Spring 1980

Evaluation Designs

Clare Rose

EVALUATION

Evaluation Designs

Clare Rose

The evaluation models described in previous issues of the *Quarterly* represent some of the major paradigms of educational program evaluation; they have been used to guide many evaluations and they have influenced the thinking of many practicing evaluators. Models provide a broad base for designing evaluation activities by offering a framework and conceptualization that guides both the focus of the evaluator and the orientation of the evaluation. But models do not provide strategies for implementation. Guidelines are provided by the design, which establishes the conditions and procedures for collecting the data required to answer the questions of concern. The design must be related to the type of program or service being evaluated; that is, the selection of a particular design is guided by the decisions that will have to be made as a consequence of the data. In turn, the adequacy of a particular design can be determined by the extent to which the results may be interpreted and the questions answered. In most cases, evaluation designs have been borrowed from research.

For example, Campbell and Stanley (1963) distinguish between three types of research designs commonly used in evaluation—pre-experimental, experimental, and quasi-experimental. The criterion differentiating the three groups of designs, as well as the quality of the designs within each group, is the extent to which the design protects against the effects of extraneous or nonprogram variables, thus legitimizing the results that are attributable to the program. More specifically, the criterion is the extent to which the design protects against eight threats to internal validity[1]—eight kinds of variables,

[1] Campbell and Stanley also describe threats to external validity that jeopardize the generalizability of the findings. Although some writers argue that generalizability is (or should be) an important consideration in program evaluation, most others feel that generalizability is not a major concern in most educational program evaluations.

POD Quarterly, Vol. 2, No. 1 (Spring 1980)

extraneous to the program, that if not controlled, will affect the outcomes of the program and thus the accuracy of the interpretations that can be made of the data.

The eight threats to internal validity are *history* (changes within the program and external events), *maturation* (of the program or target population), *testing* (effect of a pretest on subsequent tests), *instrumentation* (changes in instruments, observers or scorers), *selection biases, statistical regression* (non-program effects which can appear during statistical manipulations), and *selection-maturation interaction* as a result of selection bias.

True experimental designs protect against all of these possible threats to internal validity;quasi-experimental designs generally protect against most of them. Quasi-experimental designs require the same rigor, but they are more practical than the true experimental model in many real-world situations. Pre-experimental designs totally lack control and, according to Campbell and Stanley, are "of almost no scientific value." Examples of pre-experimental designs are: 1) the one-group, pretest-posttest design in which a single group is pretested, exposed to a program, and then posttested; depending upon the length of time between the pretest and posttest, the design is open to the threats of history or maturation; 2) the static-group comparison, in which a group that has received a program or service is compared with a group that has not—a comparison that is suspect since the original equivalence of the two groups is unknown; and 3) the one-shot case study in which a single group is studied once. More will be said about the limitations of case studies in a subsequent column.

Quasi-experimental Designs

Because of the difficulty of conducting true experiments in the real world of education, quasi-experimental designs have become more widely used in both research and evaluation projects in recent years, particularly as these designs gained respect under Campbell and Stanley's sponsorship. The designs described on the following pages are the more widely known of the quasi-experimental group, and each claims certain special features that make it appropriate in different types of evaluation settings.

The Nonequivalent Control Group Design. Probably the most commonly used design (and also the least satisfactory) is the nonequivalent control group design, in which control and experimental groups are formed without benefit of random assignment. A comparison group of available individuals or intact groups whose characteristics are similar to the experimental group are used as controls. Pretest and posttest measures are taken for both groups and the results are compared. Although obviously not as rigorous a design as a true experiment, in which comparison groups are based on random assignment, the main issue in the nonequivalent control group design is one of selection—identifying the variables that were used to place the participants in each group. The objective, of course, is to make the two groups as similar as possible. The more similar the control group is to the experimental group, the more reliable the interpretations that can be made of the data. Weiss (1972) proposes using "unawares" (people who did not hear of the program but might have joined if they had) and "geographical ineligibles" (people with characteristics similar to the experimental group who live in locations that have no similar program) as control group samples.

The Time Series Design. The time series design involves studying the behavior of an individual or a group over time. Although the statistical procedures for analyzing the data are sometimes complex, the time series design has many advantages to offer. A series of measurements are taken of the participants before, during, and after the onset of a program, with the before measures establishing a baseline performance level against which to measure changes. The measures are examined to determine an "effect pattern" or trend to show the impact of the program over time.

The *multiple* time series design provides more rigor by adding an additional group and examining the series of measurements for both groups. If the program evaluated has been effective, the effect pattern for the two groups should be markedly different. A major advantage of the time series design is that it is a fairly powerful design, providing excellent information on the effects of programs even when a comparison or control group cannot be used. Time series designs are particularly well suited for longitudinal evaluations and social action evaluations (a category into which professional development programs might likely fall) where the program cannot be withheld from appropriate participants.

Experimental Designs

Although some writers acknowledge the difficulty of applying controlled experiments to the problems of education, and more than a few add the caveat of "where conditions allow," experimental design is to many educators the cornerstone of evaluation—the ideal methodology for educational program evaluation.[2] Campbell and Stanley (1963) state unequivocally that they are

> . . . committed to the experiment: as the only means for settling disputes regarding educational practices, as the only way of verifying educational improvements, and as the only way of establishing a cumulative tradition in which improvements can be introduced without the danger of a faddish discard of old wisdom in favor of inferior novelties.

Classic experimentation design incorporates two important techniques that together rule out the possibility that something other than the program caused the observed results, and thus, they confirm the legitimacy of the interpretations made from the data. These techniques are the use of control or comparison groups and randomization. Quite simply, this means that samples of the target population are randomly selected and assigned to either the experimental group receiving the treatment (program) or the control group, which receives a different treatment or no treatment. Members of the two groups are posttested after the program has been completed, the differences are compared, and the experimental program is pronounced a success if the experimental group has more of whatever the criterion variable is than the members of the control group. That the experimental group had fewer cavities after using Crest should by now be a familiar slogan. Had a true experimental design not been used, Crest marketing strategies would likely have taken a different turn and Arthur O'Connell would not have had the opportunity for a mid-life career change to a fictional drug store owner.

Without question, experimental design can be a powerful tool. If people can be randomly assigned and if there are enough of them

[2] See Aronson and Sherwood, 1972, Campbell, 1972; Glennan, 1972; Houston, 1972; Popham, 1975; Porter, 1973; Rossi, 1972; Scriven, 1967, 1972, 1974; Stanley, 1969, 1972; Welch and Walberg, 1972; Weiss (1972a, 1972b); and Wholey *et al.*, 1972. Evans (1974) makes a compelling argument in favor of small-scale controlled experiments to test the relative effectiveness of alternative program techniques as a precurser to the introduction of massive national programs.

available to form an experimental and a control group; if the control group will not be harmed or deprived psychologically, socially, or financially by not receiving the program or by receiving a placebo program; if the program is a specific, definable entity; and if the objectives are explicit, then an experimental design is probably the best choice. If the evaluation proceeds smoothly and if the instruments and measures are valid and reliable and appropriate to the objectives, then, if the experimental group shows greater positive change than the controls, we can be fairly certain that the change is due to the effect of the program.

But programs do not exist in apolitical or ideal contexts and compromises in design are inevitable. There are innumerable occasions when forming control groups and randomization are difficult; there are many situations in which it is impossible. Sometimes programs have to be offered to intact groups; sometimes groups available for comparison are too dissimilar. And for every factor on which groups are matched, there are other equally, if not more important, variables on which they are unmatched. It is these variables that may in fact exert more influence on the outcomes than the variables on which the groups are supposedly matched.

In other situations, programs must be provided on a voluntary basis and made available to all. Few administrators, or program evaluators for that matter, would be willing to deprive people of programs that could be of benefit to them. It is difficult both to refuse service to those who seek it and to force it upon those who don't want it, as all professional developers know too well.

The many limitations of experimental design, particularly those which focus on the extent to which a program has achieved its objectives, are well documented and will not be reiterated here. For more detailed discussions, the reader is referred to Borich and Drezek (1974); Guba (1969); Riecken (1972); Rose and Nyre (1977); Stake (1975); and Wergin (1976).

Most studies carried out under experimental conditions fail to assess the impact of the program operating within functioning institutional or organizational systems. Their generally singular focus on objectives limits the evaluator's understanding of the program and, despite Scriven's exhortations, attention is seldom paid to the merit of the goals established for the program or to unanticipated outcomes that may have far more important consequences than the goals originally intended. Experimental designs do not take into account changes in goals (or procedures) that frequently take place once a program is underway, and they cannot provide the immediate formative feedback that programs often need in order to identify and correct snags in their early stages of implementation.

Most experimental designs that have been used in educational evaluation fail to consider the manner in which the program was implemented or the configuration of people, events, processes and practices, values and attitudes that surround the program, affecting the environment in which it operates and thus, at least presumably, its outcomes. It is not enough to document that a program failed to work. It is essential to identify the processes and other variables that combined to defeat it. Particularly in the case of large social action programs, but even with small-scale educational programs, the investigation of negative effects is an important issue. The capacity of communities, organizations, institutions (and people!) to resist change must be investigated and the factors that defeated a program identified so that they can be used as a base for the design of a program that is more likely to be effective.

Conversely, it is not enough to document that a program achieved its goals and the extent that it did so. Equally important as the attainment of goals is the concern with *why* the results occurred, what processes intervened between input and outcome, how the program actually operated, what nonprogram events may have affected participation, and what implications and guidelines can be derived from the evaluation for program improvement and replication. Experimental design alone cannot provide this essential information.

In broad-aim programs, such as consortial faculty development efforts and multi-project funding activities by agencies and foundations, different approaches are often used at the local level so that the programs in effect differ markedly from campus to campus. A description of the different forms and approaches as well as the forces that shaped each would be important information that cannot be obtained through traditional experimental evaluation.

Stufflebeam (1971) contends that experimental designs are only appropriate in product evaluations, and thus are of minor relevance to educational evaluation. Guba (1972) goes further, stating that experimental design actually "*prevents* rather than promotes changes" because the programs cannot be altered if the data and interpretations about the differences between them are to be unequivocal.

The same criticisms and shortcomings can be leveled against quasi-experimental designs in which the usual thrust of the study is also the degree to which desired goals have been attained. No matter how effective and useful they are in some situations, again, little attention is paid to how the program developed, what unanticipated consequences occurred, what variations exist among the program's component parts or units, what outside events affected either programming or participants, or to the adequacy of the program operation and the capability of the staff. As Stake (1972) suggests, most classical designs were developed as a means of examining "minute details"; they were not developed for portraying the "whole cloth of the program." The point is, evaluation designs must accommodate the characteristics and informational needs of the program, not the other way around.

Summary

So where does the above discussion of evaluation designs leave professional developers and evaluators of professional development programs and activities? With the recognition that most, if not all of our programs cannot be subjected to the most rigid, and thus most valid and reliable, evaluation designs. This should not be used as an excuse for not attempting to approximate the rigors of the formal models and designs, however, as some acceptable alternatives are available. They all require certain compromises, and some are naturally more suitable than others depending on the type of program being examined and the outcomes desired of the evaluation being conducted. Holistic evaluation, transactional evaluation, illuminative evaluation and case study approaches are among the strategies which will be discussed in this regard in subsequent columns.

Bibliography

Aronson, S. H. & Sherwood, C. C. Researcher versus Practitioner: Problems in Social Action Research. In C. H. Weiss (Ed.), *Evaluating Action Programs: Readings in Social Action and Education.* Boston: Allyn and Bacon, Inc., 1972. Pp. 283–293. (Originally published: *Social Work*, 1967, 12, No. 4, 89–96.)

Borich, G. D. & Drezek, S. F. Evaluating Instructional Transactions. In G. Borich (Ed.), *Evaluating Educational Programs and Products.* Englewood Cliffs, N.J.: Educational Technology, 1974.

Campbell, D. T. Reforms as Experiments. In C. H. Weiss (Ed.), *Evaluating Action Programs: Readings in Social Action and Education.* Boston: Allyn and Bacon, Inc., 1972. Pp. 187–223. (Originally published: *American Psychologist,* 1969, 24, No. 4, 409–429.)

Campbell, D. T. & Stanley, J. C. Experimental and Quasi-experimental Designs for Research on Teaching. In N. L. Gage (Ed.), *Handbook of Research on Teaching.* Chicago: Rand McNally, 1963. (Reprinted as *Experimental and Quasi-experimental Design for Research.* Chicago: Rand McNally, 1966.)

Evans, J. W. Evaluating Educational Programs—Are We Getting Anywhere? *Educational Researcher,* September 1974, 7–10.

Glennan, T. K., Jr. Evaluating Federal Manpower Programs: Notes and Observations. In C. H. Weiss (Ed.), *Evaluating Action Programs: Readings in Social Action and Education.* Boston: Allyn and Bacon, Inc., 1972. Pp. 174–186. (Originally published: Santa Monica, Calif.: Rand Corporation, 1969).

Guba, E. G. The Failure of Educational Evaluation. *Educational Technology,* 1969, 9, No. 5, 29–38. Also in C. H. Weiss (Ed.), *Evaluating Action Programs: Readings in Social Action and Education.* Boston: Allyn and Bacon, Inc., 1972. Pp. 250–266.

Houston, T. R. Behavioral Science Impact-Effectiveness Model. In P. Rossi & W. Williams (Eds.), *Evaluating Social Programs.* New York: Seminar Press, 1972.

Popham, W. J. *Educational Evaluation.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1975.

Porter, A. C. Analysis Strategies for Some Common Evaluation Paradigms. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, La., February/March, 1973.

Riecken, H. W. Memorandum on Program Evaluation. In C. H. Weiss (Ed.), *Evaluating Action Programs: Readings in Social Action and Education.* Boston: Allyn and Bacon, Inc., 1972. Pp. 85–104.

Rose, C. & Nyre, G. F. *The Practice of Evaluation.* Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement and Evaluation, 1977.

Rossi, P. H. Boobytraps and Pitfalls in the Evaluation of Social Action Programs. In C. H. Weiss (Ed.), *Evaluating Action Programs: Readings in Social Action and Education.* Boston: Allyn and Bacon, Inc., 1972. Pp. 224–235. (Originally published: Washington, D.C.: American Statistical Association, 1966. Pp. 127–132.)

Scriven, M. Evaluation Perspectives and Procedures. In W. J. Popham (Ed.), *Evaluation in Education: Current Applications.* Berkeley, Calif.: McCutchan, 1974. Pp. 3–93.

Scriven, M. The Methodology of Evaluation. In R. E. Stake (Ed.), *Perspectives of Curriculum Evaluation.* AERA Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967.

Scriven, M. Prose and Cons about Goal-free Evaluation. *Evaluation Comment,* December 1972, 3, No. 4.

Stake, R. E. The Countenance of Educational Evaluation. In C. H. Weiss (Ed.), *Evaluating Action Programs: Readings in Social Action and Eduction.* Boston: Allyn and Bacon, Inc., 1972. Pp. 31–51. (Originally published: *Teachers College Record,* April 1967, 68, No. 7, 523–540.)

Stake, R. E. Program Evaluation, Particularly Responsive Evaluation. Occasional paper #5. Kalamazoo, Mich.: Evaluation Center, Western Michigan University, November 1975.

Stanley, J. C. Controlled Field Experiments as a Model for Evaluation. In P. Rossi & W. Williams (Eds.), *Evaluating Social Programs.* New York: Seminar Press, 1972.

Stanley, J. C. Reactions to the March Article on Significant Differences. *Educational Researcher,* 1969, 20, No. 5, 8–9.

Stufflebeam, D. L. The Relevance of the CIPP Evaluation Model for Educational Accountability. *Journal of Research and Development in Education,* 1971, 5, No. 1, 19–25.

Weiss, C. H. *Evaluating Action Programs: Readings in Social Action and Education.* Boston: Allyn and Bacon, Inc., 1972a.

Weiss, C. H. *Evaluation Research: Methods of Assessing Program Effectiveness.* Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972b.

Welch, W. & Walberg, H. A National Experiment in Curriculum Evaluation. *American Educational Research Journal,* 1972, 9, 373–384.

Wergin, J. J. Evaluating Faculty Development Programs, 1976. (Unpublished paper.)

Wholey, J. S. *et al.* Proper Organizational Relationships. In C. H. Weiss (Ed.), *Evaluating Action Programs: Readings in Social Action and Education.* Boston: Allyn and Bacon, Inc., 1972. (Originally published: Federal Evaluation Policy: An Overview. A summary of the Urban Institute Study of Social Program Evaluation by federal agencies, September 1969.)