INTERNATIONAL
RESCUE
COMMITTEE

IRC

EVALUATION GUIDELINES

Research, Evaluation and Learning

June 2007

EVALUATION
EVALUATION
EVALUATION 02
PROGRAM EVALUATION
MONITORING DATA
EVALUATION REPORT-01
PROGRAM MONITORING DATA
EVALUATION 2003
EVALUATION 04
EVALUATION 05

"I can honestly say that not a day goes by when we don't use those evaluations in one way or another."

Written by *M. M. Rogers* and illustrated by *Lawson Sworh*

TABLE OF CONTENTS

I. INTRODUCTION: HOW DO WE KNOW IF WE ARE MAKING A DIFFERENCE?

Have you ever heard or said:

> *"We've been training people for 2 years and we don't know if it does anything?"*
>
> *"IRC's been in this country for 8 years; how do we know if we've made a difference?*
>
> *"Did this project accomplish its objectives?"*
>
> *"Are we really strengthening local institutions?"*
>
> *"That strategy's great. But how do we know if it works?"*

Whether you've considered specific questions about the result of trainings or the distribution of goods, or broad questions about IRC's effectiveness or the impact of an IRC project on a community, you are essentially asking evaluation questions. One of the reasons we cannot always answer them is because IRC does not use evaluation as the resource it can be.

Evaluation is one of the main tools available to NGOs to learn about their work, advance organizational learning and maintain accountability with key constituencies. Regardless of its importance, evaluation is not a panacea on its own. It is not unusual to see an evaluation Terms of Reference that expects a consultant to produce significant results regardless of how the project was designed and monitored over its life time. The fact is that the connections between design, monitoring and evaluation are not arbitrary: project design provides the roadmap for both monitoring and evaluation; data from monitoring over time should tell you how a project is going; and evaluation cannot make up for what we have failed to do from the onset – for example, ensure that we monitor our work using clear indicators and the right data collection methods.

Similarly, even if we evaluated all of IRC's work, the mere fact of doing so does not guarantee that we will be able to answer the questions above. Many organizations evaluate more than IRC does and still wonder about their impact and effectiveness. How an evaluation is actually done and whether or not its findings are used are equally, perhaps more important to organizational learning.

Taking this reality as a starting point, these guidelines are written to help improve the way IRC uses and learns from evaluation. They include the terms and concepts essential for improving the quality of IRC evaluations across sectors and country programs; they are not, however, a guide on how to do evaluation. In fact, one of the basic premises of IRC's DM&E Strategy is that activities such as evaluation are an expertise. We should treat them as such and decrease the expectation that all staff must understand and do everything.

The primary audiences for these guidelines are technical and regional units and in-country senior technical and management staff. We hope that they will demystify evaluation and help us all to ask the right questions – of ourselves, but also of the consultants and donors working with us to implement and evaluate programs that truly make a difference in people's lives.

WHAT IS EVALUATION AND WHY SHOULD WE DO IT?

The word "evaluate" means to determine the merit, worth or value of something. We use it to refer to a more narrow scope of activity known as *program evaluation,* defined here as the systematic collection and analysis of information about the activities, characteristics and outcomes of programs. Because evaluation uses social research methods, the terms *evaluation* and ***program evaluation*** are often used interchangeably with the term *evaluation research.* Regardless of the word used, the bottom line is the collection and analysis of meaningful information to inform and guide learning, decision making and practice.

Key Term

Program Evaluation

The systematic collection and analysis of information about the activities, characteristics and outcomes of programs.

The short answer to the question "Why should we do it?" is that it is an essential tool to assess whether or not we are doing what we intend to do. In this sense, it is at the heart of our need to be accountable to beneficiaries, donors and ourselves. In recent years, the search for "measurable" results (in the words of former USAID Administrator Andrew Natsios) or "demonstrable impact" seems to have taken on new urgency for private and government donors, UN agencies and NGOs alike. There are several reasons why IRC should be part of this movement to improve the evaluation of its work.

1. **Little things add up.** If we cannot and do not evaluate at the project level, we can not make any judgment about whether or not we influence "big picture" conditions (consider the aims of the Program Framework, for example) that affect the populations with which we work.
2. **Evaluation can help IRC to improve program quality.** Whether program quality is defined by our ability to meet the Program Framework aims, adhere to its principles or implement "effective" programs (i.e., programs that *work* or accomplish their objectives), we cannot improve it if we cannot see it. Evaluation provides us with a way to document and therefore "see" our work in action.
3. **Accountability is central to IRC's work and mission.** Evaluation is often characterized as an accountability tool because it can engage and share information with beneficiaries, donors and/ or other key stakeholders.
4. **Evaluation can help prioritize resources.** When done rigorously, evaluation can help IRC to identify the approaches, sectors and/or projects that are more and less effective, ultimately influencing the allocation of resources. During implementation, it can also provide essential information to staff about what adjustments can improve program effectiveness.

"Results count. If you cannot measure results, other partners will be found."

Andrew Natsios (2003)

? Want to know more?

It is interesting to note that the history of program evaluation began in an academic setting. After World War II, federal and privately funded social programs launched in the United States, and international programs in Asia, Latin America and Asia were accompanied by demands for "knowledge of results." By the 1970s, evaluation research was a distinct specialty field in the social sciences, carried out primarily by scholars skilled in social research. The continued connection between evaluation and research methods becomes very clear if you think about what motivates us to evaluate our work: to create a description of program performance that is as credible and objective as possible. Social research methods and standards of methodological quality have been developed explicitly for the purpose of constructing sound or valid factual descriptions of social phenomena. This is important to know for two reasons. First, understanding that program evaluation relies on research methods allows us to see it as a technical pursuit that requires training and experience to do well. Second, this same realization should relieve some of the expectations we have of ourselves and IRC staff in general. There is no reason to expect a staff member to be able to write a useful Terms of Reference let alone spearhead an evaluation without significant support and guidance. Likewise, there is every reason that IRC staff should expect from evaluators a thorough explanation of the methods they use, the knowledge they will produce for IRC and the degree to which IRC can learn something valuable from their work.

If there are such good reasons to evaluate our work, why don't we do more of it, more often? And when we do it, why isn't it the powerful learning and accountability tool it is supposed to be?

Evaluation at IRC: What do we do and what needs to change?

Whether you work in a country program, at headquarters in New York or London, ask a colleague who works in the same programming area if she can remember the findings of a recent evaluation. It is very possible that IRC staff will not be able to remember the last evaluation done. Some country programs have gone two years without one; some long lasting projects have gone five years without one. When evaluations are done, the chances are that the headaches of hiring the consultant, finding the budget and getting the final report to the donor on time are more memorable than their conclusions or relevance for other projects or countries. If this is the case, it is unlikely that IRC staff use evaluation findings to adjust existing or design new projects.

This exaggerated scenario is meant to suggest the poor incentives that exist for IRC (and other NGO) staff to focus on evaluation as a key resource. When asked about current practices, a diverse selection of technical and country program staff reported that IRC:

- Evaluates mostly when donors ask for it;
- Evaluates as an afterthought and without planning;
- Does not use evaluation results to make decisions;

- Cannot use the results because of the poor quality of the evaluations themselves.

Indeed, the motivation to do and use evaluations seems directly related to their quality. If you've read more than one, you have probably had the reaction that most staff do: "hmmm…. What did *that* say?" This is not surprising, considering that we often do not see the three basic ingredients to a useful evaluation. In short, evaluation is most useful when it is:

1. **Systematic**: Evaluation should be a systematic process that is planned and purposeful, not an afterthought.
2. **Well designed**: Evaluations are not all alike; every evaluation's design needs to be appropriate to the context, project and data available, and consistent with IRC's purpose for conducting it.
3. **Used**: Many people would suggest that if the intent to use an evaluation's findings is lacking, it is not worth the investment in resources. Likewise, if evaluation is solely a bureaucratic exercise – done to fulfill a donor's requirement, for example – it is possible that staff will not see enough value in its findings to use them.

Key Point

Evaluations must be systematic, well-designed and used.

II. Conceptual Foundations of Evaluation

How do we change evaluation so that it is planned, well done and used at IRC? First, we need to understand the role that evaluation plays in the project cycle and the very real differences that exist among types of evaluation.

Evaluation in the project cycle

The ***project cycle*** is a tool for understanding the sequence of tasks and management functions performed in the course of a project's lifetime. The phases in the cycle are progressive, and meant to make explicit how data collection and analysis should support decision making throughout implementation.

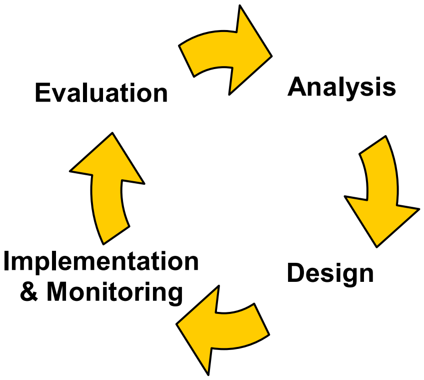
Evaluation is a key engine of the project cycle. It can and should be done during and after implementation, as a way to bring new information into decision making. To put evaluation to work for IRC, we need to understand two central relationships suggested by the project cycle.

Key Term

Project Cycle

A tool for understanding the sequence of tasks and management functions performed in the course of a project's lifetime.

1. **The relationship between design and evaluation.** To overuse an old saying: "if you don't know where you're going any road will get you there." In other words, the quality of learning IRC can acquire with any given evaluation depends in large part on the quality of the project design. The project strategy in the log frame defines what the project is conceived to accomplish. Without it, an evaluation lacks direction and focus. Even more to the point, if a project's design lacks clear logic or feasible objectives, an evaluator is likely to spend a good part of his or her time clarifying what IRC intended to do with the project, rather than assessing to what degree IRC did it.
2. **The relationship between monitoring and evaluation.** Although the term "M&E" seems to denote that they are one and the same, monitoring and evaluation activities are distinct. Monitoring is a program management function and should serve the immediate information needs of IRC staff. It takes place within a project and is carried out by staff. The resulting data should provide a regular flow of decision oriented information. If it is done well, it can capture small, but important behavioral changes (the "effect" layer in IRC's logical framework) among key populations with which we work (teachers, community health workers, refugee leaders, for example). Evaluation is intended to be more of a performance based assessment, conceived to bring into decision making information that is not as easily attained through every day data collection and analysis. It is a common misconception that evaluation can answer questions about our work without any help from monitoring data and despite the absence of it. In fact, an evaluation's final analysis can be significantly strengthened if it is complemented by data from ongoing project monitoring.



? Why is monitoring a project strategy so important?

Good monitoring can provide us with essential information about whether the project is being implemented according to plan, and if we're seeing the immediate and direct changes that result from the outputs and will lead to the objectives.

Useful monitoring, however, requires that projects have strong indicators for the different "layers" or rows of the project logical framework. Recall that the IRC log frame has five rows: goal, objectives, effects, outputs and activities. Output indicators tend to be easier to monitor, and they let us know if we are implementing our project according to plan. But the information they provide is limited to things like the number of trainings held or the number of condoms distributed – the direct goods and services resulting from our activities. Monitoring *effect* indicators, however, gives us valuable information about the direct and immediate changes in the population that result from the delivery of the project outputs. The value of monitoring the *effect* level of the IRC logical framework is twofold: effects are more immediate and direct and therefore can be monitored more regularly than objectives; and, unlike outputs, effects provide us with information about the changes in the population that we expect will cause us to achieve the project objectives.

NOT ALL EVALUATIONS ARE CREATED EQUAL

Imagine an IRC project conceived to decrease the worst form of child labor in West Africa. Four evaluators set out to measure the project's impact. They present an equal number of ways to accomplish this.

1. The first compares ***baseline*** and follow-up data from a survey of 1,200 youth living in the target communities to see if WFCL had decreased as a result of the program.
2. The second reviews project documents and monitoring data, interviews staff and conducts focus group discussions with youth to see whether the program has achieved its objectives.
3. The third conducts interviews with youth, families and teachers to ask people whether they are satisfied with the program.
4. The fourth facilitates a participatory workshop with program staff and other stakeholders to discuss accomplishments they are most proud of, challenges they face in the program and how they could resolve them moving forward.

Key Term

Baseline

An adjective that describes information gathered before a program begins. Baseline data establish the conditions in relation to the program objectives, effects and outputs before implementation.

This small selection of hypothetical evaluations illustrates a simple point: one reason why evaluation is not useful is that evaluators falsely assume that different kinds of evaluations can answer the same questions. Not specific to IRC, there is a common misunderstanding that all evaluation is good evaluation when, actually, the quality of an evaluation depends largely on if it uses the right design and data collection methods to produce the knowledge we seek.

Indeed, there is not one way to do program evaluation. Neither a standard definition of concepts like "impact" or "effectiveness," nor a one-size-fits-all approach to evaluation exists. There are, however, different types of evaluation. Each serves a different purpose, has limitations and can answer a unique set of questions about our work. Recognizing this basic fact can help

Case Study

The “M&E” connection

Evaluation is by no means the only way to answer our questions about the progress of IRC projects. In fact, there’s reason to believe that project monitoring, if done well, can be a much more useful tool for project management. Consider the following example:

IRC’s Community-driven Reconstruction (CDR) project in Liberia seeks to increase communities’ participation in civic activities, improve household assets and increase trust and cohesion. A rigorous impact evaluation (a survey conducted in program and non-program communities at the beginning and end of the project) will assess whether or not the project achieves these objectives. The results, however, will not be available until the end of the project and, even then, will fulfill a distinct learning purpose: to uncover whether or not the CDR project “worked” or was effective at achieving its aims. Regular monitoring of outputs and effects (i.e., what did IRC produce and what small behavior changes did the people with which we worked – in this case, community members, CDC leaders, maybe local government officials) is essential for day-to-day learning about what is or is not happening as a result of our activities. Monitoring data will also be important to the final analysis of impact, because they can tell us more about why or why not observed results may have been achieved (for example, how many hours did NGO staff spend in the communities that exhibit the highest level of impact? How many CDC meetings were held? Were community contributions more?). See “Monitoring and Evaluating a Project Strategy” in the concluding section of the guidelines for more information on this topic.

staff engage with evaluators and donors so that IRC gets useful information from the resources spent on evaluation.

Key Point

There are different types of evaluation. Each serves a different purpose, has limitations and can answer a unique set of questions about our work.

In this section we identify and describe **three major types of evaluation**: process, participatory and impact evaluation. The intention is to give you a starting point for any evaluation: a textbook-like categorization of the different types of evaluation, their respective purpose and the questions they can answer. In reality, these evaluations are not mutually exclusive, and it is not uncommon for a program to be evaluated using a mix of approaches at different times. What is essential is that all evaluations are purpose-driven — meaning that the evaluation design is determined based on how the evaluation will be used and what questions it seeks to (and can) answer.

Process evaluation

Process evaluation is the most frequently used type of program evaluation. It serves a crucial purpose: to improve and inform decisions about implementation by providing a better understanding of operational constraints and opportunities. A process evaluation can provide management staff with data on whether or not a program is functioning as it was intended. It focuses on the program’s structure and activities and verifies whether implementation is or is not in fact happening. It is usually directed at two key questions: whether a program is reaching the appropriate target population, and whether its implementation and support functions are consistent with the program design or other standards.

A good process evaluation does not require a lot of planning or resources. Because it is a snapshot of implementation, it can be done more than once during the life of a project. If the project lasts more than a year, process evaluations can be helpful tools mid-way through and toward the end of the project. The trick to a good process evaluation is to understand the kind of knowledge it can and cannot provide.

Key Point

The purpose of a process evaluation is to assess whether a project is being implemented according to its design. A process evaluation will not tell us if an IRC project has achieved its objectives or had its intended impact.

A process evaluation can be conducted in a variety of ways, depending on the project design, data and time available. An evaluator should always use the program's existing monitoring data as a starting place for analysis, but he/she might supplement this with additional data collection depending on the evaluation's purpose and questions: a structured assessment of the schools or local organizations with which IRC works, a series of key-informant interviews, or focus group discussions to help the program determine to what degree community members are aware of IRC's work are possible examples.

Most evaluations of IRC projects are process evaluations. This is not always by design, as the data, resources and planning usually available to us can limit the options by default. To make sure that we get the most out of this particular kind of evaluation remember the essential points below.

ESSENTIAL POINTS

- **The purpose of a process evaluation is to assess whether a project is being implemented according to its design.** A process evaluation will not tell us if an IRC project has achieved its objectives or had its intended impact.
- **A process evaluation can and should be used strategically:**
  - When there's a relatively new program and answers about how well it has implemented its services provides useful feedback to managers and donors;
  - When questions arise about how well a program is organized, whether the intervention is being delivered in a standard way

? Don't we know enough about the process just by working on the project?

To be effective in bringing about desired improvements in people's lives, an IRC project needs more than a good design and plan. Most important, it needs to implement its plan – it must carry out its intended functions in the intended way. Although implementation may seem straightforward, in practice it is often very difficult. The result can easily be substantial discrepancies between the program as intended and the program as actually implemented. Implementation is reflected in the process that it puts in place. An important evaluation function, therefore, is to assess the adequacy of the process: the program activities that actually take place and the services actually delivered.

Adapted from Rossi, Lipsey and Freeman (2004) and according to its design, and the success with which it is reaching the targeted population;

- When a program is presumed effective so that the most significant issue is whether that service is being delivered properly or according to a certain standard (Sphere standards, for example);
- As a complement to good project monitoring and all other kinds of evaluation, to assure that implementation is happening as expected and necessary for the objectives to be achieved.

?

What does satisfaction tell us?

An IRC education and vocational training program conducted a final evaluation to determine if the program had its intended impact. An external evaluator conducted interviews with staff and program beneficiaries – teachers, parents and students. Beneficiaries were asked what they liked about the program, and what changes they wanted to see. The final report indicates that "parents are extremely happy with the program" and "students were pleased with the courses, and felt that they would stand a good chance to get work." The report concludes that the program achieved its objectives and "was a great success."

Assessing people's satisfaction with a program is important – it gives people an opportunity to tell IRC what they are pleased with and what they think should be done differently. But people's satisfaction with a program is not an objective measure of IRC's intended impact or fulfillment of objectives. These would be indicated by distinct measures or indicators inferred from the objectives of the program, whereas "satisfaction" is a measure of people's feelings about IRC. In fact, it may not even provide valid information about this, as people may have many reasons to say that they are or are not happy with IRC services.

Key Point

Participatory evaluations are more about self-reflection and empowerment, than measurement of impact or analysis of process.

Participatory evaluation

Participatory evaluation involves stakeholders and beneficiaries in a collective examination and assessment of a project. It has its roots in the same schools of thought that motivate participatory development and stems from a similar dissatisfaction with "top down" aid and consultants who "fly in and fly out" to assess change in a local context.

Although participatory evaluation has been somewhat mainstreamed – promoted by USAID and the World Bank, for example – its underlying philosophy is different from those driving other kinds of evaluation. Put simply, it is more about self-reflection and empowerment, than measurement of impact or analysis of process.

The purpose of participatory evaluation is two-fold. First, it seeks to improve program effectiveness by incorporating the input of people closely involved in or affected by implementation. These can include beneficiaries, local organization leaders or members, donors, staff or any range of individuals connected to IRC's work and depending on the scope of the evaluation.

Participatory evaluation also seeks to build people's capacity to reflect on challenges confronting the program and negotiate solutions amongst themselves. As with other participatory methods, the emphasis is on the process rather than the product of the evaluation. Stakeholders and beneficiaries do more than provide information. They decide on the Terms of Reference, conduct the data collection, analyze findings and make recommendations. The evaluator is a facilitator, animating workshops, guiding the process at critical junctures and consolidating the final report based on the findings of the stakeholders. As a result of the active involvement of stakeholders in reflection, assessment and action, a sense of ownership is created, capacities are built, beneficiaries are empowered and lessons learned are applied, thereby increasing effectiveness.

Key Point

It is crucial to recognize the need for experienced, skilled facilitators when you are thinking about whether or not to plan a participatory evaluation.

Sounds easy, right?

It isn't: facilitation requires significant skill and experience, process can be overwhelming or underwhelming depending on the situation, and it is challenging to catalyze meaningful participation. It is crucial to recognize the need for experienced, skilled facilitators when you are thinking about whether or not to plan this kind of evaluation.

It is also important to understand that there is a continuum of approaches to participatory evaluation. At its most extreme, participatory evaluation represents a belief that we cannot learn anything "true" about our work or its effects on communities because there is no such thing as an observable, objective reality. Known as "naturalistic," "constructivist" or "fourth-generation," this kind of evaluation is hard to understand in terms of what it actually *looks like* in practice. It seeks to negotiate consensus among stakeholders with different perceptions of the issues and sometimes different views of the very meaning of the program.

The more mainstream version is represented by USAID's "Performance Monitoring and Evaluation Tips" (see the recommended reading for a reference). In this view, participatory evaluations are guided by a specific purpose rather than an overall ideology; they are "especially useful when there are questions about implementation difficulties or program effects on beneficiaries, or when information is wanted on stakeholders' knowledge of program goals or their views of progress."

Both views are based on the assumption that giving people the opportunity to articulate their needs, interests and expectations and engage in a shared analysis of the program can empower them toward action, resolving differences and thereby improving the quality of a program.

As with process evaluation, there are several essential points to remember about participatory evaluation regardless of which approach you consider:

ESSENTIAL POINTS

- **The purpose of participatory evaluation is twofold**: to improve program effectiveness by incorporating the input of people closely involved in or affected by implementation; and to build people's capacity to reflect on challenges confronting the program and negotiate solutions amongst themselves.
- **Qualitative data collection is not the same as participatory evaluation.** If a consultant tells you that he/she is going to do a participatory evaluation, ask what this means to them. Make sure that he/she is not confusing focus group discussions and interviews, conceived to answer pre-defined questions about the program, with participatory evaluation. Although a participatory evaluation may involve these methods, it is not defined by them and usually connotes a different approach to how they are done and by whom.
- **A participatory evaluation must be systematic.** Regardless of the type of evaluation you choose, it should follow a consistent and orderly process of analysis with attention to detail. Being a facilitator for a participatory evaluation requires a tremendous amount of skill and should not be taken as something anyone can do simply because it *sounds* less scientific or rigorous.
- **A participatory evaluation is not always the right choice, just because it is a good thing to do.** Its intent to use participation throughout all phases of the evaluation (planning and design; gathering and analyzing the data; identifying the evaluation findings, conclusions, and recommendations; disseminating results; and preparing an action plan to improve program performance) makes participatory evaluation very consistent with the Program Framework and our intuitions about how best to work with people. This does not mean that it can answer all our questions. As with other evaluations, it serves a particular purpose and can answer certain questions but not others.

IMPACT EVALUATION

What is "impact" and why is it so hard to evaluate?

Key term

Impact

The difference IRC makes; changes in outcomes experienced by people/communities for which IRC is responsible.

How do we know if IRC's work *really* makes a difference in people's lives? How do we know if one approach "works" or is more effective than another? How do we know if a program that works in one context will work in others? What would we need to do to be able to say that a given project should be "scaled up" to reach a larger population or replicated in more communities? Impact evaluation seeks to answer questions like these.

Its very definition, though, takes more than one shape in the aid community. You can ask a handful of NGOs what is *impact evaluation,* and they will give you the same number of different answers. You might hear:

> *"We measure impact through participatory evaluation techniques because nobody knows better how to assess changes in people's lives than the people themselves."*
>
> *"Measuring impact is impossible in the settings in which we work. Things are too fluid and it is unethical to use control groups."*
>
> *"We require baseline and follow up surveys of all or our programs, thereby documenting the changes that happen by comparing the situation before and after."*

The way you conceive of impact evaluation depends on how you define ***impact***. When asked, many IRC staff members say that they see impact as the changes that can be ***attributed*** to IRC's work – or the change in outcomes for which IRC is responsible. Strictly speaking, any attempt to assess whether or not IRC achieves the objectives of a given project refers to this notion of change. This is the definition we use for impact, and for the following discussion of impact evaluation and measuring effectiveness.

Key Term

Attribution

The ability to make a connection between the changes we observe or expect to observe and a specific intervention; the ability to credit the changes or the results achieved specifically to IRC.

An evaluation that can measure impact is extremely challenging in terms of the quality and quantity of data required. This is because there are so many different things that influence the outcomes we seek that it is very hard to isolate the difference IRC makes. The possibilities are almost limitless and include both inherent differences of the communities with which we work (for example, observable characteristics such as a community's proximity to the capital, as well as unobservable characteristics such as a predominant belief, or a lasting tension with another ethnic group) and external events that may

CASE STUDY

Were the project objectives achieved or... Did the project achieve its objectives?

Consider an IRC education program that seeks to improve access to education in 2 communities, both in remote areas. The project provides material support to the schools (equipment, latrines, books and teaching materials), helps to develop parent teacher associations, leads an outreach campaign to attract kids who do not attend school regularly, and creates an innovative teacher training component to improve the quality of teaching so students feel safe and protected. Monitoring data show that enrollment in both schools increases over the duration of IRC's project. Can IRC claim that its project caused enrollment to increase? Not necessarily. We can think of many factors that have nothing to do with our program that may have led to increases in enrollment. For example, during the course of the project, nearby schools with which IRC was not working may have closed, forcing children to enroll in the IRC-supported schools. The government may have reduced or abolished school fees, resulting in increased enrolments around the country. Other factors that are more difficult to ascertain may be the cause of increased enrolment as well. For example, a change in farming practices might decrease the need for children to work in the home and therefore increase their ability to go to school, or even an environmental change may make it easier for children to attend school regularly. The point is that even if IRC sees changes that our project set out to achieve, this does not mean that our project actually caused those changes to happen.

"...A systematic review of the United Nations Children's Fund (UNICEF) estimated that 15 percent of all its reports included impact assessments, but noted that "[m]any evaluations were unable to properly assess impact because of methodological shortcomings" (Victora 1995).

Center for Global Development (2006)

happen while we are implementing and have a strong influence on the change IRC seeks to effect (for example, a national government's new policy in a relevant area).

Key Term

Randomization (or Random Assignment)

The process by which the populations with which IRC works are selected randomly – think of pulling names out of a hat – to assure that there is not a selection bias that could otherwise explain changes observed in the results of the program. Those with which IRC works are known as treatment; those with which IRC does not work are the control.

This reality means that an impact evaluation has to answer almost impossible questions. How would individuals who participated in the IRC project have fared in the absence of IRC? How would people who did not participate in the IRC project have fared if they had participated in it?

The use of randomization in impact evaluation

There is a growing body of theoretical knowledge and practical experience on the use of ***randomization*** to answer these questions. In short, randomization compares a group that is given the program (treatment) to a group that is not given the program (control). The real deal breaker, though, is how these groups are selected. If you choose a control group of communities adjacent to the ones with which IRC works, how can you know that there is not something particular about where they live, for example, that makes them different from the IRC program communities? Say, for example, that the control communities are closer to a river or to a central market; or that their ethnic composition is more heterogeneous than the ones where IRC works. Any number of things – some observable, some not – can likely explain the difference in outcomes we might see between the communities where we work, and those where we don't.

The challenge is to figure out a way to compare communities that are not systematically different from one another. Then, the only systematic difference is IRC's presence: any difference in the outcome experienced by the two groups can be attributed to IRC. One situation where this so-called this becomes possible is when the two groups are selected randomly from a potential population of participants (such as individuals, communities, schools or classrooms). In this case, on average, we can be assured that those who are exposed to the program are no different than those who are not; a statistically significant difference between the groups in the program's outcomes can be confidently attributed to the program.

So what does all this language mean in practice?

? When does IRC use randomized evaluation?

Whether or not IRC can use this approach in a particular instance depends on the nature of the project, its duration and scale, the availability of partners and resources. The key to using this kind of evaluation is to identify the particular sectors, approaches and/or questions that are truly representative of what IRC – and/or the rest of the aid community – needs to know to do its work well or pursue its overall aims. Consider this option, in particular, when IRC has an opportunity for large-scale, multi-year work in non-emergency contexts. The program design should be representative of our programming and overall learning needs. These evaluations are spearheaded by or planned in consultation with REL.

The evaluation is comprised of several related data collection events that contribute to a final analysis that allows us to identify with some certainty the project outcomes that are attributable to the program itself. Rather than comparing outcomes through looking at data before and after the program is implemented, the key analysis is of the outcomes experienced by randomly selected treatment and control communities.

? Is random assignment ethical?

This is likely the most common response that people have to the idea that we randomly select the populations with which we work. It is important to distinguish between interventions that are already known to be life-saving and those that are not. IRC would never randomly select beneficiaries in an emergency context or for an intervention already known to save lives. With respect to other program areas, though, a quotation from the seminal report on impact evaluation helps to make clear the rationale:

> *"Impact evaluations that rely on collecting data from control groups are sometimes thought to be unethical because they exclude people from program benefits. But this criticism applies only when resources are available for serving everyone as soon as the program starts. In fact, whenever funds are limited or programs need to be expanded in phases, only a portion of potential beneficiaries can be reached at any time. Choosing who initially participates by lottery is no less ethical (and perhaps even more so) than many other approaches. Some programs are allocated by lottery when they are oversubscribed (school choice in the United States or voucher programs in Colombia) or for transparency and fairness (random rotation of local government seats to be set aside for women in the Indian elections). Furthermore, whenever there is reasonable doubt of a program's efficacy or concerns with unforeseen negative effects, ethics demands that the impact be monitored and evaluated. For example, in Mexico opponents of a conditional cash transfer program in the mid-1990s argued that giving funds to poor mothers might increase their vulnerability to domestic abuse. A well designed impact evaluation was able to put those serious concerns to rest. The simple truth is that many well intentioned social programs are like promising medical treatments—we cannot really know if they do more good than harm until they are tested. Finally, starting with a properly evaluated pilot program can greatly increase the number of eventual program beneficiaries, because the evidence of success will provide support for continuing and expanding an effective program… Poor quality evaluations are misleading. No responsible physician would consider prescribing medications without properly evaluating their impact or potential side effects. Yet in social development programs, where large sums of money are spent to modify population behaviors, change economic livelihoods, and potentially alter cultures or family structure, no such standard has been adopted. While it is widely recognized that withholding programs that are known to be beneficial would be unethical, the implicit corollary—that programs of unknown impact should not be widely replicated without proper evaluation—is frequently dismissed."*

Center for Global Development (2006)

The steps to a randomized evaluation resemble the following:

1. Selection of a sample size large enough to produce statistically powerful results – this depends on both the program and evaluation designs;
2. Identification of the hypotheses driving the evaluation and specific outcome measures (i.e., what are we looking to learn through this evaluation about and related to our program?);
3. A baseline household survey conducted across the sample population;
4. Random assignment of 'treatment' and 'control' communities/schools/villages (depending on the program and evaluation designs);
5. Follow up survey at the end of the project;
6. Monitoring data including and above the 'output' level to increase the data available for analysis;
7. Process evaluation to show that activities and outputs did in fact occur, giving us reason to expect the impact we seek;
8. Analyses of program impact based on all surveys, monitoring data and process evaluations.

"A review of 127 studies of 258 community health financing programs found that only two studies were able to derive robust conclusions about the impact on access to health services (ILO 2002)."
Center for Global Development (2006)

What is perhaps most important to remember is that, like other types of evaluation, this one has a particular purpose, can answer some questions and not others, and should not be considered the right way to evaluate under all or even most circumstances. Rather than serving to improve a project's implementation, impact evaluations serve the institutional purpose of knowing what works to accomplish certain objectives. They can build knowledge that we can apply in future designs and sometimes other contexts – rather than creating learning that we can use in the current program. Specifically, this means that randomized evaluation can be used to inform organizational decisions about resource allocation, replication, and scaling-up of interventions within and across countries.

They are appropriate tools to:

- Learn 'what works' in our priority sectors and contexts;
- See if a certain innovation or pilot program works;
- Test one approach against another when we have different ideas about how best to accomplish a priority IRC objective.

These evaluations require resources, planning, data and partners that the other types do not, and findings take longer to produce, making them much less useful for improving current implementation. In fact, experts *do not recommend* that all aid projects should be evaluated in this way. The time, effort and high levels of data required for randomization make it infeasible; the purpose and questions it can address make it more appropriate for strategic rather than continuous usage. Finally, it is best used when we know that IRC implements a given approach well enough to have a reasonable chance of producing the intended results (in other words, when documentation, project related data and conventional evaluation give us reason to suspect that a particular IRC approach is effective; in contrast to a one-time approach that we have never implemented and about which we know little).

What's next best? How to evaluate different levels of effectiveness

So if the applicability of randomized evaluation is so limited, what about all the other circumstances – i.e., most of IRC's programming? What is "next best" if we want to measure our impact and cannot do a randomized evaluation in a given circumstance?

The chart on page 17 shows the characteristics of three levels of program effectiveness: apparent, demonstrated and proven (which refers to the use of the randomized impact evaluation approach just described). The first column defines each level and indicates what questions it can answer about IRC's work. The second column specifies the kinds of information that is collected and the types of evaluation activities required to assess program effectiveness at this level.

Use this chart when you and your team design an IRC project and/or are working on a proposal to a donor. Any of the levels require planning before implementation starts. Most IRC evaluations should be geared to measure "apparent effectiveness." Health programs or other sectors where there is a good amount of existing knowledge about which interventions work and don't can likely seek "demonstrated effectiveness" if they fit the criteria listed in the chart. Please remember that because it is based on new methods and partners, "proven" effectiveness is something that is not pursued in isolation, but in collaboration with technical units and REL at the earliest stages of project planning.

Here are some final **essential points** to consider when thinking about these different levels of program effectiveness:

Key Point

Evaluating the apparent, demonstrated and proven effectiveness of a program requires planning before implementation begins. Most IRC programs should seek to measure apparent effectiveness.

ESSENTIAL POINTS

- **IRC defines "impact" deliberately** as the changes in people's lives that can be *attributed* to IRC's work, or the change in outcomes for which IRC is responsible.
- **A baseline study is sometimes necessary but never sufficient to measure impact.** Baseline is an adjective that describes an initial measure of something before an intervention – or a crisis, war or other event that will likely change whatever it is we're observing. It is conventional wisdom among many aid actors that baseline and follow up surveys provide a ***valid*** means to measure the difference that a given intervention makes for people's lives. Consider, though, all the things that happen in addition to the intervention and the likelihood that any of these can and probably will influence the outcome IRC seeks. The assumption that these two states of the world (before our project and after our project) are similar in every other way except for IRC's project does not hold true. There is no reason why baseline and follow-up studies will yield a good measurement of the change IRC

Key Term

Valid

Data are only useful if they actually measure what they are supposed to be measuring. The concept of validity refers to the extent to which the data we collect give a true measurement or description of "social reality".

Assessing the Impact of IRC Programs

Three Levels of Effectiveness

Level of Effectiveness	Data Collection & Evaluation Activities
Apparent A program evaluated at this level should be able to answer the following questions: Did we see the changes in the objective and effect indicators as we expected? Did the changes in the program beneficiaries and communities happen as we expected?	Requires project monitoring above the *output* level and baseline data collection (not necessarily quantitative) that gathers initial data about the relevant experience of program participants. Requires a follow up analysis that tracks the data from the baseline, good monitoring data and a process evaluation.
Demonstrated A program at this level should be able to answer the question: Do participants experience better outcomes than people who are not in the program?	In addition to the same data collection as above, demonstrated effectiveness requires; 1. objectives that are observable, very proximate to IRC activities and already sufficiently understood through research and best practice; 2. secondary data that show the objective measure for similar populations (that IRC does not work with) in the area; 3. data on the presence and activities of other actors working with the IRC populations to illustrate that IRC's influence is likely (see example in "Maternal Mortality in Pakistan" Box).
Proven A program at this level should be able to answer the following question: Are there statistically significant differences in outcomes for IRC program participants versus people in a randomized control group? The project's impact on participants has been scientifically confirmed through experimental or quasi-experimental evaluation design. This approach enables IRC to "prove" the connection between its activities and the objectives sought, assuming the objectives are feasible and measurable.	Requires an independent, external evaluator to design and conduct the evaluation. Requires random assignment of the beneficiaries with which IRC works. This could be a "true" treatment and control, as well as a phasing in the timing of implementation and comparison of populations IRC works with over time. Demographic and outcome data for both groups are collected and compared; the sample is sufficiently large to conclude statistically that the program is responsible for the differences observed. Final analysis of impact needs to be supported by ongoing monitoring of outputs and effects, and a process evaluation.

effects given everything else that happens any community. This does not mean that a baseline comparison cannot be helpful. It does mean, however, that we have to collect *more* data after the baseline during implementation, to determine whether what we expect to see as a result of our activities is in fact happening.

- **We should measure everything we can see along the logic of our design.** Given the point above, the trick is to identify the data we need to show that IRC is responsible for any changes we can observe between baseline and follow up exercises. If you consider the project strategy of your project to be a "theory" – a causal statement of what we expect to see happen as a result of what we do – there are several points along its logic that indicate the information we need to gather. This is why monitoring and thinking about how to do more than count "outputs" is so important to our understanding of our work.

- **A baseline study is not necessarily quantitative.** Though baselines are often associated with quantitative surveys, a baseline study can also capture qualitative information. The purpose of a baseline is to assess change over time by collecting data before an intervention and then again over the course of the program. Collecting qualitative baseline and follow-up data can be challenging because it requires asking the same questions to the same people over time, and analyzing changes in qualitative responses while minimizing bias. Though qualitative baseline and follow-up data will not be

Key Point

There is no reason why baseline and follow-up studies will tell us the change IRC effects given everything else that happens in any community.

CASE STUDY

Collecting qualitative baseline data in an IRC project

An IRC protection project seeks to conduct a baseline that will assess recently returned refugees' experiences with law enforcement and other government officials. The project proposal described a baseline survey that would assess the percentage of returnees who are treated with respect by local government officials. The project team planned to conduct the survey at the beginning and end of the project to tell them if the way the returnees were treated improved over time. As the team began to think through how to assess the treatment of the returnees, they realized that the information they needed was highly *qualitative* – complex, nuanced information about people's experiences and if they feel they have been respected – and perhaps a survey was not the best way to get this information. Furthermore, they realized that the number of returnees in the project communities was changing significantly everyday, and to try to assess the experiences of a representative sample of all returnees would be impossible.

Given these circumstances, the team decided to conduct a qualitative assessment involving a small sample of returnee families. The assessment would use in-depth interviews to capture information about these families' experiences and feelings upon return; interviews with the same families would be repeated several times throughout the life of the project. Although this kind of study would not allow the team to draw conclusions about the experiences of *all* the returnees, nor would it allow them to make claims about the project's impact, it would enable the IRC to capture a more complete picture of how different families' described being treated, and how their experiences changed over the course of the project.

representative of a population, they can provide valuable insight into changes that cannot be captured in a quantitative survey.

- **Seeing is believing.** There are some instances where an impact evaluation is just not necessary. Inspectors for a federal oversight agency found that in a sampling of eight rebuilding projects that the United States had declared successes, seven were no longer operating as designed because of plumbing and electrical failures, lack of proper maintenance, apparent looting and expensive equipment that lay idle (New York Times, April 29, 2007). When outcomes are as obvious as this, there's no need to consider any additional measurement efforts. Likewise, the challenge of measurement is easier when the outcomes we seek to influence are easier to see, as is the likelihood that IRC is responsible for them.

  This is the rationale underlying the "demonstrated effectiveness" level. Say, for example, that IRC is the only NGO working in a camp setting where morbidity from contaminated water is very high. IRC works to provide people with access to clean water and improve hygiene-promoting behavior. When data reveal that participants in the IRC programs are healthier and suffering less from water-borne diseases, it is within reason to conclude that IRC is responsible for the positive changes.

- **The more we know, the easier it is to measure impact.** It is much easier to know whether or not we achieve our objectives when there is a body of knowledge about the relationship between interventions and these objectives. The existence of an "evidence base" – sound research that substantiates a theory – in the health sector, for example, explains why many health outcomes are well understood as are their link to commonly understood interventions (the case study below on maternal mortality in Pakistan provides an example).

- **Design always counts.** It is a waste of time, effort and resources to attempt to estimate the impact of a program that lacks plausible,

CASE STUDY

The more we know, the easier it is to measure impact.

Imagine an IRC program that seeks to reduce maternal mortality rates in 2 areas in Pakistan. IRC provides Emergency Obstetric Care Services in the area's clinics. How do we know if we reduce maternal mortality? Research on maternal health has established that skilled delivery of babies is an essential determinant of whether or not mothers survive giving birth. Evidence of this fact provides a road map for IRC programs that work to reduce maternal mortality, guiding them to make sure that EMOC facilities provide the procedures, medicines and support that define "skilled delivery." This evidence base also guides IRC data collection. With credible surveys on maternal mortality in the areas where IRC works, data on utilization and coverage of the EMOC clinics, and consistent monitoring data on the components of "skilled delivery" offered in the clinics, IRC can investigate its program impact with some confidence without a randomized evaluation design. The resulting analysis would qualify as a project with "demonstrated effectiveness."

measurable outcomes or clearly defined program logic. A good project design (see Data Driven Issue: "The Essence of Good Design") will help make any kind of evaluation easier and more useful for IRC.

- **Information about people's satisfaction with IRC's work is not a measure of whether or not IRC achieved its objectives.** Depending on the questions and context, it may not even be a measure of people's satisfaction with IRC because the reasons people have to answer these questions are not always transparent.

? What about emergencies?

Real-Time Evaluations (RTE) have become increasingly popular tools to learn from and inform programming, primarily during emergencies. Similar to monitoring, "the key principle underlying real time evaluations is that it can affect programming as it happens" (ALNAP, 2005). Although there are many different definitions and approaches, real time evaluations tend to be: Carried out during implementation, usually in the early stages of an emergency and ideally repeated throughout the project cycle; conducted by internal teams or individuals who are not directly involved with or responsible for the program; and short in duration.

RTEs are not designed or carried out in a way that allows for any conclusions about impact as we define it here. "The judgments made in real time evaluations in general concern *how* results are being achieved: they look at process rather than impact. Impact is not only in general harder to pin down, but given the timing of a RTE, meaningful statements about impact cannot be expected" (ALNAP, 2005).

Should IRC conduct Real-Time Evaluations? It is impossible to generalize about real-time evaluations since they lack a coherent definition and consistent methodology. Properly designed and carried out real-time evaluations may tell us if we are implementing a program or emergency response well, and how we might adapt the program to improve implementation. But it will not tell us if our program is effective or having an impact on the population it is serving. Whether we call this a mid-term evaluation, process evaluation, 'real-time evaluation' or even monitoring is not important. What is important is that we are explicit about the purpose of the evaluation and careful about the claims we can make based on the evaluation design. See the Humanitarian Practice Network's *publication on Real-Time Evaluation* for more information.

Process Evaluation

Purpose:

To improve implementation.

Unique characteristics:

Can answer questions such as

- Is the program being implemented according to its design?
- Are their flaws in the program's design?
- Are the necessary inputs in place to implement the program?
- Are the appropriate staff and management structure in place to deliver the program?
- Is the program reaching the intended people (for example, vulnerable children, victims of gender-based violence, out-of-school youth)? How many people are being reached? Are they getting the intended goods and services?
- Are members of the community aware of the program?
- Are the actors (teachers, health workers, community mobilizers) being trained to use the new method/practice/tool/intervention? Are they using it?
- Are there ways of improving cost effectiveness (for example, substituting expensive inputs with less costly alternatives, substituting costly inputs with labor, delivery methods)?
- Are beneficiaries satisfied with IRC's program? What is the value of the program to the intended beneficiaries?
- Are there needy but un-served people the program is not reaching?
- Are administrative, organizational and personnel functions handled well?

Participatory Evaluation

Purpose:

To improve implementation; to strengthen participants' data collection and evaluation skills; to engage specific people in the program and enhance teamwork.

Unique characteristics:

- ***Participant focus and ownership.*** .The purpose of participatory evaluations is to build stakeholders' ownership and commitment to the results and facilitate their follow-up action.
- ***Scope of participation.*** The range of participants included and the roles they play may vary.
- ***Participant negotiations.*** Participating groups meet to communicate and negotiate to reach a consensus on evaluation findings, solve problems, and make plans to improve performance.
- ***Diversity of views.*** Views of all participants are sought and recognized. More powerful stakeholders allow participation of the less powerful.
- ***Learning process.*** Emphasis is on identifying lessons learned that will help participants improve program implementation, as well as on assessing whether targets were achieved.
- ***Flexible design.*** While some preliminary planning may be necessary, the design of the evaluation is decided by the participants, not by outside evaluators.
- ***Use of facilitators.*** Participants actually conduct the evaluation, not outside evaluators. Outside experts usually serve as facilitator – that is, provide supporting roles as mentor, trainer, group processor, negotiator, and/or methodologist.

Adapted from USAID (1996)

Impact Evaluation: Proven Effectiveness

Purpose:

To evaluate a project's impact; to measure the "difference" IRC makes; to assess "what works." To make decisions about which programs to continue, replicate, scale up.

Unique characteristics:

Can answer questions such as

- Did IRC's program make a difference?
- Are people who participated in an IRC program "better off" than if they had not participated in the program?

Methodological and other implications

- Impact evaluations must be planned as early as possible in the design stage of a project; the initial steps of an impact evaluation will take place before the project begins.
- Impact evaluations assess project outcomes for people who participated in the project compared to people who did not participate in the project; these groups – participants and non-participants, or "treatment" and "control" – must be identical (i.e., any differences between the two groups must be ruled out or accounted for).
- The best way to rule out differences between program participants and non-participants is to randomly assign people to each group. This requires an understanding of specific sampling and random assignment techniques.
- Impact evaluations require significant time, resources and technical capacity.

III. PRACTICALLY SPEAKING: SO WHAT AND NOW WHAT?

The remaining pages of this guide seek to answer the question: what do IRC staff members actually *need* to know about evaluation in order to improve its overall quality across countries and sectors?

IRC staff members need to know four things:

1. The essence of evaluation: what to expect and why
2. How to write effective Terms of Reference for an evaluation
3. What is IRC's evaluation system and how do we assess the quality of evaluations
4. What to say to donors about evaluation

1. THE ESSENCE OF EVALUATION: WHAT TO EXPECT AND WHY

There is nothing magic about an analysis of an IRC project. Through it, we can only learn (whether it's through a huge household survey, ongoing monitoring, a participatory workshop or a focus group discussion) what the data tell us. There are principles of logic, inference and sound inquiry that hold regardless of what we seek to *know* about our work. Simply put, we can't know our "impact" just by asking someone.

The **essential points** written throughout this document are summarized below and conceived to identify some of these principles in practice. They should serve as the minimum that staff involved in evaluation – those tasked with making sure it happens, writing Terms of Reference, hiring and interacting with consultants, and with learning from past evaluations – should know. Anything that is not clear or easily understood makes for a great starting point for an REL clinic in your country, technical unit or team, or among a small group of staff.

Only the key phrases are listed here; please refer back to the identified page number for the related details.

ESSENTIAL POINTS TO REMEMBER AND UNDERSTAND

1. The purpose of a process evaluation is to assess whether a project is being implemented according to its design. (page 8)
2. A process evaluation can and should be used strategically. (page 8)
3. The purpose of participatory evaluation is twofold: to improve program effectiveness by incorporating the input of people closely involved in or affected by implementation; and to build people's capacity to reflect on challenges confronting the program and negotiate solutions amongst themselves. (page 11)
4. Qualitative data collection is not the same as participatory evaluation. (page 11)

5. A participatory evaluation must be systematic. (page 11)
6. A participatory evaluation is not always right, just because it is a good thing to do. (page 11)
7. IRC defines "impact" deliberately as the changes in people's lives that can be *attributed* to IRC's work, or the change in outcomes for which IRC is responsible. (page 16)
8. A baseline study is sometimes necessary but never sufficient to measure impact. (page 16)
9. A baseline study is not necessarily quantitative. (page 18)
10. We should measure everything we can see along the logic of our design. (page 18)
11. Seeing is believing. (page 19)
12. The more we know, the easier it is to measure impact. (page 19)
13. Design always counts. (page 20)
14. Information about people's satisfaction with IRC's work is not a measure of whether or not IRC achieved its objectives. (page 20)

2. How to write effective Terms of Reference

A ***Terms of Reference*** (ToR) should present the purpose of a given evaluation and an overview of IRC's expectations. There are two common problems with IRC ToRs: they overstate the scope of knowledge an evaluation can produce and they dictate the methods to be used rather than leaving the choice up to the evaluator. A good ToR needs to do just the opposite: be concise and concrete in its scope, and leave the specific evaluation design up to the evaluator, who should be able to justify his or her choice of methods and specify what they will produce for IRC in terms of knowledge gained.

Key Term

Terms of Reference

The description IRC writes of an evaluation, intended to identify the project to be evaluated and the purpose for doing so. Terms of reference need to present a feasible task and be written concisely and clearly. Importantly, they need to leave open those issues – such as evaluation design and data collection methods – that will allow IRC staff to assess whether a specific consultant/evaluator will provide the knowledge IRC seeks and understands to be possible.

More specifically, an evaluation ToR should include information about IRC and the program; and details about the parameters of the evaluation. These include:

1. **Background information about IRC and the project to be evaluated:** How long has IRC been working in the country/community? What is the nature of the project – its sector(s), duration and objectives? Who are the important stakeholders: staff, leadership, donors, international and local partners?
2. **The purpose of the evaluation:** The TOR should state upfront why IRC is doing the evaluation and what it hopes to gain as a result. The ToR should also specify the intended user(s) and use(s) of the evaluation and how they are going to be engaged either in the process or after it is complete.
3. **The questions IRC wants answered through the evaluation:** The specific questions driving the evaluation should be clearly detailed to make sure that all parties share an understanding for the kind of

Do we use an internal or external evaluator?

Technical Unit staff members are increasingly asked to conduct evaluations in IRC country programs. Is this the right way to go? If not, why not? If so, what should we consider while planning these internal evaluations? Both internal and external evaluations are legitimate and each has strengths and limitations. In deciding between an internal or external evaluator, factors such as time, budget, and intended uses of the evaluation may lead you to decide that asking an IRC staff member to do the job is a good choice. Indeed, having an evaluation done by someone who knows IRC's issues and pressure points as well as what lessons can be applied from other country programs is a terrific opportunity. The challenge is to make sure that the TU staff member is not, in fact, being asked to evaluate their own performance or contribution to the program's quality. A way around this would be for TU staff members to evaluate projects in countries they do not support. While this would still be considered an internal evaluation, it may prevent conflict of interest and allow for greater objectivity.

knowledge IRC can hope to gain. Although there are many interesting and important questions we want answered about IRC's work, whether or not we can answer them depends on the data available and the evaluation design. The questions should be defined by technical staff with knowledge of the program; they should be as specific as possible, because vague questions usually yield vague answers.

4. **Institutional and technical priorities**: every evaluation provides an opportunity for IRC to explore how staff understand and use the Program Framework or technical approaches. Sector-specific staff in the field and headquarters should be consulted to identify the relevant priorities as well as issues stemming from similar projects in other country programs.

5. **Available data and documents**: This important section should include what data are currently produced by the project, with a particular emphasis on whether or not there is any baseline information, the quality and quantity of ongoing monitoring data and any evaluations already done. The data collection *methods* to be used in the evaluation should be left up to the evaluator and should depend on the evaluation design and purpose.

6. **The timeframe and expected outputs of the evaluation:** In addition to a final report or presentation of findings, an evaluator should be expected to present an evaluation plan or protocol before beginning the evaluation. This document is based on the purpose and questions outlined in the ToR, and specifies the data collection methods he or she will use, the resources (human, financial, logistical) needed to carry out the evaluation, how he or she will ensure ethical data collection, and draft data collection instruments (e.g., questionnaires, focus group discussion guides, observation checklists). Adequate time must be allotted for a review and approval of the plan. The final evaluation report should be as concise as possible and include the evaluation plan described above as an attachment.

Key Point

An evaluator should be expected to present an evaluation plan or protocol before beginning the evaluation, and the ToR should explain that the quality of all evaluations will be judged according to internationally recognized standards.

7. **Finally, although it doesn't need to be written in the Terms of Reference, IRC should alert the evaluator that the quality of the evaluation report** will be judged by IRC's *Research & Evaluation* team. Each evaluator should know that IRC keeps a list of recommended evaluators for ongoing usage and their place on this list will depend upon their rating according to recognized standards. A copy of all evaluation reports should be forwarded to REL for this process and inclusion in IRC's evaluation database. The next section explaining IRC's evaluation system should be made available for evaluators. IRC staff can call on Technical Unit and REL staff for support in reviewing and revising evaluation ToRs at any stage.

3. IRC'S EVALUATION SYSTEM: WHAT IS IT AND HOW DO WE ASSESS THE QUALITY OF EVALUATIONS?

The information in this section should be made available to all evaluators of IRC's work.

In order to stay abreast of program evaluations, consolidate findings and create a reliable source of consultants, IRC's *Research, Evaluation and Learning* (REL) team conducts regular reviews of all IRC evaluations. It is neither feasible nor desirable to standardize quality through the creation of a single format to which all evaluation reports must adhere. Instead, REL judges the quality of an evaluation based on the clarity of its purpose and questions and the degree to which they are answered.

REL uses internationally recognized standards to assess the quality of evaluations (standards adapted from two major sources: the World Bank and OECD DAC). These standards should be made available to consultants hired to evaluate IRC projects.

1. **Usefulness:** For evaluation to influence IRC programs, decision makers must perceive the findings as useful, timely, and geared to current concerns. The evaluation, therefore, must be practical, reflecting the purposes of the country program leadership and technical staff as well as IRC priorities overall (e.g., the Program Framework or sector-specific organizational priorities).
2. **Credibility:** The credibility of an evaluation is determined by the quality of its design and the rigor of the methods used. To be credible, an evaluation must be well conceived, with a clear purpose and questions. Data collection methods and any tools must be appropriately chosen and well used to fulfill the purpose and answer the questions. Any recommendations and all conclusions must be substantiated by the evidence found in the evaluation and presented in the final report.
3. **Feasibility:** A feasible evaluation matches its objectives and design with the resources, data and time available. Procedures related to the evaluation should be practical to keep disruption of the IRC program to a minimum while needed information is obtained.

4. **Reliability:** Evaluation findings must be as reliable as possible, referring to the degree to which repeating the same analysis would produce consistent findings. Reliability will be assessed by the rigor of the methods used, whether or not the evaluation is systematic and the degree to which information sources are assessed critically. The sources of information should be described in enough detail so that the adequacy of the information can be assessed. Data collection methods should be chosen so that they assure that the information obtained is reliable and their interpretation valid.

5. **Relevance:** Evaluation findings and any recommendations must be relevant to the project evaluated and the purpose of the evaluation.

6. **Propriety:** Evaluators should adhere to accepted standards of ethical research, respecting human dignity and worth in their interactions with people so that participants are not threatened, harmed or made uncomfortable. The evaluation should be complete and fair in its examination and recording of strengths and weaknesses of the program so that strengths can be built upon and problems addressed. Finally, the evaluation findings along with pertinent limitations should be made accessible to the people involved in the program.

7. **Independence:** Findings, analyses, and conclusions must be free from bias. This means that consultants chosen to evaluate a project should not also have responsibility for any part of that project. Evaluations must be conducted systematically and present objective, impartial findings.

8. **Completeness:** Evaluation findings must be written coherently, well organized and substantiated by data collected by the program and the evaluation process. Results should follow clearly from the evaluation questions and analysis of data, showing a clear line of logic and support that leads to the conclusions.

9. **Presentation:** The evaluation report and any related presentation must answer the questions posed in the evaluation and seeks to fulfill the intended purpose. The analysis should be structured with a logical flow. Data and information should be presented, analyzed and interpreted systematically. Findings and conclusions should be clearly identified and flow logically from the analysis of the data. Evaluations must include a clear, well written and concise executive summary that highlights the evaluation purpose, questions asked and answered, and the key conclusions.

IRC evaluation database

Consultants and/or country programs must submit evaluations to the respective Technical Unit and REL. All evaluations are saved in a central location available to staff on a protected website and the IRC intranet. Please contact REL (REL@theirc.org) to submit or find evaluations.

IRC evaluators list

Depending on the quality of his/her evaluations, consultants are either added to the IRC list of recommended evaluators, or not. Those who do not receive a strong rating are noted and the resulting lists are available for IRC staff to consult when planning evaluations.

Key Point

IRC's major donors are increasingly focused on evaluation. If you haven't yet seen evidence of it in their proposal and reporting requirements or simply in conversations with them, you will soon enough.

4. What to say to donors about evaluation

Most of IRC's major donors have been bitten by the evaluation bug. If you haven't yet seen evidence of it in their proposal and reporting requirements or simply in conversations with them, you will soon enough. They receive the same reports we do and are at the same tables discussing what to do about evaluation design. The problem will be if they increase their expectations without allocating enough financial resources or changing their time frames to allow for improved evaluation practice.

Some of what we discuss in this guide has no bearing on resources. In particular, increasing our understanding about evaluation designs and the kind of knowledge that different ones produce is more about IRC's effective use of current resources than a need for more. Although more rigorous evaluations require additional financial and human resources, there is every indication that some donors understand this and are willing to contribute the required funds.

Nevertheless, IRC's commitment and efforts to improve evaluation practice are good topics to share with donors in discussion and proposals. There is also great potential to partner with a donor on a rigorous evaluation if the timing, context and program are right. The "Three levels of Effectiveness" is something relevant staff members should understand well enough to use in planning, budgeting, proposals and other communication. Consider the following talking points to use as appropriate.

Donor talking points on evaluation

- **Situate IRC's approach to evaluation within the larger conversation about aid effectiveness that is ongoing in the lay press and policy circles**. IRC recognizes the challenge presented by the current discussion on aid effectiveness as one directly related to how it evaluates its own work. We know that the challenge is not particular to IRC, and that donors and implementing agencies alike are working hard to measure the effectiveness of their respective policies and interventions. We are doing the same with a new headquarters evaluation department, guidelines and policies on data collection and analysis, increased capacity of technical teams and partnerships with donors and, in some cases, academic evaluators.
- **Explain IRC's approach to evaluation.** IRC sees all evaluation as purpose and question-driven. We define words like "impact" and "effectiveness" explicitly each time we use them so that consultants, donors, staff and other key stakeholders share an understanding of the knowledge we hope to gain through any evaluation. We use process and monitoring data to inform programming decisions and assure that

we are accountable to ourselves and our donors. We use participatory evaluation when it serves an identified purpose and allows us to engage beneficiaries and others actively in program decisions, corrections and adjustments. Whether or not we can and do evaluate "impact" – defined as the change that can be attributed to a given IRC project – depends on the project, resources and data available, as well as on the fit of a given project with institutional learning priorities. IRC invests strategically in organizational learning about key approaches and sectors, while improving the quality of data we have about all our work. In sum, we hope to approach each evaluation opportunity with the same discretion, clear about the fit between its purpose, questions and the knowledge we hope to produce.

- **Share with donors IRC's different levels of effectiveness early on in the planning process and request resources according to the data requirements**. IRC defines program impact as the difference in outcomes for which we can say we are responsible – the changes in people's lives we can attribute to IRC's work. IRC is very committed to measuring program impact in all its work and chooses among three levels of effectiveness depending upon context, size of the beneficiary population, duration and nature of the program. Most IRC programs seek to evaluate "apparent" effectiveness. We do not deem this level to be less valid than the other levels; rather, it is an honest representation of what we can hope to learn in the usual data-scarce environments in which we work.

- **Create a feasible M&E plan with your team and share it with donors early on.** In addition to program impact, IRC seeks to inform its decisions and learning with several layers of program data, depending on the context and questions we need to ask and answer. These include:

  - Implementation monitoring – based on project outputs, this data allows IRC and its partners to assess progress and/or challenges related to the timing and details of implementation activities.

  - Outcome or "effect" monitoring – based on what IRC terms the *effect* layer in its organizational logical framework, this provides meaningful information about whether or not we are seeing the changes we expect to see in the immediate behavior of communities and those specific actors (for example: community health workers, local government officials, TBAs, parents, teachers, refugee leaders) with which we work.

  - Process evaluation at critical intervals throughout the project's lifetime – depending on the duration of the project, an external consultant conducts a *process* evaluation mid-way through the project's lifetime and upon completion. Process evaluations provide an immediate resource for improving implementation.

  - Case studies, participatory evaluations and/or other approaches to document IRC's experience, engage beneficiaries and/or receive essential input to improve IRC's effectiveness.

IV. CONCLUSION

HELPFUL HINTS

Whether within a given project, overall sector or IRC in general, evaluation is an untapped resource for organizational learning. There are several reasons why this is the case and IRC is by no means unique in the challenges it faces: the short time frames, pressure to act not reflect and high expectations placed on NGOs to design, implement, manage and evaluate their work regardless of the context or resources are among a list of explanations.

Many of these are structural constraints; in other words, they are beyond IRC's control and therefore not something we can change in the immediate future. There are, however, many low hanging fruit that we can grab immediately to dramatically improve the way we learn about our work, its process, the way people perceive it, whether or not it does what we want it to and how different program designs are more or less effective at accomplishing our objectives. In summary, these guidelines identify some of these to which we can all turn our attention:

- **Guarantee purpose and question driven evaluation**: With every evaluation, those involved should be clear about the purpose of the evaluation, what we hope to learn and how the evaluation design will produce the related knowledge. Evaluation serves several different purposes: to improve implementation, monitor quality, to incorporate key stakeholders' views and reactions and, increasingly, to measure actual impact and effectiveness. All these dimensions are complementary, but should be carefully identified and addressed separately. We too often expect that evaluation can kill all these birds with one stone, while each one requires its own approach.
- **Focus on design and monitoring:** The point was made intentionally more than once throughout these guidelines – we will not learn from an evaluation of a poorly designed project or one that has not been monitored effectively. Evaluation is not magic, and design and monitoring are critical to whether we can ultimately learn about our work.
- **Ask, ask and ask again.** Language and words can be confusing and we should feel free to ask, learn and understand exactly what is discussed and what is being asked of us. "Impact" and "evaluation" are puzzling issues for most if not all aid agencies and donors. There are several articles and books written about the topic and significant interest building about how to improve the overall measurement of aid effectiveness. IRC staff should be confident in our attempt to join this dialogue, honest about what we know and do not know, and empowered to collaborate with donors and others to find feasible solutions.

DM&E AT IRC

IRC's DM&E Strategy is available on the Research, Evaluation and Learning Intranet site and upon request (email REL@theirc.org). It reflects the same principles and approach described in these evaluation guidelines. Central among these is that IRC will not be able to measure the results and effectiveness of our work if "DM&E" continue to be thought of as one activity that anyone can do with a few trainings. The processes, methods and principles that make data meaningful and useful for decision making cannot be mastered in a few days.

Evaluation is one of these key processes. It is by no means the only resource available for IRC to assess its work, as regular project monitoring can be a powerful source of information about how people respond to IRC programs and whether our "outputs" produce the changes we aspire to see. Evaluation, however, is worth improving. If it is done well, it can serve unique learning purposes. Depending on its design, evaluation can help IRC to answer important questions, encourage reflection and analysis, provide key insight into the consequences and processes of our work and, ultimately, measure the difference IRC makes in people's lives.

SELECT ARTICLES AND BOOKS

Banerjee, A. V. (2007). *Making Aid Work.* Cambridge: MIT Press.

ALNAP (2006). *Evaluating Humanitarian Action Using the OECD-DAC Criteria*. London: Overseas Development Institute. Website: http://www.odi.org.uk/alnap/publications/eha_dac/pdfs/eha_2006.pdf

Center for Global Development (2006). *When Will We Ever Learn? Improving Lives through Impact Evaluation.* Report of the Evaluation Gap Working Group. Website: http://www.cgdev.org/content/publications/detail/7973

Conference on Evaluation and Development Effectiveness in Washington, D.C. 15-16 July, 2003.

Cracknell, B. E. (2000). *Evaluating Development Aid*. California: Sage.

Duflo, E. & Kremer, M. (2005). *Use of Randomization in the Evaluation of Development Effectiveness*. Paper prepared for the World Bank Operations Evaluation Department (OED). Website: http://econ-www.mit.edu/faculty/download_pdf.php?id=759

Egon G. &. Lincoln, Y.S. (1989). *Fourth Generation Evaluation*. California: Sage Publications.

Herson, M & Mitchel, J. *Real-time Evaluation: where does its value lie?* Web site: http://www.odihpn.org/report.asp?ID=2772. Humanitarian Practice Network.

Hofmann, C.A, Roberts, L., Shoham, J. & Harvey, P. (2004). "Measuring the Impact of Humanitarian Aid," UK: Overseas Development Institute, Research Report #17. Website: http://www.odi.org.uk/hpg/papers/hpgreport17.pdf

USAID (1996). *Conducting A Participatory Evaluation*. Performance Monitoring and Evaluation Tips. Website: http://www.usaid.gov/pubs/usaid_eval/pdf_docs/pnabs539.pdf

GLOSSARY

Attribution – The ability to make a connection between the changes we observe or expect to observe and a specific intervention; the ability to credit the changes or the results achieved specifically to IRC.

Baseline – An adjective that describes information gathered before a program begins. There are two kinds of baseline data: those that are closely related to the proposed program, clearly indicated by the goals and objectives of the program; and those that provide useful background information about the context in which a specific IRC programs works.

Impact – The difference IRC makes; changes in outcomes experienced by people/communities for which IRC is responsible.

Program Evaluation – Defined herein as the systematic collection and analysis of information about the activities, characteristics and outcomes of programs

Project Cycle – A tool for understanding the sequence of tasks and management functions performed in the course of a project's lifetime. *The phases include analysis, design, implementation and monitoring and evaluation. They* are progressive, meant to make explicit how data collection and analysis should support decision making throughout a project's implementation.

Randomization (or random assignment) – The process by which the populations with which IRC works are selected randomly – think of pulling names out of a hat – to assure that there is not a selection bias that could otherwise explain any changes we observe in the results of the program. Those with which IRC works are known as treatment; those with which IRC does not work are the control.

Terms of Reference – The description IRC writes of an evaluation, intended to identify the project to be evaluated and the purpose for doing so. Terms of reference need to present a feasible task and be written concisely and clearly. Importantly, they need to leave open those issues – such as evaluation design and data collection methods – that will allow IRC staff to assess whether a specific consultant/evaluator will provide the knowledge IRC seeks and understands to be possible.

Valid data**–** Data that measure what they are supposed to be measuring. The concept of validity refers to the extent to which the data we collect give a true measurement or description of "social reality".

IRC RESARCH EVALUATION AND LEARNING
Jodi Nelson, Director

122 E. 42ND ST.12TH FLOOR NY, NY 10168
Phone: 212-551-2735
Email: jodi.nelson@theirc.org