advice on using the methods together. Chapter 12 offers guidance for analyzing and interpreting data. It applies the same framework to quantitative and qualitative data, showing the logic of the analytic task.

Chapter 13 focuses on reporting results and disseminating them to relevant audiences. It suggests techniques for reporting and gives tips on satisfying different publics. It puts considerable emphasis on communicating and disseminating results, and it stresses the evaluator’s responsibility for encouraging the use of evaluation results and suggests ways to promote their application to issues of program and policy. Chapter 14 sums up the major themes in the book in what I hope is a rousing finale.

Acknowledgments

It is hard to thank all the people whose experience, ideas, encouragement, and criticisms have entered into my knowledge of evaluation. I would have to reproduce the international catalog of evaluation books and articles (at least the ones I’ve read), list the authors of hundreds of evaluation reports in at least a dozen countries, and name the students in my evaluation courses who shared their interests and puzzlements over many years. I limit myself here to the people who have helped me directly, but I am grateful to hundreds of others.

I had the wonderful fortune to begin writing this book when I was a Fellow at the Center for Advanced Study in the Behavioral Sciences, a place awash in fertile minds and helpful colleagueship. I finished it back at Harvard, an environment with its own brand of nutrients. Two people had the kindness and fortitude to read the whole manuscript and give excellent critiques: Will Shadish and Janet Weiss, and I am greatly indebted to them. They saw the hidden text as well as the written word, and forced me to clarify what I intended to say. John Willett read the four chapters on design and analysis with great care, taught me, argued with me, and pointed me in the right direction. Robert T. Bremmen scrutinized the chapters on measurement and data collection and made detailed suggestions to make them more accurate and complete. Other colleagues who read chapters of the manuscript and helped me improve the text were Joseph Maxwell, Mercer Sullivan, Harris Cooper, and Sharon Rallis. Lee Cronbach, David Fetterman, and Judith Singer offered references, advice, and insight.

Jay Braatz was a great help with the Glossary and the charts. I thank Stuart Yeh, Chris Mazzeo, and Mary Jo Larson for help with library references. I thank Mark Anderson-Wilk and Kathy George for typing drafts of the manuscript.

I am especially grateful to Johanna Birkmayer. Jo has been a steady source of information and advice. She read successive drafts of the manuscript and gave me the benefit of her criticism, sound judgment, and unflagging encouragement over a period of many months.

Whatever flaws remain in the book are the responsibility of this galaxy of advisors. (Not true, but it was fun to say.)

Carol Hirschon Weiss

"Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."

—Lewis Carroll (1865)

The emerging federal system for social science research reflects recognition of the need for planning and for the close relationship of knowledge to the formulation and execution of social programs.

—Gene Lyons (1969, p. 240)

Instead of trivial tokens or new Federal programs launched at full scale, [the President] could initiate a set of bold research trials.... Though large-scale problems would not be solved immediately, pilot tests would proliferate, and through evaluation of them, learning would grow.


What does an evaluation study look like? Let's start with an easy example, a program to help people stop smoking. One such program enrolls about 100 people each month and gives them three sessions of group activities. After the program has been operating for about a year, staff members ask themselves how well they are doing. They don't see the people after the program ends, and they have no way of knowing if they are successful. Do the people that they serve actually stop smoking and stay off cigarettes?

Staff decide to do an evaluation. Their aim is to find out whether the goals of the program are met: Do participants stop smoking? This is a relatively simple evaluation because there is one clear indicator of success. If a person stops smoking, that's success. You don't have to look any further, because biomedical research has conclusively shown that smoking cessation improves people's health and longevity. The staff go through the lists of people who attended the session that ended the week
before and call them up. They ask: "Have you stopped smoking?" Answers show that 47% say yes, they have stopped.

But several of the staff say that one week is too short a time interval. It's not so hard to quit for a week. The real question is whether participants stay off cigarettes. So they pull out the names of those who attended sessions during the first three months of the program, almost a year before, and call them up. They ask: "Have you stopped smoking?" The results show that 39% say yes, they have stopped. The staff look at one another in puzzlement and ask, "Is that good?" They had hoped that 100% would have stopped, but 39% is batting .390 (a terrific batting average in baseball) and that is probably 39% more than would have stopped without the program. Or is it?

One staff member reminds his colleagues that many people stop smoking without any program at all. Maybe these people would have done fine without the program. So the staff decides to call smokers they know who didn't attend the program and ask them the same question: "Have you stopped smoking?" Results show that only 2% of these people have stopped. One person says, "See? Our program did make a difference."

But another person says, "No, that's not a fair comparison. The people who came to the program were motivated to stop, and the others weren't." The problem: How do they find people to compare with their participants, people who were motivated to stop and didn't attend the program? One person suggests looking at the attendees of another stop-smoking program. "That," he says, "would be a fair comparison." But others say that such a study would only compare their program to another program, and that's not what they are really interested in. They want to know if they're doing a good job, whether they're better or worse than somebody else. Both programs might be poor or both might be wonderful, and the comparison wouldn't be able to show this.

So one staff person with research training suggests that they construct a randomized control group. They need people who are similarly motivated to stop smoking. So from the next 200 applicants to the program (people with similar levels of motivation) they will randomly accept 100 into the program and tell the others that they have no room. ("That's not unethical," she explains, because the program receives many more applications than they have room for, and instead of accepting applicants on a first-come, first-served basis, they will accept them by lottery, which is a technique of random selection.) Random selection does not mean haphazard selection; it means following very strict procedures that take advantage of the laws of probability. If each person has an equal chance of falling into the program group, the likelihood is high that people accepted into the program group and people not accepted (the control group) will be very much alike on all dimensions, such as age, length of time smoking, extent of family support for stopping, and just about everything else.

So that's what they do. Out of the next 200 applicants, they randomly take 100 into the program and tell the others that they'll have to wait for the next month's session. After a month they call both sets of people. (They can't wait any longer, because they now feel duty bound to accept the control group into the program.) Answers to the telephone interviews show that 46% of the program group say they have stopped smoking compared to 15% of the control group.

Now staff believe that they have demonstrated the unequivocal success of their program. Nonparticipants with the same level of motivation did not make it on their own.

"Wait a minute," says one of the staff. "We told them that they would be accepted into the program in a month. Why would they try to stop on their own? They're waiting for the program, enjoying their last month of nicotine. That would keep the number of stoppers low." His colleagues look at him in annoyance and say, "Why didn't you think of that before?" Another staff person says, "Or if they were really motivated, maybe they didn't want to wait and went to another program. That might push the numbers up."

Another staff person says, "There's another problem. What does it mean when someone says he has stopped smoking? Did he stop the day before we called, or has he been off cigarettes since he left the program? A lot of people probably stop and start again and then stop again. We probably should have asked a lot more questions about how much each person has smoked since leaving the program." "Yes," agrees another person. "Even if someone just reduced the number of cigarettes he smoked every day, that would be a good result, and we didn't ask questions that would find that out."

At this point staff were almost ready to chuck the whole evaluation. They were even more dispirited when the head of the agency asked: "What makes you think that people are telling you the truth? You were the people who ran the program, and some participants who are still smoking might lie to you and say they'd quit just to avoid hurting your feelings. You should check with a relative or friend or perhaps have them take a saliva test that detects nicotine."

"But we don't have the time or money for such an elaborate study," one person said. "Isn't there anything we can do now?"

By the time you have finished the book, I hope you can help them make sensible choices about what to do next. If you'd like to pursue this example, here are references to a few of the thousands of evaluations of smoking-cessation programs: Orleans et al., 1991; Ossip-Klein et al., 1991; Viswesvaran & Schmidt, 1992; Zelman, Brandon, Irenby, & Baker, 1992.

Domain of Evaluation

People evaluate all the time. Listen in on conversations and you will hear: "I loved that television program last night." "He is doing a lousy job." "That car isn't worth the price they charge." "The food at Joe's Cafe is much better now than it used to be." In more formal terms, you hear about a supervisor evaluating an employee's work performance, a teacher evaluating a student's academic progress, or Consumer Reports' evaluation of a line of products. Just as Molière's bourgeois gentilhomme finds out that he's been speaking prose all his life without realizing it, we come to realize that we have been evaluating things of various kinds without necessarily calling the procedure evaluation.

Evaluation is an elastic word that stretches to cover judgments of many kinds. 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. The yardsticks can vary widely. One criterion might be aesthetic: Is the entity beautiful and pleasing? Another yardstick is effectiveness: Does it do what it is supposed to be doing? Another is efficiency: Does it provide benefits commensurate with its costs? Other yardsticks can deal with justice and equity, acceptability in terms of community standards, enjoyment and satisfaction, contributions to social harmony, and so on.

The phenomena to be evaluated can also be diverse. In this book I will be talking about evaluation of one particular kind of phenomenon: programs and policies designed to improve the lot of people. The programs and policies are of many kinds: They can deal with education, social welfare, health, housing, mental health, legal services, corrections, economic development, highway construction, and many other fields. They can be governmental programs, run at federal, state, or local levels, or even internationally; they can be run by nonprofit organizations or for-profit firms. They can be aimed at changing people’s knowledge, attitudes, values, behaviors, the organizations in which they work, 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, I am concerned here with a specific method of evaluation—evaluation research. Rather than relying on the judgments of expert observers or periodic reports from staff, the evaluator in this scenario uses the methods of social science research to make the judging process more systematic and accurate. In its research guise, evaluation establishes clear questions for inquiry. It collects evidence systematically from a variety of people involved with the program. It sometimes translates the evidence into quantitative terms (23% of the program participants, grades of 85 or better), and sometimes it crafts the data into telling narratives. It then draws conclusions about the ways the program is run, or its near-term consequences, or its effectiveness in fulfilling the expectations of those who fund the program, manage and staff it, or participate in its activities.

In this chapter, I first offer a definition of evaluation, and then consider the types of phenomena that are usually evaluated in formal ways. Several different genres of evaluation are possible, and I discuss the difference between evaluation of program operations and the evaluation of program outcomes. Next I consider the differences and similarities between evaluation and other kinds of social science research. A subsequent section explores the history of evaluation.

### Definition of Evaluation

At this point I can venture a provisional definition of evaluation, which we will elaborate as we make our way through the book. Evaluation is the systematic assessment of the operation and/or the outcomes of a program or policy, compared to a set of explicit or implicit standards, as a means of contributing to the improvement of the program or policy.

Let’s take a look at the five key elements in that definition. The first is systematic assessment. The emphasis on system indicates the research nature of evaluation procedures. Whether the research is quantitative or qualitative, it is conducted with formality and rigor, according to accepted social science research canons. The second and third elements in the definition point to the focus of investigation: the operation and outcomes of the program. Some evaluations concentrate on studying process—that is, the way a program is conducted. Such evaluations may be interested in learning the extent to which the program is following prescribed practices (i.e., the fidelity of the program to its design), or they may aim just to learn what is going on. Some evaluations concentrate on the outcomes and effects of the program for its intended beneficiaries. Such evaluations seek to answer the questions: Are participants gaining the benefits that they were intended to receive? Or, in more open-ended fashion, what is happening to them because of the program’s intervention? Many evaluations look at both the process of the program and its outcomes for recipients.

The fourth element in the definition is standards for comparison. Once having collected evidence on process and outcomes, the evaluation assesses the merit of the program by comparing the evidence to some set of expectations. Whether the evaluation is focusing on the program’s process or outcomes, an element of judgment is present. Sometimes the criterion that is applied to make judgments comes from the official statement of goals set for the program or policy when it was enacted. If the program was established to reduce adolescent pregnancies, an explicit standard is likely to be reduction in the frequency with which adolescent program participants become pregnant. But official goals are not the only possible source for the criteria that an evaluation applies. The goals may have changed over the course of program operations, and what the program is now concentrating on may be teaching pregnant adolescents about proper prenatal health practices. If that is where program energies are being spent, then participating adolescents’ health practices and the birth weight of their babies may be important criteria. Other standards of judgment can come from the expectations of other actors on the program or policy scene. The objectives of program sponsors, program managers, practitioners, and participants can become criteria for the evaluation. In some cases, evaluation offers systematic evidence about the program in specified domains and leaves it to the reader to draw conclusions about its merit.

The fifth element in the definition of evaluation is the purpose for which it is done: contribution to the improvement of program and policy. Evaluation is a practical craft, designed to help make programs work better and to allocate resources to better programs. Evaluators expect people in authority to use evaluation results to take wise action. They take satisfaction from the chance to contribute to social betterment.

Doing evaluation through a process of research takes more time and costs more money than offhand evaluations that rely on intuition, opinion, or trained sensitivity, but it provides a rigor that is missing in these more informal activities. Rigor is apt to be particularly important when (a) the outcomes to be evaluated are complex, hard to observe, made up of many elements reacting in diverse ways; (b) the decisions that will follow are important and expensive; and (c) evidence is needed to convince other people about the validity of the conclusions.

A wide variety of social programs are operated in the United States and other countries, and new needs continually spark the development of further programs. The past decade has seen programs to feed and shelter the homeless, care for AIDS patients, educate young people about the dangers of drugs, reduce the costs of health care, and help victims of disaster around the world take control over their own lives.
Some of the new programs are logical extensions of earlier efforts; some represent radical departures from the past and a plunge into uncharted waters.

Many people want (and need) to know: How is the program being conducted? What is it actually doing? How well is it following the guidelines that were originally set? What kinds of outcomes is it producing? How well is it meeting the purposes for which it was established? Is it worth the money it costs? Should it be continued, expanded, cut back, changed, or abandoned? Should people flock to it, sign up selectively, or stay away in droves? Does it work for everybody or only some kinds of people?

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 people who have a say (legislators, boards of directors, future clients) are sufficiently removed from the program to want cogent information to help them make choices.

When programs are run by private firms, the market generally supplies the judgment. A private firm can supply a computer program or a training course. The evaluation comes when people either buy the product or service or do not. In time, good programs prosper and bad programs go out of business, or at least they do if consumers have sufficient information about their effects. But when the program is run by government or nonprofit agencies, customer satisfaction or dissatisfaction usually has little impact. Sometimes the program is the only game in town, like welfare assistance or the criminal courts and correctional facilities. Sometimes the program is the only one for a particular group of people, like training programs for members of the armed forces, or the only one that is free, like the public schools. Usually these kinds of programs continue no matter what the demand or the level of client satisfaction.

Private firms can run programs of this type, too, for their own employees. Internal programs, like day care for children of employees or safety procedures for workers, also have a captive population and tend to be relatively immune to client satisfaction or dissatisfaction.

Yet they are all obviously costly investments. Governments, businesses, nonprofit agencies, and foundations pay substantial sums to keep them going. A lot of people care whether the programs are doing what they are supposed to be doing and getting the kinds of results they are expected to get. A lot of people worry about whether they could be getting better results or spending less money, or better yet, both. Without some kind of systematic review, it is hard to know. Evaluation is the best way we have found to provide the necessary information.

What Is Evaluated? An Excursus on Terminology

To be sure we are on the same wavelength, let's agree on the meaning of a set of terms that will be cropping up in the following chapters. Some of the definitions are strictly conventional and arbitrary, but if we agree on them, the rest of the book will mean the same thing to all of us.

From now on I will call a national program, like Head Start or Superfund environmental cleanup, a program. The local operations of the program are each projects. Thus, the Head Start that is operated in the Brigham Community Center is a project. An element of the Brigham Head Start project, like involving parents through weekly meetings, is a component. Evaluations can be directed at any of these levels. We can evaluate national programs, local projects, or subproject components, using the same basic methods.

We can also evaluate policies. For our purposes, a policy is an officially accepted statement of objectives tied to a set of activities that are intended to realize the objectives in a particular jurisdiction. Thus, one federal policy is to encourage donations to charity by allowing taxpayers to take deductions for such donations on their income tax form. There's no program there, but the policy can be evaluated. Evaluative question: Does the tax deduction actually promote philanthropic giving? Another federal policy aims to ensure the health and safety of those who use medical appliances (such as heart pacemakers or wheelchairs). The government requires that manufacturers test the safety of devices before they are marketed and again after they are on sale by asking users to report defects while the devices are in use. That policy can be evaluated, too. Evaluative question: Do the testing requirements and reporting procedures increase the safety of devices? Much the same evaluative strategies are used in policy evaluation as in the evaluation of programs.

In fact, the techniques of evaluation are marvelously adaptable. They can be applied not only to social programs and policies but also to environmental policies, mass transit programs, forestry projects, and elements of military procurement. In a number of such fields, the word evaluation is not in common use, but under other guises (e.g., social audits, system reviews, performance measurement), evaluative work is undertaken.

To simplify the prose, I will generally talk about the evaluation of programs, but that is a shorthand device. In that phrase I mean to encompass the evaluation of policies, projects, and components as well.

Another convention I adopt here is to call the evaluator she, and all other players in the program world he. So policymakers, program managers, program staff, and clients of the program all turn up with the masculine pronoun, while I use the feminine pronoun for members of the evaluation team. This is by no means to suggest that all evaluators are, or should be, women. Nor do I mean to suggest that everybody else in the program arena is, or should be, male. It is simply a convention to avoid the awkward he or she wherever I use the singular form.

There are other words that come into play frequently in evaluation, many of them already familiar to you through your exposure to research. Among these are qualitative research, quantitative research, experiment, control groups, randomization, validity, reliability, measurement, indicators, variables, sampling, empirical research. If you are not certain of their meaning, it is probably a good idea to brush up on definitions and come to feel comfortable with the terms. A glossary appears at the end of the book. You might also look in the index for pages on which there is fur-
ther discussion, and you can refer to other texts on social science research for fuller discussions of the terms. In this and the next chapter I introduce terms that are unique to evaluation: *formative* and *summative evaluation*, *outcome* and *process evaluation*, and in subsequent chapters we will encounter the rest of the vocabulary.

### Outcome and Process Evaluation

What kinds of questions does evaluation ask about programs (policies, projects, components)? From the time it became a recognized activity in the 1930s, on into the 1970s, the main question had to do with outcomes. Evaluations tended to concentrate on the question: Is the program reaching the goals that it was set up to accomplish? Or, more broadly: What are the results of the program? This focus on outcomes is still a hallmark of the evaluation enterprise.

*Outcomes* refers to the end results of the program for the people it was intended to serve. I use the term *outcomes* interchangeably with *results* and *effects*. Some of the outcomes that are realized are the results that program planners anticipated—the things they wanted to happen. Other outcomes, however, are side effects that nobody expected—often effects that nobody wanted. Thus, for example, a program that serves parents who are deemed to be at risk of abusing their children may have the unintended result of labeling them as potential abusers, even to the parents themselves. Once they accept that label for themselves, they may become more, rather than less, abusive. Or a program that sets up citizens’ advisory groups for toxic waste cleanup activities may have the goal of speeding the cleanup process by dint of citizen pressure. However, when citizens are represented, they may introduce a host of other concerns into discussions of cleanup, such as issues of employment and unemployment, and thereby divert activities from cleanup to job maintenance.

Another word that sometimes comes up when people talk about the focus of evaluation is *impact*. Most of the time, the word means the same as outcomes. An impact study looks at what happens to participants as a result of the program. Sometimes *impact* is construed as long-term outcomes. Occasionally, writers use the word *impact* to mean effects of the program on the larger community. For example, the outcomes of a job training program may be studied in terms of the number of *program participants* who get jobs. Its impact, in this larger sense, might be the effect on the unemployment rate in the *whole geographic area* in which the program functions. In other cases, impact comes between program processes and outcomes (Green & Kreuter, 1991).

A further meaning of impact is the *net* effects of the program, after taking account of what would have happened in the absence of the program. In operational terms, this means looking at the outcomes for program participants (say, 9% hold jobs after the training program) and looking at outcomes for an *equivalent* population of nonparticipants (say, 3% hold jobs at the same point). Then, on the assumption that the same proportion (3%) of participants would have held jobs if they hadn’t attended the program, the evaluator estimates the impact by subtracting outcomes for nonparticipants from the outcomes for participants. She estimates the impact of the program as a 6 percentage point increase (9%-3%) in job holding. (The trick here is to be sure the two groups are equivalent. I discuss this issue in Chapter 9.)

The last meaning of *impact* is a vital concept in evaluation, but I don’t use the word *impact* in the book. I refer to *outcomes*, and in context the word often means “net outcomes”—that is, that part of the outcome that is attributable to the program. This will become clearer as we move along.

Evaluation can look not only at short-term outcomes but also (if it has world enough and time) long-term outcomes. The High/Scope study of the Perry Preschool Program has followed the children who attended preschool as 3- and 4-year-olds for 24 years (Schweinhart, Barnes, & Weikart, 1993). The long-term follow-up has enabled the evaluators to report that at age 27, young people who had attended the preschool program had significantly higher monthly earnings than an equivalent group of people who had not attended, were more likely to own their own home, had fewer arrests, and completed more schooling. These data have been extremely influential in justifying the Head Start Program, which was loosely based upon the Perry Preschool Program, and in generating support for increased funding.

Although outcomes were the original evaluative focus, evaluation questions now deal not only with outcomes but with the process of the program, too—that is going on. Evaluators need to study what the program actually does. In the early days they took for granted that the program was doing what its operators said it was doing. But they soon learned that the assumption was often ill-founded. Ralph Tyler, one of the pioneer evaluators, wrote in 1942 that his Eight-Year Study of 30 high schools found that in the first year “only two or three were operating the program described in the proposal” (Tyler, 1991: 7). Hyman and Wright (1967: 745), who were evaluating a rural health program in Egypt, found that most of the centers lacked vital categories of personnel and that even personnel on the job were putting in relatively few hours of work. Evaluators had to know what the programs were actually like before they could draw conclusions about whether they were successful or not. It might be that they never really happened.

There are other reasons for studying program process. Sometimes the key questions that the program community has about the program have to do with its process. For example, what kinds of service are participants being given? Is the service following the prescriptions of the program developer? How often are participants (say in a drug treatment program) showing up? What kinds of problems are staff encountering? Are clients happy with the program? There is a need for studies that focus on systematic assessment of what is happening inside the program.

Another reason to study program process is to help understand outcome data. The evaluator may find that some participants in the program did particularly well and others did exceptionally poorly. There are a lot of possible reasons for such a finding, but one possible reason is that people received different kinds of service or different intensity of service. One group may have received services from highly qualified staff, had the same staff person over a long period of time, and attended regularly. Another group may have had a series of different staff members who were poorly trained, and they may have failed to attend very often. If the evaluator is going to be able to analyze what conditions were responsible for different outcomes, she needs data on what went on in the program.

So we have at least three situations that call for process data. One is when the key questions concern process. Evaluation sponsors want to know what is going on.
Another is when key questions concern outcome, but we want to be sure what the outcomes were outcomes of? People often look on one project or a handful of projects as representatives of a class of programs, such as short-term psychotherapy or community health centers. The evaluator wants to know whether there actually was a community health center in place, and what kind of program it was, before she concludes that the Community Health Center Program succeeded or failed. The third situation is when the evaluator wants to associate outcomes with specific elements of program process—that is, to find out which particular features of the program were associated with greater or lesser success.

Contributions of Evaluation

Historically, evaluation research has been viewed by its partisans as a way to increase the rationality of policymaking. With objective information on the implementation and outcomes of programs, wise decisions can be made on budget allocations and program planning. The expectation has been programs that yield good results will be expanded; those that make poor showings will be abandoned or drastically modified. A quote from a congressman, made many years ago, is still a fine summary of the rationale for program 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 returning adequate dividends in terms of results. Without 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. (Dwyer, 1970)

In pursuit of the same objectives, in 1993 Congress passed the Government Performance and Results Act, which requires federal agencies to gather program performance data. The act requires that agencies make a strategic plan for their program activities; establish program performance goals in terms that are objective, quantifiable, and measurable; and then collect data that register the extent to which the program is achieving the established goals. Agencies are required to submit an annual report on program performance to the president and the Congress.

History of Evaluation

These recent efforts to institutionalize evaluation in government agencies are the latest in a long series of attempts to use data and evidence in a search for better understanding of social behavior and wiser social policy. If we are to trace evaluation back to its prehistory, we would probably start the story in the 1660s.

Evaluation is rooted in the empirical study of social problems that began in earnest in Britain in that period (Cullen, 1975). Although the intellectual origins of what was once called "political arithmetic" remain a matter of debate, it is clear that the seventeenth century saw the beginnings of a search for social laws comparable to those contemporaneously developing in the physical sciences. Perhaps the first study that can be labeled evaluative came almost two centuries later. A. M. Guerry, a Frenchman, published a statistical study in 1833 that attempted to show that education did not reduce crime (Cullen, 1975, p. 139). Other statisticians marshalled different data in an attempt to refute his findings. In a counterpoint that has remained a continuing feature of evaluation history, these statisticians not only cited different evidence but also criticized Guerry's methods in their zeal to establish that education did in fact lead to reduction in crime.

At about the same time, another Frenchman, Jules Depuit, assessed the usefulness of public works, such as roads and canals. He published a paper in 1844 that measured the value of a canal project through techniques of calcul économique. He used the maximum tolls that users would pay as evidence of the worth of the canals, and he examined the drop in demand as tolls went up as a measure of the limits of their utility (Toulemonde & Rochaix, 1994).

Despite some early forays such as these, evaluation as we know it today is a relatively recent development in the history of the world, even in the history of social programming. Early policies to improve social conditions did not include provision for evaluation. When reformers of the late 19th and early 20th centuries used social science research procedures, it was to conduct surveys in order to document the extent of problems and locate people in need (Bulmer, Bales, & Sklar, 1991). They took for granted that the remedies they provided would solve the problems. There was no evaluation of the prison reforms promoted by Dorothea Dix or of the social services provided at Jane Addams's Hull House. Very little study was done on the effects of installing electric street lighting or purification of water. When the United States passed laws in the second decade of the 20th century prohibiting child labor, nobody thought of evaluating the outcome. It was assumed that child labor would end and results would be intrinsically good. When the United States instituted a system of unemployment benefits in 1935, no evaluation accompanied it. Giving money to unemployed workers to tide them over until they found a new job was obviously beneficial.

People working in the fields of education and health were among the first to do systematic studies of the outcomes of their work. In 1912 R. C. Cabot examined 3,000 autopsy reports and compared them to the diagnoses that had been made of each case; he wrote an article in the Journal of the American Medical Association that was essentially an evaluation of the quality of medical diagnosis (Flook & Sanazaro, 1973). In 1914 Dr. Ernest Codman, a surgeon at Massachusetts General Hospital, insisted that the way to evaluate surgeons' performance was by measuring how patients fared after they left the hospital. Codman himself collected a great deal of data, but his work was largely ignored by the medical establishment (Vibbert, 1993). In education one of the best-known studies was the Eight-Year Study sponsored by the Progressive Education Association in 1933, and directed by Ralph Tyler, which looked at the outcomes of programs in 15 progressive high schools and 15 traditional high schools. An early evaluation of the effectiveness of school health programs was conducted by George Palmer, chief of the research division of the American Child Health Association; it was released in 1934. At the end of the
CHAPTER 1

Setting the Scene

In the 1940s private foundations began funding evaluations of an array of innovative social programs that they sponsored. A famous one was the Cambridge-Somerville youth worker program to prevent delinquency in suburban neighborhoods near Boston (Powers & Witmer, 1951). Early results were promising, but longer-term follow-up showed that at-risk youth who had received program services were barely doing as well as those who had not (McCord & McCord, 1959). One interpretation was that they had grown so dependent on the help of the youth workers that they had not developed the skills necessary for solving their own problems.

By the 1950s the federal government was sponsoring new curriculum efforts, such as Harvard Project Physics, in response to fears of American scientific illiteracy in the wake of the Soviets' launching of the Sputnik satellite. Evaluations were funded to find out how successful the curriculums were. In the early 1960s the President's Committee on Juvenile Delinquency funded a series of projects to reduce youth crime around the country, and federal administrators required that each project evaluate the results of its activities.

The War on Poverty in the mid-1960s marked the beginning of large-scale government-funded evaluation. The federal government began funding an array of programs to help the poor, and it started to require systematic evaluation of the results of the money spent. The Elementary and Secondary Education Act of 1965 included a requirement for evaluation in the law. Senator Robert Kennedy was the moving force behind the requirement. He wanted to be sure that the new federal money was not going to support schools' tired old practices but rather would go to help disadvantaged kids in new ways. He wanted poor parents to be well informed about what was taking place in the schools so they could prod educators to serve their children more effectively. He saw evaluation as a tool to provide parents with the necessary information (McLaughlin, 1975).

Other programs of the War on Poverty were also evaluated, including programs that provided legal services, community health services, job training, nutrition supplements for pregnant women and infants, food stamps, housing vouchers, multi-service centers for social services, preschool education, innovations in delinquency prevention and corrections, mental health services, and community action programs that mobilized residents of poor neighborhoods to determine their own priorities and demand the services they needed. Evaluators developed new methods and tools to fit the varied content and settings of the programs. The development of evaluation in this period owes a lot to poverty, just as it earlier owed a lot to illiteracy and crime.

The same period saw the rise of cost-benefit analysis in the RAND Corporation, Department of Defense, and elsewhere. Defense Secretary Robert MacNamara's policy analysis, dubbed the whiz kids, analyzed the relative advantages of weapons systems to see (in the contemporary phrase) "how much bang for the buck" they delivered. Important advances were made in methods of economic analysis.

Evaluation branched out into other areas. With new legislation came evaluation in such areas as environmental protection, energy conservation, military recruit-

ing, and control of immigration. By the end of the 1970s evaluation had become commonplace across federal agencies. Almost every department had its evaluation office, and some had evaluation offices at several levels in the hierarchy—at the level of the secretary of the department, attached to the major program area, and at the operational level.

In the field, a host of small centers and firms were established to undertake federally financed evaluations. University research centers enlarged their charters to encompass evaluation, special centers were set up, new not-for-profit and for-profit research and consulting organizations arrived on the scene, and established ones expanded into the evaluation area. Many researchers re-tooled their skills so as to take advantage of the new streams of money.

A high point in evaluation history came in the 1970s with the inauguration of a series of social experiments to test novel policy and program ideas prior to their enactment. The Negative Income Tax experiment was the largest and most widely publicized (Cain & Watts, 1973; Kershaw & Fair, 1976, 1977). It was followed by experiments with housing allowances (Carlson & Heinberg, 1978; Friedman & Weinberg, 1983; Kennedy, 1980), health insurance (Newhouse et al., 1981; Phelps & Newhouse, 1973), performance contracting in education (Rivlin & Timpone, 1975), and other smaller experiments. In these experiments pilot programs were implemented on a large enough scale to simulate actual operating conditions, and experimental results were expected to help policymakers decide whether to move ahead with the policies nationwide. As it turned out, by the time the results of the experiments became available, the political climate had generally changed. The Negative Income Tax study is a case in point. The initial ardor for change had subsided, and the steam had gone out of the reform movement. In the end, few components of the experimental programs were adopted. However, the information remained available for later policymaking. For example, in 1994 when health reform was on the agenda again for a time, results of the health insurance experiments were dusted off, a new book was published (Newhouse & Insurance Experiment Group, 1993), and the findings entered the discussions.

Evaluation continued to be a growth industry until the Reagan administration took office in 1981. At that point funding for new social initiatives was drastically cut. New and innovative programs have always been the most likely candidates for evaluation, and when the stream of new programming dried to a trickle, fewer calls came for evaluation. Nonetheless, evaluation continued at modest levels, not so much with special evaluation money as with money from the operating agencies. Evaluations were undertaken on operating programs for long-term care, teenage parents, dislocated workers, state waivers of federal rules for the Aid to Families with Dependent Children (AFDC) program, supported work, and others.

During the late 1980s and early 1990s, evaluation funding made a partial comeback (U.S. General Accounting Office, 1987). Some agencies and agencies prospered. For example, Ginsburg, McLaughlin, Fusko, and Takai (1992) wrote about the "reinvigoration" of evaluation at the Department of Education. Other agencies remained in the doldrums (Wye & Sonnichsen, 1992). Over all, evaluation held on to its place in the bureaucracy, and important new work was launched. With the Clinton presidency, more social programming and more evaluation were under-
taken, but the Republican revolution that began with the 1994 elections called for
massive downsizing of the federal government and the shift of much social pro-
gramming to the states.

In the past, both conservatives and liberals have found evaluation useful.
When conservatives are in office, the emphasis for evaluation tends to be on pro-
gram cost cutting (how does the program stack up on criteria of efficiency and cost
reduction?) and elimination of service to ineligible recipients. When liberals are in
power, standards tend to focus on the effectiveness of service in terms of improvement in
beneficiaries’ life chances.

With all the backing and forthing at the federal level, one of the notable fea-
tures of recent evaluation history is the growth of evaluation activity at state and
even local levels. For example, the Commonwealth of Massachusetts evaluated its
education and training program for welfare recipients (Massachusetts Department of
Public Welfare, 1986a, 1986b, 1986c); the state of South Carolina evaluated a pro-
ject integrating human services (South Carolina State Reorganization Commission,
1989); the city of Chicago, along with local universities and citizen reform groups,
evaluated the city’s major school reform (Bryk & Rollow, 1992, 1993; Bryk, Deabster,
Easton, Luppescu, & Thum, 1994; Sebring, Bryk, Luppescu, & Thum, 1995).

Another recent trend is the increasing use of qualitative methods for evalua-
tion. Not too long ago the only kind of evaluation with professional legitimacy, in
rhetoric at least, was quantitative evaluation, preferably using a randomized exper-
imental design. (See Chapters 8 and 9 for discussion.) But some evaluators relied
more on words than on numbers, collected their data through observation and informal
interviewing rather than through structured interview questions or quantitative
records, and their analysis probed the meaning of process and outcomes through nar-
rative analysis. (See Chapter 11.) During the late 1970s and 1980s, they burst into
the evaluation literature with an outpouring of books and journal articles, heralding
the advantages of their approach (e.g., Bogdan & Biklen, 1982; Feiterman, 1988;
Guba, 1990; Guba & Lincoln, 1989; Patton, 1980; Stake, 1975, 1995). They pro-
vided a spirited exchange with supporters of quantitative methods, and the fur flew
thick and fast.

Hard upon “the paradigm wars,” as these sometimes harshly worded critiques
and exchanges came to be known, came attempts at rapprochement. Many key fig-
ures in evaluation concluded that evaluation was a house of many mansions and had
room for a variety of approaches. In fact, qualitative and quantitative methods could
complement each other well, and studies using both kinds of procedures began to
thrive.

Out of the melee came increased awareness of the advantages of qualitative
techniques and a heightened legitimacy for qualitative work. More evaluations are
incorporating a qualitative component, especially in education, and textbooks are
now including chapters on qualitative approaches to evaluation. (Including this one; see
Chapter 11.)

Another noteworthy development has been the development of professional
associations in evaluation. Professional associations—the American Evaluation
Association in the United States and similar associations in Canada, Europe, Great
Britain, Australia and New Zealand, and elsewhere—provide a forum for evaluators to
share their work and their concerns. At annual meetings and through publications, they
offer opportunities for evaluators to disseminate their findings, keep abreast of new
techniques, consider larger issues regarding the role of evaluation in society, propagate
standards of professional conduct, and generally advance the state of the field.

Comparison between Evaluation and Other Research

Evaluation applies the methods of social science research, both quantitative and quan-
titative. 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

Utility Evaluation is intended for use. Where basic research puts the empha-
sis on the production of knowledge and leaves its use to the natural processes of dis-
semination and application, evaluation starts out with use in mind. In its simplest
form, evaluation is conducted for a client who has decisions to make and who looks to
the evaluation for information on which to base his decisions. Even when use is
less direct and immediate, utility of some kind provides the rationale for evaluation.

Program-Derived Questions The questions that evaluation considers derive
from the concerns of the policy and program communities—that is, the array of peo-
ple involved with, or affected by, the program. Unlike basic researchers who formu-
late their own hypotheses, the evaluator deals in the currency of program concerns.
Of course, she has a good deal of say about the design of the study, and she
approaches it from the perspectives of her own knowledge and discipline. She can
choose the way the questions are posed, stimulate further questions, and exercise
control over how many questions the evaluation can comfortably address. She is
usually free to incorporate inquiries on subjects of particular concern to her. But the
core of the study represents matters of administrative and programmatic interest.

Judgmental Quality Evaluation tends to compare “what is” with “what
should be.” Although the investigator herself usually tries to remain objective, she is
typically concerned with phenomena that demonstrate how well the program is func-
tioning and whether it is achieving its intended purposes. Wherever the questions for
study are formulated, somewhere in the formulation appears a concern with mea-
suring up to explicit or implicit standards. This element of judgment against criteria
is basic to evaluation and differentiates it from many other kinds of research.

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 like-
ly 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 assign-
ment of participants to program activities and locations. Not infrequently, research requirements (for "before" data, for control groups) run up against established program procedures, and there is tension about which is to prevail.

**Role Conflicts** Interpersonal frictions are not uncommon between evaluators and practitioners. 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 is not likely to make such contributions to program improvement that it is worth disruptions and delays. Often, practitioners 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 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 desired purposes, 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 intent on moving on the next evaluation contract, that they submit the required report to the agency and go on to a new study.

Fortunately, in the past decades new publication channels have opened. Almost a dozen periodicals now publish articles reporting evaluation studies or discussing evaluation methods, philosophy, and uses. Among them are *Evaluation Review*, *Educational Evaluation and Policy Analysis*, *Evaluation and the Health Professions*, *New Directions for Evaluation*, *Evaluation Practice* (renamed *American Journal of Evaluation*), *Evaluation and Program Planning, Studies in Educational Evaluation*, and *Evaluation: The International Journal of Theory, Research, and Practice*. These journals, and journals in substantive fields such as substance abuse or criminology, provide a place where evaluators can share the results of their work and also discuss innovative developments in their craft.

The journals also make study results visible to people concerned with program and policy. If progress is to be made in understanding how programs should best be conducted and where and when they can be improved, a cumulative information base is essential. Only through publication will results build up. Even when evaluation results show that a program has had little effect, a situation that generally makes authors and editors reluctant to publish, it is important that others learn of the findings so that ineffective programs are not unwittingly duplicated again and again. When program results are mixed, some good, some not so good, people who run programs would profit from learning about the components of the program associated with greater success.

---

3 In this technologically advancing age, computerized services are being created to supplement journal publication.

---

**Setting the Scene**

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 the results do not generalize beyond the immediate project, there is little to report to others. To avoid these limitations, evaluators need to keep the needs of broad audiences in mind when they plan their work. Then published reports can add to the stock of program knowledge.

**Allegiance** The evaluation researcher has a dual, perhaps a triple, allegiance. She has obligations to the organization that funds her study. She owes it a report of high quality and as much usefulness for action as she can devise. She wants it to be helpful to the interests of policymakers, managers, practitioners, and program participants. Beyond the specific organization, she has responsibilities to contribute to improvement of programming in the field she has studied (science education, firearms regulation). Whether or not the organization supports the study's conclusions, the evaluator often perceives an obligation to work for the application of the conclusions in order to advance program and policy in the field. On both counts, she has commitments in the action arena. She also has an obligation to the development of knowledge and to her profession. As a social scientist, she 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 constraints that attend its practice.

**Similarities**

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

Evaluators, like other researchers, 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 processes and effects. The data collection scheme to be used depends on the information needed to answer the specific questions that the evaluation poses.

The classic design for evaluations has been the randomized experiment. 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, postprogram surveys, time series, correlational studies, and so on.

There is no cut-and-dried formula to offer evaluators for the "best" or most
suitable way of pursuing their studies. The programs and agencies with which evaluation deals are so diverse and multi-faceted that the specifics of the particular case exert significant influence. Much depends on the uses to be made of the study: the decisions pending and the information needs of the decision-making community, or the uncertainties in the field and the need for better understanding of how programs work. 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. If the evaluator is on the payroll of the agency running the program, there may be limits on her freedom to explore negative aspects or negative outcomes. 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 sometimes seen as the drones of the research fraternity, drudging away on dull issues and compromising their integrity out in the corrupt world. But as any working evaluator will fervently tell you, evaluation calls for a higher level of skills than research that is designed and executed under the researcher’s control. It takes skill to make research simultaneously rigorous and useful when it is coping with the complexities of real people in real programs run by real organizations—and it takes some guts.

The evaluator has to know a good deal about the formulation of the research question, study design, sampling, data collection, analysis, and interpretation. She has to know what is in the research methodology texts, and then she has to learn how to apply that knowledge in a setting that is often inhospitable to important features of her knowledge. If she persists in her textbook stance, she runs the risk of doing work irrelevant to the needs of the agency, antagonizing the program personnel with whom she works, and seeing study results ignored—if indeed the work is ever completed. So she sometimes has to find alternative ways of conducting the study, while at the same time she stands ready to defend to the death those elements of the study that cannot be compromised without vitiating the quality of the research. Then she needs finely honed skills in disseminating the results of the study in ways that encourage and support the application of findings to the improvement of policy and program.

Summary

This chapter defines evaluation as the systematic assessment of the operation and/or the outcomes of a program or policy, compared to explicit or implicit standards, in order to help improve the program or policy. Evaluation is undertaken when the program is complex, hard to observe, composed of elements interacting in multiple ways; when the decisions to be made are important and expensive; when evidence is needed to convince others about the merit and/or failings of the program. Evaluation can also be a tool of accountability. Program sponsors and operators can use evaluative evidence to report their operations and outcomes to broader publics.

Evaluation can address the process of the program (i.e., how it is being imple-