a practical look at evaluation

David Logsdon

Frequently Extension agents are asked to help local citizens judge program activities. A home economics or 4-H program may be ending and questions are being asked about the merits of the program. Did we accomplish what we set out to do? Did the participants benefit from it? What should we do differently next time? Or, maybe a funding agency is asking for an evaluative report on the program.

At this point, the agents and local citizens must turn to evaluation models in search of a set of techniques that's realistically applicable and promise useful results. The outcome of this search may well point to the experimental model, a direction which may not suit the purposes in mind. There are other models, however, which in some cases offer greater potential for manageability and usefulness. Extension agents and citizen leaders can learn a great deal by becoming familiar with the pros and cons of different evaluation models. Then they can make more informed decisions about the techniques of evaluating program activities.

Evaluation methodology is changing from the rather rigid adherence to the experimental model to a more flexible and practical position of doing whatever meets the needs and constraints. Therefore, practitioners no longer can say that evaluation methodology is awesome and impractical. [1]

Evaluative research represents a field in itself, and to respond to the peculiar research problems associated with program activity, evaluation must develop its own criteria and methodologies. To be useful, it must relate closely to the subjective world of programmers, participants, and funders for whom it's performed. [2]

Evaluation is a pluralistic phenomenon that can be flexible to meet different kinds of program needs and conditions, whether they be formal-scientific, informal-subjective, or both. Social scientists, practitioners, program participants, and funders

David Logsdon: Evaluation Coordinator, Catholic Service Bureau, Inc., Miami, Florida.

have developed different models practitioners and researchers can choose from to best serve their needs and meet practical conditions. Each model offers certain practical points of usefulness and carries its own set of advantages and disadvantages, depending on the evaluation setting and purposes.

Evaluation Purposes

When embarking on the research task, the first questions we have to ask are: Why are we evaluating? What are the most important points of practical usefulness this research should perform to help decision makers? The evaluation purposes we encounter most frequently are four-fold:

1. To provide an account of the program's effectiveness—sometimes an unquestionably convincing account, sometimes a general picture of benefits. In some cases, decision makers demand a high degree of scientific rigor, above all other considerations. In other cases, a general picture of possible benefits is enough.
2. To improve programs—to identify the strengths and weaknesses of the program process as they relate to program effects.
3. To train staff, participants, and supervisory board members in the program's effects and processes to develop an ongoing evaluation process directed at program improvement. If a program is to be repeated or operates continually, a built-in evaluation system may be essential to channel a flow of feedback into the program. If staff, participants, and board members take part in data collection and assessment, then all are trained to carry out the monitoring system.
4. To help get funding—it's a proven fact that evidence of accomplishments and attempts to improve programs convinces funders of returns on their money.

Any one or a combination of these may stand out as the primary purpose for evaluation. Usually two or three are salient, in which case they can be ranked to give a clearer picture of the model best suited to meet the purpose.

Evaluation Models

Experimental Model

The experimental model is characterized by random selection from a large pool of candidates and their assignment into a program (experimental) group and control group. The program itself is viewed as the experimental treatment. Statistical tests of significance are used to determine whether the program group achieved significantly greater goal-oriented progress than the control group.

This design carries the potential of delivering results that are the most convincing to behavioral scientists, others dedicated to pure science, and in some cases, funders. Since results may be generalizable to a large population from which the program

group was sampled, the findings may add to the wealth of knowledge about the problems under investigation. On the other hand, there's such an array of limitations associated with this model that at times its practical application is almost impossible and, at other times, totally inappropriate for the evaluation purpose.

To list some of the specific obstacles to adapting the experimental design to an action situation:

1. Programmers may find it very difficult to define and operationalize goals.
2. Randomization and control groups are precluded since the researcher maintains minimal influence over selection into the programs.
3. Developing and pre-testing goal indicators for validity and reliability may not be feasible, if standardized or pre-tested indicators aren't available.
4. The study population may be inaccessible to testing, biased against questionnaires, and experience a high dropout rate from the program and/or study—especially if the population is low income.
5. Very few controls, if any, may be possible if the evaluation is sought after the program has run its course.

Even if it were possible to manipulate these controls, the product may not be adequately useful. The main focus of the experimental design is directed toward effects, when a close look at the program process and areas of possible improvement may be equally important. Characteristically, the research report isn't ready until months after the program is over, and the document winds up being read by very few and ignored by most.

The richness of the program's effects, as well as the participants' and staff's feelings, have no place in a controlled experiment. If you want to use the evaluation as a learning experience for staff and participants—learning about the program's benefits, strengths, weaknesses, and the actor's feelings, as well as the evaluation process itself, then the experimental design won't serve the purpose.

Survey of Subjective Opinions and Skills Learned Model

Programmers frequently need an evaluation that can be rapidly and easily accomplished, a study that provides a readable and usable assessment of program inputs as well as benefits. They need an evaluation that relates directly to the staff's and participants' experience. And, it's not necessary that the results be "unquestionably convincing." With this model, objective tests of the skills learned can be given before and after the program, as well as periodically thereafter, to determine the degree to which the developed skills were retained and put to use. Participants' and staff's reactions to program inputs can also be gathered.

To illustrate further, a home economics program may be directed at helping low-income homemakers plan and serve more nutritious meals to their families. In collaboration with the programmers, this goal could be defined in terms of a percentage of participants who, over time, serve their families the foods that represent desired amounts of protein, vitamins, and minerals. The goal could be that at least 67% of the participants serve, for at least 1 year after the program, meals that, according to nutrition standards, are high in protein, low in carbohydrates, and high in needed vitamins and minerals.

Evaluation is a pluralistic phenomenon that can be flexible to meet different kinds of program needs and conditions, whether they be formal-scientific, informal-subjective, or both.

To measure the program's effectiveness, before and after the program, we determine the number of participants who serve meals that meet the nutrition standards. This can be done by simply asking the homemakers before and after the program what they usually serve. The homemaker's spouse and children can also be interviewed to validate these reports and get their reactions to the changes. The after measures should be continued periodically for about one year. In addition, participants' opinions about the program benefits, strengths, weaknesses, and needed changes should be collected.

The evaluation would address itself to the following questions:

1. What percentage of participants were serving nutritious meals before the program, immediately after the program and up to a year thereafter? What changes occurred after the program?
2. To what degree can we attribute these changes to the program? In other words, what percentage of participants report that the program helped them to make the changes?
3. What percentage of the homemakers' families reacted favorably to the changes?
4. What do the participants report are the strengths, weaknesses, and needed changes in the program?

The main limitations to this design concerns the questionable validity of the results. If before/after tests are employed, the test itself may get in the way, as might other contaminating factors associated with testing and questionnaires in general. If participants' opinions represent the basis of evidence, then you may find some decision makers doubting the accuracy of the testimonies.

After working with this design for several years, we have our own opinions about the question of validity. In the first place, little reason exists to doubt a participant's assertion that, "this tutoring program has really helped my child because he reads a lot more now," or "I know it (the program) helped him because he gets along better with his brothers and sisters." Maturation doesn't cause such significant changes over a brief period of time.

One opinion can become more convincing when it's supported by two additional opinions from two other sources. For instance, if a child's parents *and* teachers agree that she's benefiting. We've found that when the reader of the evaluation report is confronted with one testimony after another of the program's benefits, or lack of benefits, and when these opinions are accompanied by frequency counts indicating the percentage of respondents benefiting, the reader is convinced that something closely approximating these testimonies and ratings did occur.

The advantage of this model is that it brings evaluation closer to the lives of the actors involved in the program at the expense of a degree of validity.

Group Process Model

The group process model draws the evaluation, the actors, and the program even closer together. Evaluation completely loses the aura of an esoteric exercise arbitrarily manipulated by awesome researchers. It becomes a learning experience for everyone, especially the researchers. From their involvement in the evaluation process, the actors ideally gain first-hand experience in research and assessment, thereby being trained to integrate and continue the evaluation mechanism vis-a-vis future programming.

This model was used rather effectively to evaluate four community action commission (CAC) projects: one urban and one rural community development project, a youth project, and an adult education project. A group process seemed to meet the practical needs and limitations of the evaluation setting. The CAC staff wanted to know: After the investment of hundreds of contact hours, how had the low-income residents benefited? The funding cycle was coming to an end and a thorough report was needed for the Office of Economic Opportunity.

The situation called for an evaluation model that was manageable, adaptable to a brief time span, and inexpensive to implement. It was especially important that the research process parallel the program approach of staff, board, and participants working closely to solve problems.

The essential components of the model are a series of skillfully led group discussions in which board members, staff, and participants sit down together and review program goals, processes, benefits—including agency records, other statistical data, and everyone's perceptions, problems, and recommend-

ations for improvement. All nonrepetitious comments are recorded on "tear sheets" or "butcher paper" and later compiled for review and reporting. These discussions may occur over a series of three or four meetings with a process like this:

1. First meeting.
   a. Orientation in full group.
      Researcher explains the purpose and procedures of the evaluation process. Everyone introduces himself and identifies his role. Ground rules for discussion are reviewed:
         Everyone should contribute; all comments are brief and to the point; the group makes certain that all nonrepetitious comments are recorded; comments in no way identify any participant, staff, or board member by name; no comments are directed personally at any group member.
   b. Small groups formed and assigned to meeting rooms.
   c. Small groups chaired by discussion leaders.
      1) Recording secretary is appointed.
      2) Evaluation questions are presented and answers recorded:
         a) What are the program goals?
         b) What evidence of goal attainment is available in terms of pre-collected agency records?
         c) Records are summarized to the group and submitted to recording secretary.
         d) What are the program participants' perceptions of program benefits, strengths, weaknesses, and needed changes?
            To answer this question, members give their opinions and are asked to call 10 randomly pre-selected program participants each during the next week to gather their perceptions of benefits, strengths, weaknesses, and needed changes.

. . . it's a proven fact that evidence of accomplishments and attempts to improve programs convinces funders of returns on their money.

2. Second meeting (in small groups)—one week later. All participants return with data gathered from phone calls. If feasible, results from each interview are read aloud and summarized before the group. Participants react to data and general picture that emerges.
3. Third meeting (in small groups)—three weeks later. All data from above meetings have been roughly

compiled into a report organized around the evaluation questions. Copies of report are distributed to all members: All data are reviewed. All are included and accurately reported.

4. Fourth meeting (full group session).
   Discussion led by one of the group leaders and recorded: Now that we have a general picture of accomplishments, what does this say about the needs and problems the program is supposed to be attacking? Is the program really meeting the needs it was intended to meet? What should be the future course of the program? Assessment of evaluation process: participants, staff, and board members report on what they've learned—strengths, weaknesses, and needed changes in process. Should the process be repeated? How can an evaluation process be built into regular functioning?
5. Data from the meeting are added and report is finished in the next few weeks. An evaluation committee of researchers, participants, staff, and board members is formulated to carry out plans for an ongoing evaluation system.

Since this design, as in the case of the previous one, relies on individuals' opinions, it's limited by the questionable validity of the findings. The same strategies to deal with this limitation apply here. In addition, it may be possible, if the program continues, to augment the opinions and agency records with case study and survey data on skills learned. Another limitation concerns the requirement that all actors must gather together in an open, lively, yet manageable, series of meetings, which may prove practically impossible and strategically unwise.

Summary

A precipitating reason for conducting evaluation may well be pressure from funding sources. Although evaluation guidelines are offered in some cases, we've found that the design can be constructed within a degree of flexibility that views both the values and limitations to implementing the experimental design. Consequently, the question of design becomes a matter of doing the best you can in terms of valid results and meeting your own information needs. There are some funding sources, however, especially at the federal level, that will support primarily experimental design evaluation.

Although these three designs are reviewed in exclusion of one another so we could analyze points of primary usefulness associated with each design—two or three designs may be applied together for maximum evaluation usefulness. For instance, in a group process setting, both experimental data and subjective reactions may be introduced as evidence of goal attainment. The experimental findings would have been collected earlier and the subjective data could be gathered during the group

meetings, as well as from the participants' assigned interviews. If group meetings aren't possible, experimental data may be augmented by participants' perceptions to yield a richer view of the program's impact.

To aid the decision-making process, evaluation can assume a variety of forms depending on the utility the evaluation is to serve and the practical constraints involved in implementing a design.

If the primary purpose is to produce unquestionably valid results and the design controls are manageable, the controlled experimental model should be used.

If the needs are for program improvement as well as accounting for effects, with some flexibility about validity, then a survey of subjective opinions and skills learned may be appropriate.

When the research is sought as a tool for integrating evaluation into program planning and development with the possibility of gathering the actors together for constructive discussion, the group process model may best serve the purpose.

And, if the greatest usefulness of evaluation is wanted, the three designs may be combined to yield sound data, a basis for program improvement, and a learning process.

Footnotes

1. Edward A. Suchman, *Evaluative Research: Principles and Practice in Public Service and Social Action Programs* (New York: Russell Sage Foundation, 1967), p. 12 and Tom R. Houston, "Behavioral Sciences Impact Effectiveness Model," in *Evaluating Social Programs, Theory, Practice and Politics*, Peter H. Rossi and Walter Williams, eds. (New York: Seminar Press, 1972), p. 52, 63.
2. Sara M. Steele, "Validity, Reliability, and Appropriateness in Program Evaluation As Determined by Value to Adult Education" (Paper presented at the Adult Education Research Conference Workshop on Evaluation, Montreal, Canada, 1973).
3. When it's not possible to develop a control group, we can at least ask these kinds of questions.