PSR 27(1979):5-10

OF BLACKBIRDS AND BOXES: AN INTRODUCTION TO EVALUATIVE RESEARCH

FRANK LYNCH, S.J.*

My aim is to describe and illustrate just three features of evaluative research, namely, the basic concept, the most common foci, and the favorite designs. For if at the outset of this seminar we review what is meant by evaluative research, what problems it most frequently addresses, and in what ways it is best conducted, we should be well prepared to profit from the presentations that follow.

The Meaning of Evaluative Research

Everyone of us is at heart an evaluator. Indeed, to be truly human and alive one must be critical; scratch even a saint, and a critic bleeds. For the making of discriminative judgments, which comes so easily to all of us, is of itself a good thing. The problem is this, that in the hustle-bustle of ordinary life we now and then evaluate without reflection; such is our compulsion in this regard. In these instances, unfortunately, we feel no obligation to specify the criteria that underlie our offhand pronouncements, or to offer acceptable proof that the person or thing on trial has indeed failed or passed the test of our standards. Life is too fast, perhaps, or what we owe one another too easily forgotten.

Precisely because any critical judgment, however loose, naive, vicious, or irresponsible, can be called an evaluation, some authors are at pains to use a special phrase, "evaluative research," for evaluations done with scientific discipline, system, and accountability. The label "evaluative research," which we have adopted, is consciously restrictive: it emphasizes the research aspect of the process, and informs us at once that we are dealing with a scientific operation of one particular kind -- concretely, a species of explanatory research.

Most commonly, evaluative research is now employed to weigh the merits of some organized intervention, generally a social action program undertaken to produce certain beneficial effects in the target population. Examples that come easily to mind in the Philippine context are programs for the improvement of nutrition, housing, family planning and welfare, rural electrification, out-of-school youth, manpower training, adult literacy, or agrarian reform. Applied to any such organized activity, evaluative research may be defined as *a process to determine how successful (or unsuccessful) a program has been in achieving its predetermined objective or objectives.* Operationally, the process will include at least the following five steps:

1. Identifying and operationalizing (devising measurable indicators of) those objectives whose attainment will be evaluated – the so-called *program effects,* or *outcomes;*
2. Similarly identifying and operationalizing the program *inputs;*
3. Devising and implementing a data-gathering plan;
4. Analyzing and interpreting the findings; and
5. Making summative or formative suggestions regarding the program which has been studied.

*Resident Consultant, Institute of Philippine Culture, Ateneo de Manila University and other institutions until his untimely death in 1978.

Consider a simple example. Imagine that we have been invited to help plan a modest literacy program to be conducted in a single municipality. While discussing matters with the originators of the idea, we come to several agreements. The first is that the major program objective will be functional literacy for those out-of-school youths or adults who live in the municipality and ask to be enrolled in the proposed two-month training course. Second, the attainment of this goal is to be measured by a straightforward test of the ability to read and write a simple message in any language (though the languages of instruction will be Tagalog and English). The principal inputs, on the other hand, will be the efforts of the literacy instructors (administrative/program variables) and the attendance and performance of the enrollees (participant variables).

Having defined both the expected outcome and the programmed inputs, and provided as well for their measurement, we now design a data-gathering scheme. Since in this instance we are dealing with an enlightened program manager (any manager who calls you in during the program-planning stage *must* be enlightened), he or she will certainly be open to our suggestions for the selection of students for the literacy course. Although eventually all eligible candidates will be trained, we shall propose that they be admitted to the program on a batch-by-batch basis. This arrangement will allow us to employ a very creditable research design.

Concretely, then, the target population will be all those out-of-school youths or adults who reside in the municipality and apply for admission to the literacy program. The experimental or test group will be those chosen at random to be admitted to the first round of the program. The controls will be those who are waiting their turn. Input data will be gathered by simple records made of the attendance and performance of both instructors and participants during the two months the course is in session. Output data will consist of literacy-level measurements made both before and after the course on members of the experimental and control groups alike.

The analysis and interpretation of data could proceed in a number of ways. The basic questions we might answer, however, are these:

1. What percentage of enrollees achieved functional literacy? Did they do so more often than the controls did?
2. What about gain scores – did the enrollees show greater average *improvement* (posttest scores minus pretest scores) than the controls?
3. Did instructors and students perform as they were supposed to?
4. Was improvement in literacy correlated with instructor and/or participant inputs?

On grounds of the replies given to these four questions, we might recommend that the program be continued or terminated (a summative evaluation), or be modified in one way or another to make it more effective (the formative mode).

The Most Common Foci

The example just given illustrates the fact that at least two aspects of a social program may be evaluated: inputs and outputs, or – to use another set of terms – efforts and effects. The third and last fundamental focus for studies of this kind is generally labeled *process*, a shorthand term for whatever makes sense of, or explains, the earlier evaluative findings. The study of process is, as it were, a rooting around in the "black box" that houses those secret workings whence spring the results we call effects (or non-effects).

But effects, or outcomes (the second basic focus) are most often seen, not in isolation, but in reference to some total goal to be achieved. And again, however impressive the program's output, the sobering question will commonly arise: could the same results have been attained at less expense or trouble, or in a shorter period of time? With these added criteria of success, then, we have five foci in all, to one or more of which attention is regularly paid in the course of evaluative research. I shall review them, adding Suchman's (1967) memorable analogy of the bird in flight; you may forget the foci, but I guarantee that you will remember the bird. The five concerns of

evaluative research are these.

1. *Effort.* This is an assessment of the quantity and quality of program-related activity, regardless of output. How many hours were spent in the classroom, how many homes visited, how many outpatients treated, how many crop loans extended? This is an evaluation of *inputs.*

Evaluation at this level is compared to measuring the number of times a bird flaps his wings, with no attempt to determine how far the bird has flown (Suchman,1967:61).

2. *Performance.* Here we measure the *results* of effort, rather than the effort alone. What did the students learn, what changes occurred in homes visited by the social worker or nurse, how many outpatients were relieved of the symptoms that brought them to the clinic, how many crop loans were used for the intended purpose – and to what effect? This is an evaluation of effects, or *outcomes.*

Here the focus is on how far the bird has flown, and not merely on the frequency of his wing flapping.

3. *Adequacy.* The question is now no longer how far have we come but how much further must we go. The outcome is seen in reference to the total envisioned goal. What percentage of the syllabus have the students mastered, what further changes must be effected in the homes of the client population, what percentage of outpatients treated remains uncured, and what proportion of crop loans were improperly used? The measure of adequacy tells us how effective a program has been in terms of the denominator of total need.

How far, in other words, has the bird flown with reference to where he has to go?

4. *Efficiency.* Compared to the program under scrutiny, we ask, is there a better way to achieve the same results? "Efficiency is concerned with the evaluation of alternative patterns or methods in terms of costs – in money, time, personnel, and public convenience. In a sense, it represents a ratio between effort and performance – output divided by input." (Suchman, 1967:64).

Could the bird have gotten where he was going with less flapping of his wings? Did he fly too high, or not high enough? Did he take sufficient advantage of favoring air currents? Indeed, could he have travelled more efficiently still, by sitting on a freight car roof and letting the train do his work?

5. *Process.* This is the most difficult of all questions to answer: how and why a program succeeds or fails, or works well under some conditions but not others.

Strictly speaking, questions of this kind need not be part of a program evaluation. However, properly answered, they may result in the saving of a program which might otherwise have been closed down because of its ratings on the other four criteria.

The analogy of the bird is a bit strained here, unless we wish to speak of meteorologists teaming up with animal physiologists and psychologists to explain the flight characteristics, migration patterns, and homing behavior of our little feathered friend. Consequently, we might return instead to the black-box metaphor mentioned earlier, and think of the study of process as an attempt to discover what makes the program tick – or run down.

These then are the five specific problems to which evaluative research commonly addresses itself: effort, performance, adequacy, efficiency, and process.[1]

Preferred Research Designs

Since the appearance of Campbell and Stanley's classic catalogue of research designs (1963, 1966), the three-way division of study plans which they employ has become increasingly popular. Further, despite frequent assurances that the second and third categories of design – that is, the quasi-experimental and non-experimental – are acceptable in certain situations, the experimental model remains the undisputed first choice, even for evaluation research. This is so largely because the principal feature of this kind of design is the *random assignment* of study units to experimental and control groups. Properly done, this procedure solves one of the greatest threats to the study's internal validity: preexistent differences bet-

ween those who were subsequently exposed to the experimental treatment and those who were not.[2]

Although there are three relatively common experimental designs, two of them are merely variations on the basic plan found in the first. This elementary arrangement is called the *pretest-posttest control-group design,* and is represented as follows in the notation used by Campbell and Stanley (1966).

$$R\,O_1\ X\,O_2$$

$$R\,O_3\quad O_4$$

By this plan, units are assigned at random (R) to the experimental (upper line) or control (lower line) group. Pretest observations are made of the members of each group (O_1 and (O_3, then the experimental treatment (X) is applied to one group; posttest measurements follow ($O_2 - O_4$). The change that occurred in the experimental group ($O_2 - O_1$) is then compared with that experienced by the control group ($O_4 - O_3$), and conclusions drawn regarding what effects might be traceable to the experimental treatment. This is the basic experimental design.[3]

Especially where the number of study units is large, random assignment to treatments makes the use of pretests superfluous. Hence a simpler form of the basic model may be used. Here the first set of observations is eliminated, and one has instead the *posttest-only control-group design,* as follows.

$$R\ X\ O_1$$

$$R\quad O_2$$

Now if we quite literally combine the first and second plans into a single strategy involving two experimental and two control groups, we have the *Solomon four-group design.* Its attractiveness consists in this, that the minor difficulties that would arise from using either design alone are corrected by using the two of them together. In the notation below, the pretest-posttest control-group design will be recognized in the first two lines, the posttest-only plan in the lines that follow:

$$R\,O_1\ X\,O_2$$
$$R\,O_3\quad O_4$$
$$R\quad X\ O_5$$
$$R\quad\quad O_6$$

Less desirable, but often either appropriate or necessary, are the so-called quasi-experimental designs. The label derives from the fact that we here apply an experimental approach to bodies of data which are not susceptible to full experimental control. Most important, the assignment of treatments cannot be made, or in any event is not made, in random fashion. Compensating somewhat for the absence of this paramount feature are two provisions, multiple observations and comparison groups, which appear singly or in combination in the three most common quasi-experimental designs. In descending order of acceptability, I would rank the plans as follows: first, the *multiple time-series design* (with a nonequivalent control[4]); second, the simple *nonequivalent control-group design* (without a time series); and third, the *time-series experiment* without a comparison group (for the Campbell and Stanley notations, see Figure 1, frames 1-3).

Weakest of all, but acceptable nonetheless for the preliminary "soft" assessment of a program's worth, are the so-called nonexperimental (or pre-experimental) designs. These plans possess the virtue of simplicity, and can detect the presence of at least gross before-after differences. What they cannot discover is the program's role in the changes we observe. They can tell us, in other words, whether or not some effect is being produced, but they cannot eliminate the possibility of non-program explanations for the pretest-posttest differences.[5] Indeed, in two of the three most common designs, there is no pretest against which to measure the posttest results. Consequently, if I were asked to state my first preference among these three plans, I would name the *one-group*

pretest-posttest design; after that I would place the *static-group comparison,* followed by the *one-shot case study* (Figure 1, frames 4-6).

Figure 1. Quasi-experimental (1-3) and non-experimental (4-6) research designs in the notation of Campbell and Stanley (1966).

$\underline{O_1\ O_2\ X\ O_3\ O_4}$ $O_5\ O_6\ \ O_7\ O_8$ (1) Multiple time-series design	$\underline{O_1\ X\ O_2}$ $O_3\ \ O_4$ (2) Nonequivalent control-group design
$O_1O_2O_3O_4\ X\ O_5O_6O_7O_8$ (3) Time-series experiment	$O_1\ X\ O_2$ (4) One-group pretest-posttest design
$\underline{X\ O_1}$ O_2 (5) Static-group comparison	$X\ O$ (6) One-shot case study

Practical Suggestions

If one or more of you approached me now and asked what concretely you might do to increase the likelihood that the evaluations you did would be worth the trouble you took to do them, I would answer with these five pieces of advice.

1. When the opportunity presents itself, reinforce the growing consensus that a social program with no provision for its own evaluation is at best deficient, and at worst, irresponsible.
2. When the occasion permits, make the point that a program evaluation should be planned when the program itself is planned; it should not be tacked on as an afterthought.
3. In planning an evaluation with the program manager, or whoever is responsible for its conduct, strive mightily for an arrangement in which an experimental design can be used; objections to this kind of plan are more easily met than many people realize.
4. If an experimental approach is impossible, then push at least for a design in which pretest data are gathered, preferably over time, from both a program and comparison group.
5. But be flexible, open to the advantages of a two-stage approach featuring a soft evaluation followed, where indicated, by a more rigorous appraisal.

The strategy outlined in these five suggestions is clearly realistic: we do what we can. Equally down to earth will be your acceptance of the fact that the results of your evaluative research will not always endear you to those for whom you did it. Indeed, the more skilled you are, the more likely you are to incur the displeasure of program managers and sponsors. To paraphrase Peter Rossi (1972: 38, n. 32): "No good evaluation goes unpunished."

Notes

[1]Among the many aspects of process which might be investigated is what is currently referred to as social soundness, a phrase referring principally to the qualities of social (distributive) justice and cultural compatibility. A discussion of the concept and of means to assure, as much as one can, its presence in a social program is found in Lynch, Illo, and Barrameda (1976).

[2]Aside from the problem of differential selection, seven other threats to internal validity are identified: history, maturation, instrumentation, statistical regression, experimental mortality, and selection-maturation interaction (Campbell and Stanley, 1966). The same authors also describe four threats to a study's external validity, or generalizability: interaction effect of testing, interaction effects of selection biases and the experimental variable, reactive effects of the experimental arrangements, and multiple-treatment interference.

[3]With one added feature, this is the research plan which appears in the literacy-program example presented in the first section of this paper. The basic design was slightly modified by provision for the "staging of experimental treatments" (Weiss, 1972: 63-65) through the use of a so-called holdout sample (Boruch, 1976: 52-56).

[4]A "nonequivalent control group" is not, strictly speaking, a control group. It is rather a *comparison* group. However, I am retaining the design labels of Campbell and Stanley (1966).

[5]In view of these considerations, Rossi suggests (1972: 48) "a strategy of evaluation in which soft methods are used to eliminate ineffective projects and to detect potentially effective ones. Those found to be potentially effective then need further and more precise evaluation through controlled experiments or close approximations to such designs."

References

Boruch, Robert F.
1976 Coupling randomized experiments and approximations to experiments in social program evaluation. *In* Validity issues in evaluative research, I.N. Bernstein, ed. Beverly Hills, Sage Publications. Pp. 35-57.

Campbell, Donald T., and Julian C. Stanley
1963 Experimental and quasi-experimental designs for research on teaching. *In* Handbook of research on teaching, N.L. Gage, ed. Chicago, Rand McNally.
1966 Experimental and quasi-experimental designs for research. Chicago, Rand McNally.

Lynch, Frank, J.F.I. Illo, and J.V. Barrameda, Jr.
1976 Let my people lead: rationale and outline of a people-centered assistance program for the Bicol River Basin. Social-soundness submitted to the U.S. Agency for International Development. Quezon City, Social Survey Research Unit, Institute of Philippine Culture, Ateneo de Manila University.

Rossi, Peter H.
1972 Testing for success and failure in social action. *In* Evaluating social programs: theory, practice, and politics, P.H. Rossi and W. Williams, eds. New York, Seminar Press. Pp. 3-49.

Suchman, Edward A.
1976 Evaluative research: principles and practice in public service and social action programs. New York, Russell Sage Foundation.

Weiss, Carol H.
1972 Evaluation research: methods of assessing program effectiveness. Englewood Cliffs, New Jersey, Prentice-Hall, Inc.