measurement strategy* might use a combination of quantitative measures, such as sample surveys, and qualitative measures, such as stories and vignettes from focus groups and interviews.

3.4 How it all fits together

Hopefully you now understand why it makes sense to separate "monitoring" from "evaluation", and to distinguish between beneficiary assessment, performance evaluation, and impact evaluation. Monitoring describes what the program has delivered. Beneficiary assessment tracks outcomes among those targeted by the program. Performance assessment focuses on decision-making and management structures. Impact evaluation generally measures differences between beneficiaries and a comparison group. Prior to commissioning an evaluation of any type, a program manager may want to assess the *evaluability* of a program. Given constraints on time, budget, and beneficiary selection, what kind of evaluation is feasible? As a manager, will that satisfy your accountability and learning needs?

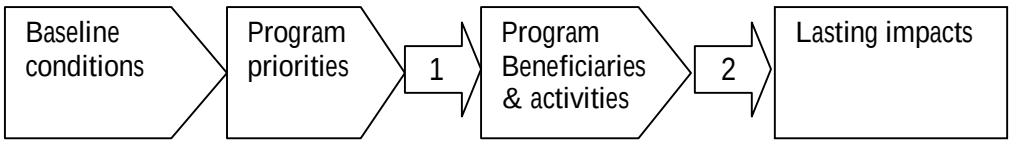
A tool used by evaluation professionals to organize these tasks is a *logic model*. A logic model illustrates how different evaluation tasks fit into a program. Figure 2 shows components of a logic model. An assessment of baseline conditions feeds into program priorities. These priorities are translated into activities. The activities are proposed to lead to impact. Performance evaluation is primarily concerned with the arrow 1. Monitoring and beneficiary assessment is primarily concerned with keeping track of "Program activities & beneficiaries" and may collect some information associated with arrow 2. Impact evaluation is all about arrow 2, studying outcomes among beneficiary and comparison group communities.

Other terms are sometimes used by evaluation professionals to describe clusters of evaluation activities. "Process evaluation" and "program evaluation" are sometimes used to refer to beneficiary assessment and sometimes also performance evaluation. This can be confusing, but we feel that the distinction between monitoring, beneficiary assessment, performance evaluation, and impact evaluation provides the clearest delineation.

Finally, there exists some debate on what impact evaluation strategies are appropriate under different conditions. The centrality of randomized control trials has been questioned, e.g. by economists James J. Heckman and Angus Deaton. They worry about the method dictating the questions, wherein evaluators focus on questions amenable to randomization, rather than adapting

the method to the question. For these commentators, there is great need for quasi-random strategies such as those described above and also for a tighter link between evaluation and theories of change. Another area of debate is over the use of quantitative versus qualitative indicators. Here, however, it seems that most evaluators would agree that the two sources of information complement each other.

Figure 2: Template for a Logic Model



4. Lessons learned

What are the main reasons that evaluations do not deliver what managers hope they do?

For an evaluation to deliver what you want, you need to be clear about your goals. Are you trying to simply monitor output and ensure that beneficiaries are generally happy? Are you trying to draw lessons for future management practice? Are you trying to assess cost-effectiveness? Each of these things implies a different set of tasks. The first is monitoring and beneficiary assessment, the second performance evaluation, and the third requires impact evaluation. Sometimes, an evaluation will be commissioned to external consultants, and the resulting report will be unsatisfactory to the program managers. The reason is usually because the goals were not spelled out. Use the concepts that we have developed above to refine your goals and spell them out clearly to your consultants.

What if I have a mandate to demonstrate "effectiveness" and therefore want an impact evaluation?

If you want an impact evaluation, you need to assess whether your program is flexible enough to allow for randomization, randomized roll-out, or the use of a numerical index to determine beneficiary eligibility. If not, you need to determine whether there is enough quality data available to construct a plausible matched comparison, or whether there is some quasi-random source of variation that you can use. Program managers sometimes commission impact evaluations in situations where causal attribution is nearly impossible. The program may be over already. Beneficiaries may have been selected according to vague criteria, with poor records kept on how this happened. There may be no clear theory of change or no clear sense of what are the outcomes of interest. There may be too little information to construct a plausible control group. The number of beneficiaries may be small. Under these conditions, an impact evaluation may be hopeless. If this is the case and you still need to do an impact evaluation, you may have to change some features of the program itself to make impact evaluation possible. This means that you accept the learning and accountability objectives as being as high a priority as the proximate goals of the program itself. It is usually a good idea to seek the assistance of social scientists, public health scholars, or other relevant experts to help design an impact evaluation. The International Initiative for Impact Evaluation (www.3ieimpact.org) tries to connect government and inter-governmental organizations with scholars to facilitate impact evaluations.

How can evaluations be most effective for accountability and learning?

The answer to this question is surprisingly clear: use a prospective evaluation. That is, bring in the evaluation consultants and establish a data collection system before the program begins its field operations. This holds for all types of evaluation. Sometimes the program itself may have to be modified. Evaluation objectives should be clarified from the very beginning, before you finalize program activities or procurement contracts. The evaluation design and objectives need to be institutionalized in all program documentation to ensure that implementing partners do not take decisions that may undermine evaluation goals. For example, suppose a prospective impact evaluation is to be done using a design based on a numerical index for beneficiary selection. Implementing partners *must not* be allowed to deviate from using this index in their beneficiary selection. In order to ensure that, the beneficiary selection processes must spelled out in the implementing partners' contracts.

What if my program is almost over, and I now realize that I need an evaluation?

Your only choice is a retrospective evaluation. Retrospective monitoring and beneficiary assessment would involve auditing program documentation to account for outputs and interviewing beneficiaries to ask about how things have changed. Retrospective performance evaluation would involve interviewing current and past program staff about their experiences and using whatever records are available to reconstruct key events during the program. Retrospective impact evaluation would involve trying to identify some plausible control group and examining how they differ from beneficiaries. For all types of evaluations, you will not be able to find all the people and all the data that you need to reconstruct fully what happened. People will have forgotten what really happened and may misreport things. You will have to accept that the picture that you obtain will be incomplete and possibly biased. Nonetheless, this may be the best that you can do.

5. Checklist

- What do you want to learn from your evaluation? What will principals and members of the beneficiary population want to know?
- Do your evaluation goals imply that you only need monitoring and, perhaps, beneficiary assessment, or do you want to go further and have a performance evaluation or impact evaluation?
- Are your evaluation goals and your methods for reaching them written into all of your program documentation?
- If you want an impact evaluation, is your program flexible enough to use a design-based strategy?

6. Relevant UN documents

United Nations Development Programme, *Handbook on Planning, Monitoring, and Evaluating for Development Results.* Posted to http://www.undp.org/evaluation/handbook/

United Nations Evaluation Group, *Guidance Documents.* Posted to http://uneval.org/papersandpubs/index.jsp

United Nations Office for the Coordination of Humanitarian Affairs, *Evaluation Studies Information*, posted to http://ochaonline.un.org/ToolsServices/EvaluationandStudies/tabid/1277/language/en-US/Default.aspx

United Nations Office of Inspection and Oversight Services, *Guidance to Programmes for Developing an Evaluation Policy*. Posted to http://www.un.org/Depts/oios/pages/ied_guidance_for_dev_ep.pdf

7. Bibliography

Abdul Latif Jameel Poverty Action Lab, *What is Evaluation?* Posted to http://www.povertyactionlab.org/methodology/what-evaluation

Francois Bourguignon and Luiz A. Pereira da Silva (editors), *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools*, World Bank and Oxford University Press, 2004.

OECD DAC, *Evaluating Development Cooperation: A Summary of Key Norms and Standards, Second Edition*, June 2010.

OECD DAC, *Guidance on Evaluating Conflict Prevention and Peacebuilding Activities (Working Draft for Application Period)*, 2008.

Peter H. Rossi, Mark W. Lispey, and Howard E. Freeman. *Evaluation: A Systematic Approach.* Sage Publications, 2003.

William R. Shadish, Thomas D. Cook, and Donald T. Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Wadsworth Publishing, 2001.

World Bank, *Monitoring and Evaluation: Some Tools, Methods, and Approaches*. Posted to http://www.worldbank.org/oed/ecd/tools/

Case study 1

The International Rescue Committee, a major international NGO, and the United Kingdom Department for International Development (DFID) commissioned a fully randomized evaluation of the impact of a two-year community-directed reconstruction (CDR) program in Lofa County in Eastern Liberia (2006-8). The program constructed a pool of 83 communities there were "equally deserving" to receive assistance through the CDR program. A lottery was used to randomly select 42 communities to receive the program. The 41 non-selected communities formed the control group. The recipient and control group communities were assessed at the beginning and the end of the two years of the program. The theory of change was that CDR programs would help to reunite communities divided by war, and that this would allow communities to promote their own welfare more effectively. The impact evaluation examined indicators of "community reunification" and "social cohesion" as well as indicators of material welfare and governance. The findings suggests that "the program reduced social tension, increased the inclusion of marginalized groups, and enhanced individuals' trust in community leadership." However, "evidence is much weaker that the program positively reinforced support for democracy, had an impact material well-being or resulted in an increased ability of the community to act collectively and provides no evidence that the attitudes of traditional leaders towards decision making were affected in any way."

The evaluation was exceptional in its rigor, owing mostly to the use of full randomization as the method for creating a control group. In addition, the evaluation used a combination of sample surveys, behavioral games, and interviews with community leaders to measure outcomes of interest. This kind of "triangulation" of measures helps to boost one's confidence in accuracy of the findings.

Reference

James Fearon, Macartan Humphreys, and Jeremy Weinstein, *Community-Driven Reconstruction in Lofa County: Impact Assessment*, submitted to the International Rescue Committee, December 2008.

Case study 2

In 2008, the United Nations Office of Internal Oversight Services Inspections and Evaluations Division commissioned an evaluation of the impact as of Autumn 2008 of the United Nations Mission in Liberia (UNMIL) deployments on community-level security, economic recovery, and democracy promotion. The evaluation team used a "matched comparisons" strategy to construct a control group. Twelve communities that hosted deployments were matched to twelve communities that were far removed from deployment bases but resembled the deployment communities on the following criteria: ethno-linguistic region, agricultural region, proximity to roads, number of households, number of schools per 100 households, number of health posts per 100 households, and the number of conflict events that occurred in or near the community during the war. The theory of change was that deployments created a local security bubble, and that within this bubble, economic reconstruction and democracy promotion could flourish. The evaluation team used sample surveys of people living in these twenty four communities to measure local security conditions, economic recovery among households, and political perceptions. The key findings from the IE were that the security contribution of UNMIL was not in providing local law and order, but rather a more general suppression of the likelihood of conflict recurrence nationwide. Deployments seemed to stimulate local labor markets, but there seemed to be no impact of the deployments on democracy promotion.

A few remarks are in order on the quality of this impact evaluation. Of course, since deployments cannot be randomized, we should be concerned about the possibility of hidden factors that taint the comparison between the communities that were close to deployments and those that were far away. Also, one cannot directly measure security conditions; one has to rely on the answers that survey subjects are willing to offer, and these may suffer from "courtesy bias"--meaning that respondents say what they think the survey interviewers want to hear. The same goes for measuring democracy promotion.

Reference

Eric Mvukiyehe and Cyrus Samii. *Quantitative Impact Evaluation of the United Nations Mission in Liberia: Final Report*. Submitted to UNOIOS-IED, February 2010.