Chapter 9
Evaluation

Orlando Hernández

This chapter provides guidance to EE&C project managers who work with evaluators. It will introduce you to some of the techniques and terms evaluators use, but most importantly it will show you how to design a project that can have meaningful evaluations, not only at the end of the project but throughout its life to keep it on course.

Most project managers make the mistake of not bringing in the evaluator until the end of a project and then not giving him/her a goal against which to evaluate performance. Asking an evaluator at this late stage, "What should we be evaluating?" is meaningless. The evaluator can only measure whether you have stayed on course—he/she cannot suggest destinations.

When involved in a project from the beginning, a good evaluator can regularly tell the manager whether the program is on course or, if not, in what direction it has strayed. With this information, the manager can decide how to get back on track (see Box 9.1).

A mantra for managers is: "Start with the results." If you don't have a precise vision from the outset of how things will look at the end of a successful project, you will have trouble with the evaluation.

Developing this vision is not easy. Indeed, it may be the most difficult part of management. The process should be participative, at least with a management team, sometimes with a wider group representing the target audience. It is usually a long, and sometimes exhausting, process at the end of which everyone commits to the vision and wants to be assessed in terms of it. Once agreed upon, the vision become the program's North Star.

Evaluation is usually categorized as *summative* evaluation, which measures the project success or failure by comparing outcomes with the original goals, or as *formative* evaluation, which measures project progress against ongoing benchmarks and allows the manager to make course corrections.

Formative evaluation is more useful to a program manager, because it provides information that helps the program succeed. Summative evaluation, coming after the program is over, gives a verdict about whether the program achieved its goals, but is of no help to the manager in achieving those goals. (Of course, the results of summa-

BOX 9.1
Keeping the Desired Results in Mind

In Nicaragua, sea-turtle experts were convinced that if local residents just understand the rapid sea-turtle population decline, they would be less likely to harvest eggs. A storybook was written and approved by the biologists that did an excellent job of explaining all the potential disasters that stalk the young turtles until they reach maturity: egg predation by herons, crabs, and coyotes on the shore; sharks; shrimp nets, and even turtle hunters. After giving the story to readers, the program manager developed an evaluation survey that asked about their attitudes regarding egg collection and abiding by the quota system. Only then was it clear that the storybook was not about egg collection and did not even mention the quota system. Fortunately, there was still time to rewrite the story.

A mantra for managers is: "Start with the results."

tive evaluation can be useful for people designing new projects.)

DEVELOPING DESIRED RESULTS

The statement of the project's vision—or more specifically its *desired results*—guide the evaluation process, just as they have driven the program development. By operationalizing desired results into measurable statements, the evaluator can reflect upon the degree to which the program achieves these results. Well-stated desired results for educational programs are specific to the situation and share these elements:

1. Each objective targets one and only one thing: a fact, an attitude, a skill. Limit the statement to only one measurement.
2. Each objective specifies an outcome that the participant will be able to perform. The objective is not written from the perspective of the leader (teach about turtles) or the program coordinator (host the workshop). Use appropriate action verbs to define the outcome.
3. Each objective spells out what will be measured in order to meet certain criteria (80% success, three out of five reasons).
4. Each objective is set in a context or a condition (when asked, when given a list of 10 items, where ascertained, which population...).

OBTAINING BASELINE MEASURES

Since the evaluation is designed to measure change, some technique to measure the "baseline" situation is necessary. The following activities may provide this initial information.

- Use the literature or existing data in the agency
- Survey people
- Observe people
- Interview people
- Use information from a comparable site or a former program
- If you didn't do a baseline study, at least ask people at the end of the study how they think they've changed

TOOLS FOR COLLECTING INFORMATION

Each information-collecting tool has a niche in evaluations, and just like an organism in a functional ecosystem, each is best suited to a particular condition. The program manager must match the tool to the need. A variety of equally good evaluation designs can use different tools. As a rule of thumb, choose the tool that is least expensive in time and resources. There are many ways to maximize the advantages and minimize the disadvantages of each option (see Table 9.1).

What is a Research Design?

To evaluate is to compare. Comparisons are needed to determine if an intervention had the desired impact. A research design tells the researcher how many measurements should be done to determine impact, and when those measurements should take

Table 9.1 Information Needs and Evaluation Tools

Data Collection	records, logs, journals, clicker counts
Program Quality	expert review, observation, staff self-analysis, staff performance
Participant Reaction	drawings, photographs, journals, logs, post-it boards, suggestion boxes, comment cards, testimonials, anecdotes, observation
Participant Knowledge and Behavior	surveys, interviews, concept maps, observation, artifacts, photographs, focus groups
Action Research	journals, tape-recorded sessions, observation, etc. to support participant reflection and analysis
Media Impact	phone or mail surveys, count calls, visits
Materials Quality	readability tests, pre-tests, observation
Participant Involvement	participatory rapid appraisal techniques such as discussion groups, engineering models, mapping, sorting photographs, calendars, time lines, trend lines, ranking, pie chart, matrix

place. The various comparisons needed to determine net effects of an intervention make up a research design. Designs also dictate whether or not comparisons will be limited to study groups exposed to the intervention or if they will also include groups not exposed to it (control groups).

Three Commonly Used Research Designs

GreenCOM has used three well-known research designs (listed below) to evaluate how education and communication programs changed the audiences' knowledge, attitude, skills, or beliefs. Each one of these designs has different advantages and disadvantages regarding the sources of error. Each can be used in formative or summative evaluation.

Design 1: Pre-Test and Post-Test (Before and After) Studies

This design compares the same type of study participants at two points in time, separated by a period of participation in a program. Differences in scores between one point in time to the other are taken as an estimate of the net effects of an intervention (Rossi and Freeman, 1988).

There are two versions of this design. One version known as the "one-group pre-test post-test design," uses the same group for both measurements. The other version, known as a "separate sample pre-test post-test design," tests people from different groups at each measurement point.

The one-group version is commonly used in education and in communication. It can be used when an intervention affects a specific target group. Despite its popularity, this design embodies several confounding factors that can jeopardize the validity of its results. For example, it does not clearly establish that the intervention caused the measured change in the population. Other variables may have caused any difference detected between the two measurement points. As Rossi and Freeman (1988) have concluded, "the main deficiency of such designs is that they ordinarily do not permit disentangling the effects of extraneous factors from the net effects of the intervention." See "Cautions to the Evaluator," below.

The "separate sample design" offers some improvement over "one-group pre-test post-test designs." If study participants are randomly selected for each measurement, the effect of testing is controlled for. *Maturation issues* (see below) are controlled if the distribution of age is the same in both samples. However, in a separate group pre-test/post-test design there is still a questions as to whether external events that affected all participants might have had an influence.

Design 2: Pre/Post-Test with Experimental and Control Groups

This design is similar to the pre-test/post-test design, but a control group has been added. Thus, the experimental and control groups are both measured before and after the intervention. If an external event influences all participants, it will show up in results from the control group as well as from the experimental group. As before, this design has two forms, one where study participants have been randomly assigned to the study group (the pre-test, post-test control group design) and one where they have not, (the non-equivalent control group design.) In both of these designs it is imperative that the same study participants take the pre-test and the post-test (Fisher, Laing, Stoeckel and Townsend, 1995).

Design 3: Post-Test Only Control Group Design

Post-test only designs are appropriate when baseline data have not been collected, are lost, or are inaccurate, or when the introduction of new subject areas makes pre-testing impossible. This design requires that two study groups be researched after an intervention has ended. The experimental group is exposed to the EE&C intervention and the control group is not.

There are two types of post-test only designs that differ in how study participants are chosen. When there is no random assignment of participants to each study group, the design is called a "static-group" comparison. When participants are randomly assigned to the study group, the design is called the "post-test only control group."

Campbell and Stanley (1966) argue that under the static-group comparison there must be a method of assuring that the two groups would be equivalent had it not been for the treatment. The randomization element added to the post-test only control group design corrects that deficiency. Campbell and Stanley (1966) also argue that randomization can suffice without the pretest in the case of the post-test only control group design.

Rossi and Freeman (1988) define randomization as the chance assignment of potential targets in order to obtain equivalent treated and comparison groups. Randomization requires that every unit in a target population has the same chance to be selected for either the experimental or the control group. An important aspect of randomization is the elimination of the possibility of self-selection. Randomization is different from random sampling. Random sampling allows the selection of units in an unbiased manner to form a sample from a population. Random sampling can be used to choose individuals to participate in a study. Randomization is used to assign each member of the resulting sample to the experimental or control group.

Cautions to the Evaluator: Common Sources of Error

Research designs are chosen based on the sources of error that must be avoided. Common sources of error are listed below.

Contextual events

Contextual events are between two measurements taken to evaluate an intervention that may have influenced the knowledge, attitudes, beliefs, intentions, and behaviors targeted by the EE&C intervention. The changes that may be observed between the two measurements may be due to these events not to the intervention.

Maturation of Study Participants

Study participants may change over time and those changes may influence the results. If there is a time difference between measurements, study participants may have gotten tired or hungry, or, if there is a long time between measurements, gotten older and more mature.

Loss of Study Participants

All participants involved in the beginning of the study may not be available at the end because of migration, loss of interest, or even death. The key question is: are the remaining subjects in subsequent measurements representing either the best, the worst, or the average study participants of the first sample? (See Box 9.2.)

Repeated Testing

The more individuals are exposed to the same questions, the better they may become at answering correctly. When an evaluation instrument is applied before and after an intervention, the first evaluation has an impact on the second one. Responses obtained during the second measurement may be better than those obtained during the first measurement, simply because of the testing effect. Repeated exposure of study participants to study instruments may invalidate research findings. Campbell and Stanley (1966) report that on achievement and intelligence tests, "students tak-

BOX 9.2

Evaluating over Time

A three-year study was conducted to evaluate a program promoting soil conservation practices. Measurements were done at each cropping season to see if study participants were using soil conservation practices such as minimum tillage and contour plowing. At each measurement point, 10–15 percent of the study participants were lost. It was difficult to determine if the participants that were lost were the best or the worst soil conservation farmers. Consequently, it was impossible to determine what impact their loss had on research findings and changes observed over time.

Evaluation is difficult because it involves a great deal of thinking, planning, and imagining the future.

ing the test for a second time, or taking an alternate form of the test, usually do better than those taking the test for the first time. These effects, occur without any instruction as to scores or items missed on the first test."

Modifications of Evaluation Instruments and Increased Experience of Evaluators

Evaluation instruments may be refined or modified between measurements either by accident or intention. From one measurement to the next, an original question such as "Can you mention the days when waste is collected in this neighborhood?" can be changed to "Are you aware when waste is collected in this neighborhood?" The changes observed between measurements may be due to the way in which the question was asked each time and not the result of an awareness-related intervention. The experience of evaluators, interviewers and observers can also have a great impact on results. Observers may differ in their accuracy and severity. Both factors can affect results and invalidate findings.

CONCLUSION

Evaluation is difficult because it involves a great deal of thinking, planning, and imagining the future. At the beginning stages of program design, it is often challenging to identify measures of success for each activity. Each of these measures could become a desired result that will guide the development of the program and determine how the program is evaluated.

The broadest definition of the evaluation process begins with program planning. As needs are assessed and formative research conducted to determine initial knowledge, attitudes, and behaviors, a type of evaluation is in progress. Baseline data, collected before the intervention, will help measure changes that can be attributed to the project.

As the project evolves, pretesting is critical for keeping activities on track, by testing elements, making revisions, trying new techniques, and reorganizing activities to best meet the desired results. Observations and interviews help record information about the experiences of the participants.

At the conclusion of the project, a summative evaluation can measure its merit.

Remember, start planning by imagining the results you want.

References

Campbell and Stanley. (1966.) *Experimental and Quasi-Experimental Design for Research*. Boston: Houghton Mifflin Company.

Fisher, A.A., J.E. Laing, J.E. Stoeckel, and J.W. Townsend. (1995.) *Manual para el Diseño de Investigación Operativa en Planificación Familiar*. Segunda Edición. New York: The Population Council.

Rossi, P. And H. Freeman. (1988.) *Evaluation: A Systematic Approach*. Fourth Edition. Beverly Hills, CA: Sage Publications.