CAROL H. WEISS
Columbia University

EVALUATION RESEARCH

Methods for Assessing Program Effectiveness

PRENTICE-HALL, INC., Englewood Cliffs, New Jersey

1

Introduction

Evaluation is an elastic word that stretches to cover judgments of many kinds. People talk about evaluation of a worker's job performance, evaluation of a movie script, evaluation of the sales potential of a new detergent. What all the uses of the word have in common is the notion of judging merit. Someone is examining and weighing a phenomenon (a person, a thing, an idea) against some explicit or implicit yardstick.

In this book we will be talking about evaluation of one particular kind of phenomenon: social programs designed to improve the lot of people. The programs are diverse; they can deal with education, social welfare, health, housing, mental health, legal services, corrections, economic development, and many other fields. They can be aimed to change people's knowledge, attitudes, values, behaviors, the institutions with which they deal, or the communities in which they live. Their common characteristic is the goal of making life better and more rewarding for the people they serve.

Furthermore, we are concerned here with a specific method of evaluation—evaluation research. The tools of research are pressed into service to make the judging process more accurate and objective. In its research guise, evaluation establishes clear and specific criteria for success. It collects evi-

dence s, .natically from a representative sample of the units of concern. It usually translates the evidence into quantitative terms (23 percent of the audience, grades of 85 or better), and compares it with the criteria that were set. It then draws conclusions about the effectiveness, the merit, the success, of the phenomenon under study.

The research process takes more time and costs more money than off-hand evaluations that rely on intuition, opinion, or trained sensibility, but it provides a rigor that is particularly important when (1) the outcomes to be evaluated are complex, hard to observe, made up of many elements reacting in diverse ways; (2) the decisions that will follow are important and expensive; and (3) evidence is needed to convince other people about the validity of the conclusions.

In the past decade social programs at all levels have expanded enormously. Some are logical extensions of earlier efforts, some represent radical departures from the past and a plunge into uncharted waters. Decision makers want (and need) to know: How well is the program meeting the purposes for which it was established? Should it be continued, expanded, cut back, changed, or abandoned? The answers are hard to come by through informal means. The best informed people (the staff running the program) tend toward optimism and in any case have a stake in reporting success. Many programs provide a variety of services and deal with large numbers of participants. A handful of "consumer testimonials" or a quick tour of inspection can hardly gauge their effectiveness. Decisions about future operations will affect the fate of many people and involve sizable sums of money, and the decision makers are often people (legislators, boards of directors) sufficiently removed from the program to want hard facts on which to base their decisions. Under these conditions, evaluation research appears well suited to the task of producing the requisite information, and in recent years it has become a growth enterprise.

Contributions to Rational Decision Making

Evaluation research is viewed by its partisans as a way to increase the rationality of policy making. With objective information on the outcomes of programs, wise decisions can be made on budget allocations and program planning. Programs that yield good results will be expanded; those that make poor showings will be abandoned or drastically modified. The following excerpt from Congresswoman Dwyer's (Republican, New Jersey) *Report to the People*, although it does not mention evaluation research, captures the rationale of the case for evaluation:

> It is becoming increasingly clear that much of our investment in such areas as education, health, poverty, jobs, housing, urban development, transportation and the like is not returning adequate dividends in terms of results. Without for a moment lessening our commitment to provide for these pressing human needs, one of Congress' major, though oft-delayed, challenges must be to reassess our multitude of social programs, concentrate (indeed, expand) resources on programs that *work* where the needs are greatest, and reduce or eliminate the remainder. We no longer have the time nor the money to fritter away on non-essentials which won't produce the needed visible impact on problems.[1]

Both on the national and the local scale, the application of social science knowledge and methodology is expected to have beneficial effects: improve decision making, lead to the planning of better programs, and so serve program participants in more relevant, more beneficial, and more efficient ways. The production of objective evidence is seen as a way to reduce the politicking, the self-serving maneuvers, and the log-rolling that commonly attend decision making at every level from the Congress to the local school. Data will replace favors and other political negotiations, so that the most rational decisions will be reached.

In these terms, the history of evaluation research to date has been disappointing. Few examples can be cited of important contributions to policy and program. Part of the reason lies in the remarkable resistance of organizations to unwanted information—and unwanted change. Even evidence of outright failure can leave some institutions figuratively and literally unmoved. Part of the fault lies in the way evaluation itself is structured, staffed, and operated. There are fissures between the intended purposes of evaluation and the kinds of studies conducted. That indeed is the subject of much of this book.

But part of the disillusionment with the contributions of evaluation derives from the unrealistic nature of the expectations. An evaluation study does not generally come up with final and unequivocal findings about the worth of a program. Its results often show small, ambiguous changes, minor effects, outcomes influenced by the specific events of the place and the moment. It may require continued study over time and across projects to speak with confidence about success and failure.

Furthermore, for decision makers, evaluation evidence of outcome is only one input out of many. They must consider a host of other factors, from public receptivity and participant reaction, to costs, availability of

[1] Rep. Florence P. Dwyer, *Report to the People*, 12th District New Jersey, XIV, No. 1, January 22, 1970.

staff and facilities, and possible alternatives. Those who look to evaluation to take the politics out of decision making are bound to be disappointed. Within every organization, decisions are reached through negotiation and accommodation, through politics. This is the system we have for attaching *value* to facts. Different actors bring different values and priorities to the decision-making process. Evaluative facts have an impact on collective decisions only to the extent that program effectiveness is perceived as valuable. And program effectiveness—inevitably and justifiably—competes for influence on decisions with considerations of acceptability, feasibility, and ideology. Sometimes it is emotionally and politically rewarding to run a program even when it has been shown to have little effect if the alternative is to do nothing for a particular group. Sometimes the existing ideological climate precludes the adoption of more effective programs if these violate cherished assumptions and values.

It is within this context that evaluation should be viewed. What evaluation *can* do is provide data that reduce uncertainties and clarify the gains and losses that different decisions incur. In this way, it allows decision makers to apply their values and preferences more accurately, with better knowledge of the trade-offs that alternative decisions involve.

Purpose of Evaluation Research

The purpose of evaluation research is to measure the effects of a program against the goals it set out to accomplish as a means of contributing to subsequent decision making about the program and improving future programming. Within that definition are four key features: "To measure the effects" refers to the *research methodology* that is used. "The effects" emphasizes the *outcomes* of the program, rather than its efficiency, honesty, morale, or adherence to rules or standards. The comparison of effects with goals stresses the use of explicit *criteria* for judging how well the program is doing. The contribution to subsequent decision making and the improvement of future programming denote the *social purpose* of evaluation.

Programs are of many kinds. Not only do they range over a gamut of fields; they also vary in scope, size, duration, clarity and specificity of program input, complexity of goals, and innovativeness. These differences in programs have important consequences for the type of evaluation that is feasible and productive. It is one thing to evaluate the effects of a small, short-term, specific, well-defined program, such as a training film. It is a far different and more difficult matter to evaluate the effects of the national antipoverty program, with its diversity of methods, actions, and goals. The evaluator may find it rewarding to become aware of some of the differences

among programs so that he can think about ways to shape evaluative approaches and method to suit.

Scope. The program being evaluated may cover the nation, a region, state, city, neighborhood, or be limited to one specific site (a classroom). Some programs turn up in scattered locations (a methadone treatment program for drug addicts in ten hospitals around the country).

Size. Programs can serve a few people or reach thousands or even millions.

Duration. A program can last a few hours, days, or weeks, a specified number of months or years, or go on indefinitely (the Boy Scout program, public school education).

Clarity and specificity of program input. What it is that the program actually *does* may be well-defined and precise; for example, brighter street lights may be installed on given streets in an attempt to reduce crime. Many programs have some degree of clarity (a new science curriculum, foster home placement), since a particular method or specific materials are being employed, but different staff members may vary in style and skill in administering them. At the extreme there are programs that are diffuse, highly variable, and difficult even to describe (a program of interagency planning).

Complexity and time span of goals. Some programs are intended to produce a clear-cut change or changes (improvement in reading skills, placement in a job). Others seek more complex goals (make children better citizens, improve mental health, improve family functioning) that are harder to define and measure. A goal such as "improving the quality of urban life" contains within it not only a large number of subgoals (that must be made explicit) but also ambiguous subgoals (improving the esthetics of the urban scene) that pose awesome problems of conceptualization and measurement.

Another issue is the time span of the goals. It is easier for the evaluator to deal with intended changes that manifest themselves quickly than with those that become evident or sure only after half a lifetime.

Innovativeness. At one end of the continuum are programs that mark a drastic shift from accustomed methods of operation. At the other are regular ongoing programs of established agencies.

The characteristics of the program will affect the kind of evaluation that can be done and the purposes that evaluation can serve. In Chapter 2 we will examine the subject of purpose in greater detail. One of the prob-

lems in doing good evaluation research is that different people see different purposes for the evaluation and want to use its results in different ways. Unless and until the evaluator finds out specifically who wants to know what, with what end in view, the evaluation study is likely to be mired in a morass of conflicting expectations.

Comparison Between Evaluation and Other Research

Evaluation applies the methods of social research. Principles and methods that apply to all other types of research apply here as well. Everything we know about design, measurement, and analysis comes into play in planning and conducting an evaluation study. What distinguishes evaluation research is not method or subject matter, but intent—the purpose for which it is done.

Differences

Use for decision making. Evaluation is intended for use. Where basic research puts the emphasis on the production of knowledge and leaves its use to the natural processes of dissemination and application, evaluation starts out with *use* in mind. In its ideal form, evaluation is conducted for a client who has decisions to make and who looks to the evaluation for answers on which to base his decisions. Use is often less direct and immediate than that, but it always provides the rationale for evaluation.

Program-derived questions. The questions that evaluation considers are the decision maker's questions rather than the evaluator's. Unlike the basic researcher who formulates his own hypotheses, the evaluator deals in the currency of program concerns. He has a good deal of say about the shape of the study, and he approaches it from the perspectives of his own knowledge and discipline. He is usually free to embroider it with investigations of particular concern to him. But the core of the study represents matters of administrative and programmatic interest. The common evaluation hypothesis is that the program is accomplishing what it set out to do.

Judgmental quality. Evaluation compares "what is" with "what should be." Although the investigator himself remains unbiased and objective, he is concerned with phenomena that demonstrate whether the program is achieving its intended goals. However the questions for study are formulated, somewhere in the formulation appears a concern with measuring

up to stated criteria. This element of judgment against criteria is basic to evaluation and differentiates it from other kinds of research. The statement of program goals by the staff of the program is therefore essential to evaluation. It comes as a particular blow to discover that programs do not generally have clear statements of goals. In Chapter 3 we will explore the problems involved.

Action setting. Evaluation takes place in an action setting, where the most important thing that is going on is the program. The program is serving people. If there are conflicts in requirements between program and evaluation, priority is likely to go to program. Program staff often control access to the people served in the program. They may control access to records and files. They are in charge of assignment of participants to program activities and locations. Not infrequently, research requirements (for "before" data, for control groups) run up against established program procedures, which tend to prevail.

Role conflicts. Interpersonal frictions are not uncommon between evaluators and practitioners. The practitioners' roles and the norms of their service professions tend to make them unresponsive to research requests and promises. As they see it, the imperative is service; evaluation research is not likely to make such contributions to the improvement of program service that it is worth disruptions and delays. Often, they believe strongly in the worth of the program they are providing, and see little need for evaluation at all. Furthermore, the judgmental quality of evaluation research means that the merit of their activities is being weighed. In a sense, as they see it, they are on trial. If the results of evaluation are negative, if it is found that the program is not accomplishing the purposes for which it was established, then the program—and possibly their jobs—are in jeopardy. The possibilities for friction are obvious.

Publication. Basic research is published. Its dissemination to the research and professional fraternity is essential and unquestioned. In evaluation, probably the majority of study reports go unpublished. Program administrators and staff often believe that the information was generated to answer their questions, and they are not eager to have their linen washed in public. Evaluators are sometimes so pressed for time, or so discouraged about the compromises they have made in research design, that they submit a mimeographed report to the agency and go on to the next study. Yet if progress is to be made in learning which types of programs work and which do not, a cumulative information base is essential. Only through publication will results build up. Even when results show that the program

has had little effect, it is important that others learn of the findings so that ineffective programs are not duplicated again and again.

Of course, not all evaluation studies are worth publication. Poorly conducted studies are more misleading than useful. Further, if the evaluator has addressed the issues in such concrete and specific terms that his results are not generalizable beyond the immediate program, there is little to report to others. Hovland makes a distinction between "program testing" and "variable testing." If only the specific program has been tested and not the concepts or the approaches (variables) on which it is based, the study makes little contribution to developing knowledge.

Allegiance. The evaluation researcher has a dual, perhaps a triple, allegiance. He has obligations to the organization that funds his study. He owes it a report of unqualified objectivity and as much usefulness for action as he can devise. Beyond the specific organization, he has responsibilities to contribute to the improvement of social change efforts. Whether or not the organization supports the study's conclusions, the evaluator often perceives an obligation to work for their application for the sake of the common weal. On both counts, he has commitments in the action arena. He also has an obligation to the development of knowledge and to his profession. As a social scientist, he seeks to advance the frontiers of knowledge about how intervention affects human lives and institutions.

If some of the differences between evaluation research and more academic social research have made the lot of the evaluator look unduly harsh, there are compensations. One of the most rewarding is the opportunity to participate actively in the meeting of scientific knowledge and social action and to contribute to the improvement of societal programs. It is this opportunity that has attracted so many able researchers to the field of evaluation research despite the disabilities that attend its practice.

Similarities

There are important similarities, too, between evaluation and other brands of research. Like other research, evaluation attempts to describe, to understand the relationships between variables, and to trace out the causal sequence. Because it is studying a program that intervenes in people's lives with the intention of causing change, evaluation can often make direct inferences about the causal links that lead from program to effect.

Evaluators use the whole gamut of research methods to collect information—interviews, questionnaires, tests of knowledge and skill, attitude inventories, observation, content analysis of documents, records,

examination of physical evidence. Ingenious evaluators can find fitting ways of exploring a wide range of effects. The kind of data-collection scheme to be used depends on the type of information needed to answer the specific questions that the evaluation poses.

The classic design for evaluations has been the experimental model. This involves measurement of the relevant variables for at least two equivalent groups—one that has been exposed to the program and one that has not. But many other designs are used in evaluation research—case studies, post-program surveys, time series, correlational studies, and so on. The experimental model that has long reigned as the ideal (if often neglected) design for evaluation research has recently been challenged on several grounds. We will discuss these issues further in Chapter 4.

There is no cut-and-dried formula to offer evaluators for the "best" or most suitable way of pursuing their study. Much depends on the uses to be made of the study, the decisions pending, and the information needs of decision makers. Much also depends (unfortunately) on the constraints in the program setting—the limits placed on the study by the realities of time, place, and people. Money is an issue, too. Textbooks rarely mention the grubby matter of funding, but limited funds impose inevitable restrictions on how much can be studied over how long a period. Thus evaluation methods often represent a compromise between the ideal and the feasible.

Evaluation is sometimes regarded as a lower order of research, particularly in academic circles, than "basic" or "pure" research. Evaluators are looked down on as the drones of the research fraternity, technicians drudging away on dull issues and compromising their integrity out in the corrupt world. But as any working evaluator will heartfeelingly tell you, evaluation calls for a higher level of skills than research that is under the researcher's complete control. It is relatively easy to run experiments in an insulated laboratory with captive subjects. But to make research work when it is coping with the complexities of real people in real programs run by real organizations takes skill—and some guts. The evaluator has to know a good deal about the formulation of the research question, study design, sampling, measurement, analysis, and interpretation. He has to know what is in the research methodology texts, and then he has to learn how to apply that knowledge in a setting that is often inhospitable to important features of his knowledge. If he persists in his textbook stance, he runs the risk of doing work irrelevant to the needs of the agency, antagonizing the program personnel with whom he works, and seeing his study results go unused—if indeed the work is ever completed. So he sometimes has to find alternative ways of conducting his study, while at the same time he stands ready to defend to the death those elements of the study that cannot be compromised.

2

Purposes of Evaluation

In this chapter, we will discuss the purposes, acknowledged and unacknowledged, for which people decide to undertake program evaluation. We suggest that the evaluator find out what decision makers really seek from the study and how they expect to use the results. With this knowledge, he can most effectively tailor the study to provide information for decision making. The location of the evaluation unit—where it fits into the organizational structure—can make a difference in whether the study has sufficient latitude to be useful.

Before we get on with these matters, let us raise a prior question. Is evaluation always warranted? Should all programs if they are good little programs go out and get themselves evaluated? The answer, heretical as it may seem, is No. Evaluation as an applied research is committed to the principle of utility. If it is not going to have any effect on decisions, it is an exercise in futility. Evaluation is probably not worth doing in four kinds of circumstances:

1. When there are no questions about the program. It goes on, and decisions about its future either do not come up or have already been made.

2. When the program has no clear orientation. Program staff improvise activities from day to day, based on little thought and less principle, and the program shifts and changes, wanders around and seeks direction. There is little here to call "a program."
3. When people who should know cannot agree on what the program is trying to achieve. If there are vast discrepancies in perceived goals, evaluation has no ground to stand on.
4. When there is not enough money or no staff sufficiently qualified to conduct the evalution. Evaluation is a demanding business, calling for time, money, imagination, tenacity, and skill.

There are those who argue that even in such dismal circumstances, evaluation research can produce something of value, some glimmering of insight that will light a candle for the future. This is a fetching notion, and from time to time in this volume, we succumb to it. But experience suggests that even good evaluation studies of well-defined programs, directed to clear decisional purposes, often wind up as litter in the bureaucratic mill. It will be a rare study indeed that provides illumination under unfavorable conditions.

Overt and Covert Purposes

People decide to have a program evaluated for many different reasons, from the eminently rational to the patently political. Ideally, an administrator is seeking answers to pressing questions about the program's future: Should it be continued? Should it be expanded? Should changes be made in its operation? But there are occasions when he turns to evaluation for less legitimate reasons.

Postponement. The decision maker may be looking for ways to delay a decision. Instead of resorting to the usual ploy of appointing a committee and waiting for its report, he can commission an evaluation study, which takes even longer.

Ducking responsibility. Sometimes one faction in the program organization is espousing one course of action and another faction is opposing it. The administrators look to evaluation to get them off the hook by producing dispassionate evidence that will make the decision for them. There are cases in which administrators know what the decision will be even before they call in the evaluators, but want to cloak it in the legitimate trappings of research.

Public relations. Occasionally, evaluation is seen as a way of self-glori-

fication. The administrator believes that he has a highly successful program and looks for a way to make it visible. A good study will fill the bill. Copies of the report, favorable of course, can be sent to boards of trustees, members of legislative committees, executives of philanthropic foundations who give large sums to successful programs, and other influential people. Suchman [1] suggests two related purposes: eyewash and whitewash. In an eyewash evaluation, an attempt is made to justify a weak program by selecting for evaluation only those aspects that look good on the surface. A whitewash attempts to cover up program failure by avoiding any objective appraisal.

The program administrator's motives are not, of course, necessarily crooked or selfish. Often, there is a need to justify the program to the people who pay the bills, and he is seeking support for a concept and a project in which he believes. Generating support for existing programs is a common motive for embarking on evaluation.

Fulfilling grant requirements. Increasingly, the decision to evaluate stems from sources outside the program. Many federal grants for demonstration projects and innovative programs are tagged with an evaluation requirement; for example, all projects for disadvantaged pupils funded under Title I of the Elementary and Secondary Education Act are required to be evaluated.

From the point of view of the funders, who are taking a chance on an untried project, it is reasonable to require that there be some evidence on the extent to which the project is working. To the operators of a project, the demands of starting up and running the new program take priority. Plagued as they often are by immediate problems of staffing, budgets, logistics, community relations, and all the other trials of pioneers, they tend to neglect the evaluation. They see it mainly as a ritual designed to placate the funding bodies, without any real usefulness to them.

Evaluation, then, is a rational enterprise often undertaken for nonrational, or at least noninformational, reasons. We could continue the catalog of the varieties of covert purposes (justifying a program to Congress, "getting" the program director, increasing the prestige of the agency), but the important point is that such motives have consequences for the evaluation that can be serious and bleak.[2]

[1] Edward A. Suchman, "Action for What? A Critique of Evaluative Research," in *The Organization, Management, and Tactics of Social Research*, ed. Richard O'Toole (Cambridge, Mass.: Schenkman Publishing Co., Inc., 1970).

[2] See Sar Levitan, "Facts, Fancies, and Freeloaders in Evaluating Antipoverty Programs," *Poverty and Human Resources Abstracts*, IV, No. 6 (1969), 13–16; Richard H. Hall, "The Applied Sociologist and Organizational Sociology," in *So-*

An evaluator who is asked to study a particular program usually assumes that he is there because people want answers about what the program is doing well and poorly. When this is not the case, he may in his naiveté become a pawn in intraorganizational power struggles, a means of delaying action, or the rallying point for one ideology or another. Some evaluators have found only after their study was done that they had unwittingly played a role in a larger political game. They found that nobody was particularly interested in applying their results to the decisions at hand, but only in using them (or any quotable piece of them) as ammunition to destroy or to justify.

Lesson No. 1 for the evaluator newly arrived on the scene is: Find out who initiated the idea of having an evaluation of the program and for what purposes. Were there other groups in the organization who questioned or objected to the evaluation? What were their motives? Is there real commitment among practitioners, administrators, and/or funders to using the results of the evaluation to improve future decision making? If the real purposes for the evaluation are not oriented to better decision making and there is little commitment to applying results, the project is probably a poor candidate for evaluation. The evaluator might well ponder whether he wishes to get involved in the situation or whether he can find more productive uses for his talents elsewhere.

Intended Uses

Even when evaluation is undertaken for bona fide purposes (that is, to learn how well the program is reaching its goals), people can have widely differing expectations of the kinds of answers that will be produced. If the evaluator is not to be caught unawares, it behooves him to know from the outset what kinds of answers are expected from his study.[3]

---

ology in Action, ed. Arthur B. Shostak (Homewood, Ill.: Dorsey Press, Inc., 1966), pp. 33–38; Joseph W. Eaton, "Symbolic and Substantive Evaluative Research," *Administrative Science Quarterly*, VI, No. 4 (1962), 421–42; Lewis A. Dexter, "Impressions About Utility and Wastefulness in Applied Social Science Studies," *American Behavioral Scientist*, IX, No. 6 (1966), 9–10.

[3] Downs makes the point that the extent of applied research should be economically justified by the value of the information it produces for decision making. Evaluators, like other researchers, can become fascinated with the problem and do more research than the program needs. But he also stresses the point that clients frequently need redefinition of the problem and the suggestion of alternative approaches. Anthony Downs, "Some Thoughts on Giving People Economic Advice," *American Behavioral Scientist*, IX, No. 1 (1965), 30–32. Of course, far more common than spending too much money is trying to conduct evaluation with funds grossly inadequate for the extent and precision of the results expected.

Who expects what?

Expectations for the evaluation generally vary with a person's position in the system.[4] Top policy makers need the kind of information that will help them address the broad issues: Should the program be continued or dropped, institutionalized throughout the system or limited to a pilot program, continued with the same procedures and techniques or modified? Should more money be allocated to this program or to others? They want information on the overall effectiveness of the program.

The directors of the program face other issues. They want to know not only how well their program is achieving the desired ends, but also which general strategies are more or less successful, which are achieving results most efficiently and economically, which features of the program are essential and which can be changed or dropped.

Direct-service staff deal with individuals and small groups. They have practical day-to-day concerns about techniques. Should they spend more time on developing good work habits and less time on teaching subject matter? Put more emphasis on group discussions or films or lectures? Should they accept more younger people (who are not already set in their ways) or more older people (who have greater responsibilities and more need)? Practitioners, who are accustomed to relying on their own experience and intuitive judgment, often challenge evaluation to come up with something practical on topics such as these.

Nor do these three sets of actors—policy makers, program directors, and practitioners—exhaust the list of those with a possible oar in the evaluation. The funders of evaluation research, particularly when they are outside the direct line of operations, may have an interest in adding to the pool of knowledge in the field. They may want answers less to operating questions than to questions of theory and method. Can social group work help improve the parental performance of young couples? Does increasing the available career opportunities for low-income youth result in less juvenile deliquency? If coordination among community health services is increased, will people receive better health care? Here is another purpose for evaluation—to test propositions about the utility of concepts or models of service. The public too has a stake, as taxpayers, as parents of schoolchildren, as contributors to voluntary organizations.[5] They are concerned that their money is wisely and efficiently spent.

[4] A useful discussion appears in Louis Ferman, "Some Perspectives on Evaluating Social Welfare Programs," *Annals of the American Academy of Political and Social Science*, Vol. 385 (September 1969), 143–56.

[5] Edward Wynne, in "Evaluating Educational Programs: A Symposium," *Urban Review*, III, No. 4 (1969), 19–20.

Recently, another actor has entered the decision-making arena—the consumer of services. He may see a use for evaluation in asking "client-eye" questions about the program under study. Is the program serving the goals that the intended beneficiaries of service value?[6] Recently, there has been rising opposition, particularly in some black communities, to traditional formulations of program goals.[7] Activists are concerned not only with how well programs work to improve school achievement or health care, but also with their political legitimacy. They are interested in community participation or community control of programs and institutions. When such issues are paramount, evaluative questions derive from a radically different perspective.

Compatibility of purposes

With all the possible uses for evaluation to serve, the evaluator has to make choices. The all-purpose evaluation is a myth. Although a number of different types of questions can be considered within the bounds of a single study, this takes meticulous planning and design. Inevitably not even the best-planned study will provide information on all the questions that people will think of. In fact, some purposes for evaluation are incompatible with others. Let us consider the evaluation of a particular educational program for slow learners.

The teaching staff wants to use the results to improve the presentations and teaching methods of the course, session by session, in order to maximize student learning. The state college of education wants to know whether the instructional program, based on a particular theory of learning, will improve pupil performance. In the first case, the evaluator will have to examine immediate short-term effects (learnings after the morning drill). He need not be concerned about generalizing the results to other populations, and needs neither control groups nor sophisticated statistics. He will want to maximize feedback of results to the teachers so that they can modify their techniques as they go along.

On the other hand, when evaluation is testing the proposition that a program developed from certain theories of learning will be successful with slow learners, it is concerned with long-range effects. It requires rigorous design so that observed results can be attributed to the stimulus of the

[6] Philip H. Taylor, "The Role and Function of Educational Research," *Educational Research*, IX, No. 1 (1966), 11–15; Edmund deS. Brunner, "Evaluation Research in Adult Education," *International Review of Community Development*, No. 17–18 (1967), 97–102.

[7] David K. Cohen, "Politics and Research: Evaluation of Social Action Programs in Education," *Review of Educational Research*, XL, No. 2 (1970), 232.

program and not to extraneous events. The results have to be generalizable beyond the specific group of students. The instructional program should be insulated from alterations during its course in order to preserve the clarity of the program that led to the effects observed.

In theory, it is possible to achieve both an assessment of overall program effectiveness and a test of the effectiveness of component strategies. Textbooks on the design of experiments [8] present methods of factorial design that allow the experimenter to discover both total effect and the effects of each "experimental treatment." In practice, evaluation can seldom go about the business so systematically. The constraints of the field situation hobble the evalution—too few clients, demand for quick feedback of information, inadequate funds, "contamination" of the special-treatment groups by receipt of other services, drop-outs from the program, lack of access to records and data, changes in program, and so on.

Some researchers say that to try to satisfy a multiplicity of demands and uses under usual field conditions invites frustration. The evaluator who identifies the key decision pending and gears his study to supplying information relevant to that issue is on firmer ground. Others believe that there are ways—not necessarily formal and elegant—to study a range of issues concurrently.[9] Some of these methods will be discussed in Chapters 3 and 4. Nevertheless, it remains important for the evaluator to know the priority among the purposes. If the crunch comes, he can jettison the extra baggage and fight for the essentials.

Formative and summative evaluation

We have identified several types of uses for evaluation. Evaluation can be asked to investigate the extent of program success so that decisions such as these can be made:

1. To continue or discontinue the program
2. To improve its practices and procedures
3. To add or drop specific program strategies and techniques
4. To institute similar programs elsewhere

[8] A good example is B. J. Winer, *Statistical Principles in Experimental Design* (New York: McGraw-Hill Book Company, 1962). F. Stuart Chapin, W. G. Cochran and G. M. Cox, D. R. Cox, A. L. Edwards, R. A. Fisher, R. E. Kirk and E. F. Lindquist, among others, have also written useful texts on experimental design. Some of these are listed in the third section of the bibliography.

[9] See Robert E. Stake, "Generalizability of Program Evaluation: The Need for Limits," and James L. Wardrop, "Generalizability of Program Evaluation: The Dangers of Limits," *Educational Product Report*, II, No. 5 (1969), 38–40, 41–42.

5. To allocate resources among competing programs
6. To accept or reject a program approach or theory

A useful distinction has been introduced into the discussion of purpose by Scriven.[10] In discussing the evaluation of educational curriculums, he distinguishes between *formative* and *summative* evaluation. Formative evaluation produces information that is fed back during the development of a curriculum to help improve it. It serves the needs of developers. Summative evaluation is done after the curriculum is finished. It provides information about effectiveness to school decision makers who are considering adopting it.[11]

This distinction can be applied to other types of programs as well, with obvious advantages for the clarification of purpose. Many programs, however, are never "finished" in the sense that a curriculum is finished, and continued modification and adaptation will be necessary both at the original site and in other locations that use the program. The evaluator still has some hard thinking to do.

In practice, evaluation is most often called on to help with decisions about improving programs. Go/no-go, live-or-die decisions are relatively rare. Even when evaluation results show the program to be a failure, the usual reaction is to patch it up and try again. Rare, too, is the use of evaluation in theory-oriented tests of program approaches and models. These are more readily studied under controlled laboratory conditions. It is the search for improvements in strategies and techniques that supports much evaluation activity at present.

Even when decision makers start out with global questions (Is the program worth continuing?), they often end up receiving qualified results ("There are these good effects, but . . .") that lead them to look for ways to modify present practice. They become interested in the likelihood of improved results with different components, a different mix of services, different client groups, different staffing patterns, different organizational structure, different procedures and mechanics. One of the ironies of evaluation practice is that it has performed well at assessment of overall impact, suited to the uncommon go/no-go decision; it is relatively undeveloped in designs that produce information on the effectiveness of comparative strategies. We shall return to this point in Chapter 4.

[10] Michael Scriven, "The Methodology of Evaluation," in *Perspectives of Curriculum Evaluation*, ed. Ralph W. Tyler, Robert M. Gagné, and Michael Scriven, AERA Monograph Series on Curriculum Evaluation, No. 1 (Chicago: Rand McNally & Co., 1967), pp. 39–83.

[11] See also Thomas J. Hastings, "Curriculum Evaluation: The Why of Outcomes," *Journal of Educational Measurement*, III, No. 3 (1966), 27–32.

Whose Use Shall Be Served?

Some possible users of the evaluation have been mentioned:

1. A funding organization (government, private, foundation)
2. A national agency (governmental, private)
3. A local agency
4. The directors of the specific project
5. Direct-service staff
6. Clients of the program
7. Scholars in the disciplines and professions

Which purposes shall the evaluation serve and for whom? In some cases, the question is academic. The evaluator is on the staff of some organization—national organization, pilot program—and he does the job assigned to him. But more often, the evaluator has a number of options open. If he is on the staff of an outside research organization that is being asked to undertake the evaluation, he may have the opportunity to negotiate the purpose and focus of the study. Even if he is more closely attached to the project, there is commonly such an amazing lack of clarity among the other parties that he has wide room to maneuver.

If he can help shape the basic focus of the study, the evaluator will consider a number of things. First is probably his own set of values. A summer program for ghetto youth can be evaluated for city officialdom to see if it cools out the kids and prevents riots and looting. The evaluator may want to view the program from the youths' perspective as well and see if it has improved their job prospects, work skills, and enjoyment. The data such a study produces can give a wider frame of reference to the decision of whether or not to continue the summer programs. It is important that the evaluator be able to live with the study, its uses, and his conscience at the same time.

Beyond this point, the paramount consideration in what use the study should be designed to serve is: What decision has to be made? The pending question may be one of extending a small pilot program in one hospital ward to other wards in the same hospital. It may be allocating money to one project or to another. There may have to be a decision on the adoption of one technique (reduced case loads, nonprofessional aides) throughout the system. Perhaps the upcoming decisions have to do with staffing, structure, or target populations. Once the evaluator finds out what key decisions

are pending and when they will come up, he can gear his study to provide the maximum payoff.

Often there is no critical decision pending, at least that anyone can identify at the moment. There are, however, "users" who are interested in learning from the study and applying the results and others who are not. When the local program managers are conscientiously seeking better ways to serve their clients while the policy makers at higher levels are looking primarily for "program vindicators," the local managers' questions may deserve more attention. On the other hand, if the locals want a whitewash and the higher levels want to know where to put further appropriations, the evaluator should place more emphasis on comparative assessment of overall outcome.

The next task, then, is designing the evaluation to provide the answers that are needed. Finding out what answers are needed is not always an easy job. As we shall see in Chapter 3, it is the rare program that is articulate about goals, objectives, criteria, and bases for decision. Nevertheless, based on his best estimate of intended use, the evaluator has to make decisions on the measures to be used (see Chapter 3), sources of information (Chapter 3), and research design (Chapter 4). He will be abetted or hindered by the location of the evaluation within the organizational structure. It is to this issue that we now turn.

Structure of the Evaluation

An evaluation study can be staffed and structured in different ways. A research unit or department within the program agency can do the evaluation, or special evaluators can be hired and attached to the program. (This is often the way federally funded demonstration projects handle their evaluation requirement.) Outsiders, usually university faculty members, are sometimes paid to serve as consultants, and either advise the evaluators on staff or carry out some of the evaluation tasks themselves in close cooperation with staff. These kinds of arrangements can be lumped together as "in-house."

Another approach is for the agency to contract with an outside research organization to do the study. The research organization, whether it is an academic group, a nonprofit organization, or a commercial firm, is responsible to the persons (and the level in the program agency) who commission it. Still another kind of arrangement is for a national agency (such as the U.S. Office of Education or the national YMCA) to employ a research organization to study a number of the local programs it supports or oversees.

Inside vs. outside evaluation

There is a long tradition of controversy, mainly oral, about whether in-house or outside evaluations are preferable.[12] The answer seems to be that neither has a monopoly on the advantages. Some of the factors to be considered are administrative confidence, objectivity, understanding of the program, potential for utilization, and autonomy.

Administrative confidence. Administrators must have confidence in the professional skills of the evaluation staff. Sometimes agency personnel are impressed only by the credentials and reputations of academic researchers and assume that the research people it has on staff or can hire are second-raters. Conversely, it may view outside evaluators as too remote from the realities, too ivory-tower and abstract, to produce information of practical value. Occasionally, it is important to ensure public confidence by engaging evaluators who have no stake in the program to be studied. Competence, of course, is a big factor in ensuring confidence and deserves priority consideration.

Objectivity. Objectivity requires that evaluators be insulated from any possibility of biasing their data or its interpretation by a desire to make things look good. Points usually go to outsiders on this score, although fine evaluation has been done by staff evaluators of scrupulous integrity. It even happens that an outside research firm will sweeten the interpretation of program results (by choice of respondents, by types of statistical tests applied) in order to ingratiate itself with a program and get further contracts. In any event, safeguarding the study against even unintentional bias is important.

Understanding of the program. Knowledge of what is going on in the program is vital for an evaluation staff. They need to know both the real issues facing the agency and the real events that are taking place in the program if their evaluation is to be relevant. It is here that in-house staffs chalk up points, although outsiders too can find out about program proc-

[12] See Elmer Luchterhand, "Research and the Dilemmas in Developing Social Programs," in *The Uses of Sociology*, ed. P. F. Lazarsfeld, W. H. Sewell, and H. L. Wilensky (New York: Basic Books, Inc., Publishers, 1967), pp. 513–17; Rensis Likert and Ronald Lippitt, "The Utilization of Social Science," in *Research Methods in the Behavioral Sciences*, ed. Leon Festinger and Daniel Katz (New York: Holt, Rinehart & Winston, Inc., 1953), pp. 581–646; Martin Weinberger, "Evaluating Educational Programs: Observations by a Market Researcher," *Urban Review*, III, No. 4 (1969), 23–26.

esses if they make the effort and are given access to sources of information.

Potential for utilization. Utilization of results often requires that evaluators take an active role in moving from research data to interpretation of the results in a policy context. In-house staff, who are willing to make recommendations on the basis of results and advocate them in agency meetings and conferences, may be better able to secure them a hearing. But sometimes it is outsiders, with their prestige and authority, who are able to induce the agency to pay attention to the evaluation.

Autonomy. Insiders generally take the program's basic assumptions and organizational arrangements as given and conduct their evaluation within the existing framework. The outsider may be able to exercise more autonomy and take a wider perspective. While respecting the formulation of issues set by the program, he may be able to introduce alternatives that are a marked departure from the status quo. The implications he draws from evaluation data may be oriented less to tinkering and more to fundamental restructuring of the program.[13] However, such a broader approach is neither common among outsiders nor unknown among insiders.

All these considerations have to be balanced against each other. There is no one "best site" for evaluation. The agency must weigh the factors afresh in each case and make an estimate of the way which the benefits pile up.

Level in the structure

Whoever actually does the evaluation, the evaluation staff fits somewhere in the organizational bureaucracy. The evaluator reports to a person at some level of authority in the program organization or its supervisory or funding body, and he is responsible to that person and that position for the work he does. If the evaluator is an insider, he reports on a regular basis. The outsider researcher also receives his assignment and reports his results to (and may get intermediate advice from) the holder of a particular organizational position.

The important distinction in organizational location for our discussion is the difference between the policy maker and the program manager. To abridge our earlier catalog of users of evaluation and the decisions they have to make, the key points are these:

[13] Robert K. Merton, "Role of the Intellectual in Public Bureaucracy," in *Social Theory and Social Structure* (New York: The Free Press, 1964), pp. 207–24.

User	*Decision*
Policy maker	Whether to expand, contract, or change the program
Program manager	Which methods, structures, techniques, or staff patterns to use

The evaluation should be placed within the organizational structure at a level consonant with its mission. If it is directed at answering the policy questions (How good is the program overall?), evaluators should report to policy makers. If the basic shape of the program is unquestioned and the evaluation issue centers on variations in specific features, the evaluator should probably be responsible to the program managers.[14]

Real problems arise when the evaluation is inappropriately located in the structure. An evaluation that is initiated by and responsible to program managers is under all kinds of pressure not to come up with findings that disparage the effectiveness of the whole program. If it does, the managers are likely to stall the report at the program level and it will never receive consideration in higher councils.[15] On the other other hand, when top policy makers initiate and oversee the evaluation, their questions are paramount, and questions about operations may get the short end of the budget. Nor do the evaluators have the easy, informal contact with program managers and practitioners that allows them to hear and understand the problems and options they face. It sometimes becomes difficult to study the effectiveness of different program components because staff see the evaluators as "inspectors" checking up on them and become vary of divulging information that might reflect poorly on their performance. Nor are they always cooperative in maintaining the conditions necessary for evaluation research, particularly if there is competition among program levels and the evaluation is viewed as an effort to assert the priorities of the higher level.

The problem of structural location becomes more complex when the evaluation is serving both masters. By and large, it appears best to report in at the higher level. In that way, the evaluator maintains greater autonomy. But then he has to make special efforts to learn enough about critical issues in day-to-day program operations to incorporate them into the study and to maintain the support of local program managers for appropriate research conditions.

[14] This rule of thumb applies whether the evaluation is performed by an in-house evaluation unit or by an outside research organization. Either one should report in at the level of decision to which its work is addressed. The outsiders probably have greater latitude in going around the organizational chain of command and finding access to an appropriate ear, but even they will be circumscribed by improper location.

[15] This point is discussed in Likert and Lippitt, *op. cit.*

Good placement in the structure is important. A recent report of Wholey et al. on federal evaluation practice [16] discusses this issue in terms of federal agencies' responsibilities. It recommends that a central evaluation staff in each agency should have responsibility for planning and coordinating all evaluation work in the department, but that staff at different levels should be responsible for direct supervision of evaluation studies depending on their scope and purpose.

> Policy makers are most often called upon to make choices among national programs; program managers are most often called upon to make choices of emphasis or decisions on the future of individual projects *within* national programs. To the extent possible, program impact evaluations, designed to discover the worth of an entire national program, should be directed by persons not immediately involved in management of the program and operation. Program strategy evaluation should be directed by persons close enough to the program to introduce variations into the program.[17]

Wherever the evaluation project sits in the structure, it should have the autonomy that all research requires to report objectively on the evidence and to pursue issues, criteria, and analysis beyond the limits set by the program in order to better understand and interpret the phenomena under study.

[16] Joseph S. Wholey et al., *Federal Evaluation Policy* (Washington, D.C.: The Urban Institute, 1970), pp. 54–71.

[17] *Ibid.*, p. 65.

3

Formulating the Question and Measuring the Answer

The traditional formulation of the evaluation question is: To what extent is the program succeeding in reaching its goals? Variations are possible: Is program *A* doing better than program *B* in reaching their common goals? How well is the program achieving results *X*, *Y*, and *Z* with groups *F*, *G*, and *H*? Which components of the program (*R*, *S*, or *T*) are having more success? But the basic notion is the same. There are goals; there is a planned activity (or several planned activities) aimed at achieving those goals; there is a measure made of the extent to which the goals are achieved. In evaluation there is also the expectation that controls are set up so that the researcher can tell whether it was the *program* that led to the achievement of goals rather than any outside factors (such as the maturing of the participants, improvement in the economy, and so on). The issue of study design—how controls can be instituted in research on an action program—is the subject of the next chapter.

The evaluation question sounds simple enough in the abstract. All the researcher has to do, it seems, is:

1. Find out the program's goals.

2. Translate the goals into measurable indicators of goal achievement.
3. Collect data on the indicators for those who participated in the program (and for an equivalent control group who did not).
4. Compare the data on participants (and controls) with the goal criteria.

And voilà!

But what looks elementary in theory turns out in practice to be a demanding enterprise. Programs are nowhere near as neat and accommodating as the evaluator expects. Nor are outside circumstances as passive and unimportant as he might like. Whole platoons of unexpected problems spring up. This chapter deals with four:

1. Program goals are often hazy, ambiguous, hard to pin down. Occasionally, the official goals are merely a long list of pious and partly incompatible platitudes.
2. Programs not only move toward official goals. They accomplish other things, sometimes in addition and sometimes instead. The evaluator has a responsibility to take a look at these unexpected consequences of program activities.
3. The program is a congeries of activities, people, and structures. Some of its elements are necessary for the effects it achieves; others are irrelevant baggage. Decision makers want to know what the basic and essential features of the program are, so that (if successful) they can reproduce them or (if unsuccessful) avoid them. How do you identify and separate out the elements that matter?
4. The evaluation question as posed ignores the issue of why the program succeeds or fails. The *why* is often just as important to know as *how well* the program works.

In addressing these issues, we will recommend a series of strategies. Possibly the most important theme (and we return to it in the next chapter when we discuss design) is the classification of the component parts of the program. Each element (of activity, approach, structure, participant, and so on) that is presumed likely to affect outcomes is observed, defined, and classified. The differences that evolve between groups, between activities, and so on give increasing information about what works and does not work in reaching program goals.

In this chapter, then, we consider these core issues:

1. Formulating the program goals that the evaluation will use as criteria
2. Choosing among multiple goals
3. Investigating unanticipated consequences
4. Measuring outcomes

5. Specifying what the program is
6. Measuring program inputs and intervening processes
7. Collecting the necessary data

FORMULATING PROGRAM GOALS

It is a common experience for an evaluator to be called in to study the effects of a program and not be told its purpose. If he presses for a statement of goals, program administrators may answer in terms of the number of people they intend to serve, the kinds of service they will offer, the types of staff they will have, and similar information. For program implementers, these are "program goals" in a real and valid sense, but they are not the primary currency in which the evaluator deals. He is interested in the intended *consequences* of the program. When he pursues the question, "What is the program trying to accomplish?" many program people give fuzzy replies, often global and unrealistic in scope. They may hazard the statement that they are trying to "improve education," "enhance the quality of life," "reduce crime," "strengthen democratic processes." Thus begins the long, often painful, process of getting people to state goals in terms that are *clear, specific,* and *measurable*.

The goal must be clear so that the evaluator knows what to look for. In a classroom program, should he look for evidence of enjoyment of the class? interest in the subject matter? knowledge of the subject matter? use of the subject matter in further problem solving?

The goal has to be specific. It must be able to be translated into operational terms and made visible. Somebody has to *do* something differently when the goal is reached. Thus, if the goal is to interest students in new materials, they are likely to talk more often in class, or raise their hands more often, or do more outside reading on the subject, or tell their parents about it, or any of several other things.

For evaluation purposes, the goal has to be measurable. This is not as serious a restriction as it may seem at first glance. Once goal statements are clear and unambiguous, skilled researchers can measure all manner of things. They can use the whole arsenal of research techniques—observation, content analysis of documents, testing, search of existing records, interviews, questionnaires, sociometric choices, laboratory experiments, game playing, physical examinations, measurement of physical evidence, and so on. With attitude tests and opinion polls, they can measure even such relatively "soft" goals as improvements in self-esteem or self-reliance. But since few programs set out only to change attitudes, the evaluator will also want to find and measure the behavioral consequences of changed at-

titudes—the things participants do because they feel different about themselves, other people, or the situation.

Some programs find it extremely difficult to formulate goals in these terms. David Kallen tells of working with an advisory committee to plan for evaluation of a detached worker program for gang youth. Asked to specify the program's goals, the committee members came up with such things as improving the behavior of the youth, helping them become better citizens, and improving their school work. When they tried to translate the goals into operational criteria of program success, "behavior" and "citizenship" were too vague to use, and school grades were too likely to be influenced by teachers' stereotyped perceptions of the youngsters. The discouraging story continues:

> Finally, it turned out that a number of the area residents objected to the young people's use of swear words, and it was decided that one measure of behavioral improvement would be the reduction in swearing, and that this was something the detached worker should aim for in his interaction with the youngsters he was working with. [Was the group identifying program goals or making up new ones?] It was therefore agreed that part of the criteria of success would be a reduction in swearing. I might add that this was the only measure of success upon which the evaluation team and the program advisory committee could agree.[1]

Fuzziness of program goals is a common enough phenomenon to warrant attention. Part of the explanation probably lies in practitioners' concentration on concrete matters of program functioning and their pragmatic mode of operation. They often have an intuitive rather than an analytic approach to program development. But there is also a sense in which ambiguity serves a useful function: It may mask underlying divergences in intent. Support from many quarters is required to get a program off the ground, and the glittering generalities that pass for goal statements are meant to satisfy a variety of interests and perspectives.

However, when there is little consensus on what a program is trying to do, the staff may be working at cross-purposes. One side benefit of evaluation is to focus attention on the formulation of goals in terms of the specific behaviors that program practitioners aim to achieve. The effort may force disagreements into the open and lead to conflict. But if differences can be reconciled (and the program may not be viable if they are not), the clarification can hardly help but rationalize program implementation. It may reveal discrepancies between program goals and program

[1] Personal letter from David J. Kallen, January 10, 1966.

content, in which case either the content or, as Berlak notes,[2] the goal statement should be changed. When a sense of common purpose is reached, the logic and rationality of practice are likely to be enhanced.

What does an evaluator do when he is faced with a program that cannot agree on a statement of specific and meaningful goals? Four courses are open to him:

1. He can pose the question and wait for program personnel to reach a consensus. But as Freeman and Sherwood[3] note, he should bring books to the office to read while waiting for them to agree. And they still may not develop a statement that provides an adequate basis for evaluation.
2. Another thing he can do is read everything about the program he can find, talk to practitioners at length, observe the program in operation, and then sit down and frame the statement of goals himself. Sometimes this is a reasonable procedure, but there are two dangers. One is that he may read his own professional preconceptions into the program and subtly shift the goals (and the ensuing study) in the direction of his own interests. The other risk is that when the study is completed, the program practitioners will dismiss the results with the comment, "But that's not really what we were trying to do at all."
3. He can set up a collaborative effort in goal formulation. This is probably the best approach. Sitting with the program people, the evaluator can offer successive approximations of goal statements. The program staff modifies them, and the discussion continues until agreement is reached.
4. He can table the question of goals, and enter not upon evaluation in the traditional sense, but on a more exploratory, open-ended study. In complex and uncharted areas, this may be a better strategy than formulating arbitrary and superficial "goals" in order to get on with the study while the really significant happenings around the program are allowed to take place unstudied, unanalyzed, and unsung. Evaluations based on too-specific goals and indicators of success may be premature in a field in which there is little agreement on what constitutes success.[4]

[2] Harold Berlak, "Values, Goals, Public Policy and Educational Evaluation," *Review of Educational Research*, XL, No. 2 (1970), 261–78.

[3] Howard E. Freeman and Clarence C. Sherwood, "Research in Large-scale Intervention Programs," *Journal of Social Issues*, XXI, No. 1 (1965), 11–28.

[4] See Cyril S. Belshaw, "Evaluation of Technical Assistance as a Contribution to Development," *International Development Review*, VIII (1966), 2–6, 23, for a situation in which this was the case. He goes on, however, to recommend a theoretical framework and a series of possible criteria of success for technical assistance programs, such as an increase in the range of commodities produced or increased division of labor. He offers an approximation of goal statements that can be progressively modified by other researchers, operators, and scholars.

The experienced evaluator also searches for the hidden agenda, the covert goals of the project that are unlikely to be articulated, but whose achievement sometimes determines success or failure no matter what else happens. For example, if a program of interdisciplinary studies in a university fails to win the support of the departmental faculties and the university administration, even consummate educational results may not be enough to keep it alive. The evaluator, if he is to study the attainment of goals, is well advised to keep an eye on the "system" goals (those that help maintain the viability of the program in its environment) as well as the "outcome" goals. He will learn much that explains why the program makes the adaptations it does and where the real game is.[5]

Some researchers have even proposed that the goal model of evaluation should be junked in favor of a system model.[6] The elements of such a model are not yet clear; there are almost as many interpretations as there are participants in the discussion. But the common recognition is that organizations pursue other functions besides the achievement of official goals. They have to acquire resources, coordinate subunits, and adapt to the environment. These preoccupations get entangled with, and set limits to, attainment of program goals. According to system model proponents, an evaluation that ignores them is likely to result in artificial and perhaps misleading conclusions.

What would a system model look like? Etzioni, and Schulberg and Baker suggest that the system model should be based on the evaluator's extensive knowledge of the organization and his understanding of the optimal allocation of resources among organization-maintenance and goal-achievement functions. The key question then becomes: "Under the given conditions, how close does the organizational allocation of resources approach an optimum distribution?"[7] Provocative as the notion is, it sets

[5] Andrew C. Fleck, Jr. "Evaluation Research Programs in Public Health Practice," *Annals of the New York Academy of Science*, CVII, No 2 (1963), 717-24, recommends that evaluators have intimate knowledge of the organization and its relative emphasis on short-run stability versus long-run survival

[6] See Edward A. Suchman, "Action for What? A Critique of Evaluative Research," in *The Organization, Management, and Tactics of Social Research*, ed. Richard O'Toole (Cambridge, Mass · Schenkman Publishing Co , 1970); Amitai Etzioni, "Two Approaches to Organizational Analysis: A Critique and a Suggestion," *Administrative Science Quarterly*, V, No. 2 (1960), 257–78, Herbert C Schulberg and Frank Baker, "Program Evaluation Models and the Implementation of Research Findings," *American Journal of Public Health*, LVIII, No 7 (1968), 1248-55; Perry Levinson, "Evaluation of Social Welfare Programs Two Research Models," *Welfare in Review*, IV, No. 10 (1966), 5–12, Herbert C. Schulberg, Alan Sheldon, and Frank Baker, "Introduction" in *Program Evaluation in the Health Fields* (New York Behavioral Publications Inc , 1970).

[7] Etzioni, *op cit* , p. 262.

such demanding requirements for the evaluator (knowing more about the organization than the organization knows itself) that it is difficult to imagine its practical application, at least in these terms. Perhaps future development will bring its genuine insights into the realm of practicality. For the time being, most evaluators will probably stick with the goal model, which is certainly justifiable on its own grounds, and give as much attention to the organizational and community systems that affect the program as the situation seems to warrant.

Choices Among Goals

Once the goals of the project are clearly, specifically, and behaviorally defined, the next step is to decide which of them to evaluate. How does the evaluator make the decision?

Usability and practicality

Part of the answer lies in the potential for utilization. How will the evaluation findings be applied, and which goals are relevant to that decision? Part of the answer lies in the hard realities of time, money, and access. How far off in time the evaluator can study is limited by how long the project—and the evaluation—last; how much he can study is at least partly a function of money; whether he can examine certain classes of effects depends on whether he is permitted access to people and agencies. A tendency endemic in all kinds of research is to study what is easy to study rather than what ought to be studied. It is particularly important for the evaluator to avoid this kind of cop-out and to concentrate on key concerns of the program.

Relative importance

There remains still another factor—the relative importance of different goals. This requires value judgment, and the program's own priorities are critical. The evaluator will have to press to find out priorities—which goals the staff sees as critical to its mission and which are subsidiary. But since the evaluator is not a mere technician for the translation of a program's stated aims into measurement instruments, he has a responsibility to express his own interpretation of the relative importance of goals. He doesn't want

to do an elaborate study on the attainment of minor and innocuous goals, while vital goals go unexplored.[8]

Incompatibilities

In some cases there are incompatibilities among stated goals. A model cities program, for example, seeks to increase coordination among the public and private agencies serving its run-down neighborhood. It also desires innovation, the contrivance of unusual new approaches to services for the poor residents. Clearly, coordination among agencies will be easier around old, established, accepted patterns of service than around new ones. Innovation is likely to weaken coordination, and coordination is likely to dampen the innovating spirit. Which goal is more "real"? Evaluation cannot stick its head in the sand and treat the two goals as equal and independent.

Short-term or long-term goals?

Another issue is whether short- or long-term goals are more important. Decision makers, who by professional habit respond to the demands of the budget cycle rather than the research cycle, usually want quick answers. If they have to make a decision in time for next year's budget, there is little value in inquiring into the durability of effects over 24 months. It is this year's results that count.

But decision makers can often be persuaded to see the utility of continuing an investigation over several years, so that the program's long-term effectiveness becomes manifest. Clearly, it is good to know whether early changes persist, or on the other hand, whether the absence of early change reflects a "sleeper effect," the slow building up of important changes over time. Evaluations, wherever possible, should look into long-term effects, particularly when basic policies or costly facilities are at stake. A comparison of short- *and* long-term effects provides additional information about how, and at what pace, effects take place.

The evaluator is well-advised to thrash out the final selection of goals for study with decision makers and program managers. They are all involved. It is he who will have to live with the study and they who will have to live with the study results and—one would hope—their implementation.

[8] Robert E. Stake discusses the evaluator's responsibility for evaluating proffered goals. "The Countenance of Educational Evaluation," *Teachers College Record*, LXVIII, No. 7 (1967), 523–40.

Yardsticks

Once the goals are set, the next question is how much progress toward the goal marks success. Suppose a vocational program enrolls 400, graduates 200, places 100 on jobs, of whom 50 are still working three months later. Is this success? Would 100 be success? 200? 25? Without direction on this issue, interpreters can alibi any set of data. A tiny change is better than no change at all. No change is better than (expected) retrogression. Different people looking at the same data can come up with different conclusions in the tradition of the "fully-only" school of analysis. "Fully 25 percent of the students . . ." boasts the promoter; "only 25 percent of the students . . ." sighs the detractor.

Only on a comparative basis does the question really make sense. How do the results compare with last year's results, with the results for those who did not get the special program, or better still, with the results from programs with similar intent?[9] If comparable data are not available, the evaluator can present his results and let others draw their own conclusions. Or he can get into the act by drawing on past experience, the opinions of administrators and staff, and perhaps outside experts, in reaching a judgment of his own.[10] Early attention to standards of judgment—before the data come in—can forestall later wrangling.

Unanticipated Consequences

The program has desired goals. There is also the possibility that it will have consequences that it did not intend. The discussion of unanticipated results usually carries the gloomy connotation of undesirable results, but there can also be unexpected good results and some that are a mixture of good and bad.

Undesirable effects can come about for a variety of reasons. Sometimes the program is poorly conceived and exacerbates the very conditions it

[9] This of course limits the question rather than settles it. How *much* better must the program be before it is considered a success? Statistically significant differences do not necessarily mean substantive significance. Perhaps cost-benefit analysis brings the wisest question to bear: How much does it cost for each given amount of improvement? Carol H. Weiss, "Planning an Action Project Evaluation," in *Learning in Action*, ed. June L. Shmelzer (Washington, D.C.: Government Printing Office, 1966), pp. 15–16.

[10] Stake, *op. cit.*, pp. 527, 536–38, suggests comparisons with absolute standards, with other programs, and with the opinions of experts for judgment of success.

aimed to alleviate. A loan program to inefficient small businessmen may only get them deeper into debt. Or a program can boomerang by bringing to light woes that have long been accepted. Some programs raise people's expectations. If progress is too slow or if only a few people benefit, the results may be widespread frustration and bitterness. Occasionally, a program that invades the territory of existing agencies generates anger, competition, and a bureaucratic wrangle that lowers the effectiveness of services.

Good unanticipated consequences are not so usual, because reformers trying to sell a new program are likely to have listed and exhausted all the positive results possible. Nevertheless, there are occasions when a program has a happy spin-off, such as having its successful features taken over by a program in a different field. There can be spillovers of good program results to other aspects of a program participant's life. For example, pupils who learn reading skills may become more cooperative and less disruptive or aggressive in school and at home. Contagion effects appear, too. People who never attended the program learn the new ideas or behaviors through contact with those who did.

Sometimes programs tackle one aspect of a complex problem. Even if they achieve good results in their area, the more important effect may be to throw the original system out of kilter. Thus an assistance program to underdeveloped areas introduces a new strain of rice that increases crop yield—the goal of the program. But at the same time, the effect is to make the rich farmers richer (because they can afford the new seed and fertilizer and can afford to take risks), widen the gulf between them and the subsistence farmers, and lead to social and political unrest. Fragmented programs all too often fail to take into account interrelationships between program efforts and the overall system in which people function. What are originally conceived as good results in one sphere may be dysfunctional in the longer view. It is because of such complex interlinkages that the notion of a systems approach to evalution is appealing.

The evaluator has to keep an eye on the "other" consequences of the program he is studying. Although decision makers have not articulated them as goals, he must unearth and study consequences that have significant impact on people and systems. Like the formulation of goals, this exercise requires thought and attention. A wise evaluator brainstorms in advance about all the effects, good, bad, and indifferent, that could flow from the program. Envisioning the worst as well as the best of all possible worlds, he makes plans for keeping tabs on the range of likely outcomes. What were "unanticipated consequences" are now—if he judged well—unintended but anticipated. He also has to remain flexible and open enough to spot the emergence of effects that even his sweeping imagination had not envisioned.

If he or his evaluation staff is close enough to the scene to observe what goes on, informal observation may be sufficient for the first look at unplanned effects. In more remote or complex situations, he will have to develop measures and data-gathering instruments to pull in the requisite information. Once trends become clear and side effects are seen to be a strong possibility, he will want as precise measures as he can devise of what may become the most important elements in the program field. He never wants to be caught saying, "The program (on our outcome measures) was a success, but the patient died."