CHAPTER 12

Evaluation

ROSS KOPPEL

PREVIEW

Evaluation is not just for experts. You already know a lot about it. You evaluate people, programs, and organizations all the time. You evaluate teachers, restaurants, possible dates, former relationships, cars, school, sports teams, local policies, and probably, yourself.

Think about what you consider when you evaluate anything:

1. Did you get the product, outcome, or service you thought you would get?
   a. How did it work? Does it still work?
   b. Did you get more, less, or something different than what you expected?
   c. Which issues, characteristics, or aspects were more important to you in your evaluation? Which were less important to you?
2. What about the *process*?
   a. Was it timely, thoughtful, responsive, nasty, helpful?
   b. Did you have any say over what went on?
3. Did you fully understand the goals of the persons or organization? For example:
   a. Were the school's aims clear to you?
   b. Did you understand if the restaurant was interested more in entertainment than in food?
   c. Did you know your date's goals?
4. Did the goals of the persons or organizations change sometime after you started? For example:
   a. You could have wanted a fast car at first but concluded you preferred a comfortable one.
   b. Your school could have shifted orientation to favor business programs.
   c. Your dates could have decided they wanted something different from what they sought in the first flowering of the relationships.
5. Did you use systematic data collection and analysis? For example:
   a. Some car evaluations are based on surveys of former owners.
   b. Many schools use student evaluations in considering professors' promotions.

ROSS KOPPEL • Social Research Corporation, Wyncote, Pennsylvania 19095.

Handbook of Clinical Sociology, Second Edition, edited by Howard M. Rebach and John G. Bruhn. Kluwer Academic / Plenum Publishers, New York, 2001.

   c. Many colleges are evaluated in terms of the job successes of their graduates, the number of publications of their faculty, the sizes of their libraries, and other seemingly numeric information.
6. You have probably employed "expert" evaluations. For example:
   a. Consumer reports on cars.
   b. Newspaper reviews for restaurants and movies.
   c. Friends' insights on possible dates (recognizing the serious methodological and other dangers in such practices).

You have also undoubtedly recognized that most evaluation tasks are not "yes" or "no," but rather are more nuanced or contingent, as in: "good for X but not for Y," or "ideal for beginners but boring for experts," or "fun if in that kind of mood." For example, think about the following statements:

"Ideal car if you commute across the Mojave Desert but lousy around town."
"Not someone for me now but perfect for those who want to feel bad about themselves."
"Great clothes for people who are over 6′2″ and are under 90 pounds."

As part of this natural evaluation, you have undoubtedly weighed several different factors—recognizing that there are many characteristics you might seek or seek to avoid. In restaurants, you might seek value for the money *and* convenience. You might want a school with access to skiing or with access to beaches. In selecting a class, you might seek a subject in which you have some interest *and* one that fits into your schedule *and* one that has a dynamic teacher. How you balance, or weigh, each of these factors reflects your needs, values, and constraints.

Last, in our preview of evaluation methods, you have almost certainly communicated, or responded to, the findings of your assessments:

1. You've probably told a waiter the food was not great, and hopefully found other places to eat.
2. You may have realized that trying to learn an entire course the night before a final is not a wise pedagogic policy.
3. You've told friend X that he can keep his movie recommendations to himself but sought the opinion of friend Y on films.
4. You've probably concluded that some dates are not worth another evening—and made those feelings clear to your dates.

INTRODUCTION

All of the issues and trade-offs we have just discussed come into play in evaluation research. Of course the question is *not* about which restaurant, car, or date. It is about *understanding how a policy, program, or organization is helping (or failing to help) some group of people or, perhaps, society in general.* But the concerns we just discussed have exact parallels in evaluation research. Consider these examples of evaluation questions:

1. Could the state have spent its money more wisely in helping the elderly? Did anyone find out what the elderly clients of the program needed or wanted? Do the clients know the available options?
2. Did the policy to assist people in the transition form welfare to work have clear goals and strategies? That is, did the effort pay attention to the kind of work people might

get? Do the jobs pay enough to get a family out of poverty? How long do people stay in those jobs? Do the former welfare recipients lose so many benefits that the jobs fail to make sense for them?

3. Did the policy to reduce the cost of long-distance calls help the average person if we remember that most calls are local? Perhaps businesses benefit most?
4. Did stakeholders[1] involved in a program share the program's goals? Were the plans to accomplish those goals well articulated? Did they make sense? Did the organization have the funding and resources needed to run the program? Did someone try this before? Are we learning from the efforts of others and of this program?
5. How many people are helped by an alcohol treatment program? For how long are they helped (for how long do they stay sober)? What proportion are helped? Which types of people are more likely to be helped?

With any program, agency, or policy, we ask: What does it cost? What are the benefits? What is the value of the benefits? Who benefits? What are the incentives for making alternative choices?

Why You Should Read about Evaluation

Evaluation research—of programs, policies, agencies, and of any "intervention"—is increasingly required by all sponsors. Governments and foundations demand it, leaders use it to justify their work, community groups use it to determined if they are being well served, and wise administrators request it for feedback on improvements and on allocation of resources. You will undoubtedly be involved in some work that is being evaluated or is evaluating some other project. Even if you will not conduct an evaluation yourself, you must be able to read and respond to evaluation reports and methods. By learning about evaluation methods you will know what to look for, how to defend yourself, who to hire, and what to request. Note also that evaluations are increasingly ongoing events, not just something that comes at the end to see if the work was acceptable. They can provide frequent feedback to help long-range strategy and immediate plans. Last, understanding the principles and practices of evaluation research will probably help you make better decisions in your own life.

Outline of the Chapter

We have already introduced the basic concepts of evaluation. In the remaining pages of this chapter we elaborate on these concepts.

In the next section we revisit our short definition of evaluation, offering, we hope, more clarity. Then, we very briefly indicate some of the history of evaluation. In the fifth section, we provide an overview of four different types of evaluation. This is the major portion of this chapter and we also examine the purposes of evaluation, looking at the differing needs of stakeholders, clients of the programs or agencies, and so on. We provide examples throughout this discussion from some of the evaluations we and others have conducted.

Good evaluation is based on rigorous research methods and research designs. In the sixth section we discuss the relevant designs and methods of evaluation research. We also look at

[1]Stakeholders in the this context are the people, groups, neighbors, or organizations that are affected by the actions of the policy, program, intervention, or agency. This includes clients, staff, funders, and administrators.

some of the adjustments and threats to research design that evaluation involves. In the last section we discuss evaluation in context: the political and value conflicts implicit in any evaluation and the audiences for the reports and findings. Throughout the discussion we include short sidebars (boxed texts) about the work of a few evaluators.

DEFINITIONS

Weiss (1998) and Rossi *et al.* (1999) offer useful, if not graceful, definitions of evaluation research.

> Weiss: [E]valuation 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.
>
> Rossi *et al.*: [Evaluation is] the use of *social research procedures* to *systematically investigate* the *effectiveness* of *social intervention programs* that is *adapted to their political and organizational environments* and *designed to inform social action* in ways that *improve* social conditions.

Let us examine the key terms in these definitions (in italics). To save time and space, we discuss the terms in both quotations together, whenever possible.[2]

Systematic assessment or *systematically investigate*: While some evaluations may consist of casual observations, the kinds of evaluation research that are most compelling for program administrators and participants and for funders involve rigorous research design, clear measurement, careful data collection, and intelligent analysis. Good evaluation incorporates good and systematic *research procedures*.

Operation of the program refers to analysis of *process*—how the program or agency conducts its work. While not all evaluations focus exclusively on process or operations, all evaluations address process because one cannot improve outcomes just by wishing them to be better. The outcome is the result of what the agency or program is doing—its operations or process. Note that many evaluations examine both outcomes and operations.

Outcomes or *effectiveness of social intervention programs*: The product or service the program, intervention, or agency provides is a critical measure in any evaluation. As we have seen and will see again very soon, what is defined as a good outcome can vary greatly for different observers, audiences, and stakeholders. Also the assessment of the outcome will always depend on how it is measured.

Compared to a set of standards: As in most research, one is making a comparison between something and another thing, including possibly a comparison between a service or product and some standard. What did we expect to happen? What happened? Assuming we can agree on measurement (never easy) and on a set of goals (often fuzzy), we can see how close the program or policy came to meeting its promises. Remember, too, that goals change. Perceived needs change. We also improve our ability to measure outcome when we collect data over time and/or when we learn from our previous efforts.

Adapted to their political and organizational environments: Programs, agencies, policies exist in the real world. They must negotiate their actions and goals in a dynamic setting of other agencies, legislators, regulations, staff, neighborhood groups, and so on. While this is happening and while the service or product is being offered, the norms of what is acceptable and appropriate may change. We increasingly seek to include the views of participants and

[2]I am following Weiss's format of explicating the definition, although I do not necessarily use her explanations or examples.

other stakeholders; to include the perceptions of clients and others who have been without a voice in the design and evaluation of programs. All research, and especially evaluation research, is political and is embedded in the complexities of previous interventions, of economic power, of clients' experiences, and of the programs' inertia, plans, and internal conflicts.

Improvement: One's view of *improvement* is entirely related to one's views of the program's or the policy's goals, outcomes, and constraints. If you think your program or agency is doing wonderful work under difficult circumstances, your recommendation for improvement would be to give it more resources. If you think the program or agency is beyond repair—a total waste of money and time—your view of improvement may be to terminate the program or agency. In short, all of the issues raised in the definitions must be confronted before we can address the ultimate goal of improvement.

Designed to inform social action: In general, people do not like to be evaluated, and even friendly evaluators and evaluation reports are viewed with trepidation. You may have a carefully researched and insightful analysis of a program or policy, but the people in charge think it is a threat to their jobs, to their good work, to their clients. Evaluation research is supposed to provide information for a useful purpose. Based on the evaluation, administrators can change the target population to those with greater needs; staff can be more responsive to clients; neighbors and funders can consider the best uses of a resource. Programs can benefit significantly from evaluation, and evaluation reports are often welcomed with great appreciation. But not always. While not common, evaluation findings can be used to end a program or to seek very new goals, even if those in charge are vigorously opposed to such changes.

Too gloomy? While we have used these definitions to discuss several challenges faced by evaluators, we do not want to suggest that good evaluation research is unattainable or even rare. Under recent federal and private funder guidelines, almost every program must be evaluated. Most evaluations provide useful insights and recommendations, helpful discussions of measurement, and a chance for programs and agencies to examine their goals and outcomes in light of additional data and analysis. Most evaluations, in short, are very beneficial.

A VERY SHORT HISTORY OF EVALUATION

The *idea* of figuring out if some program, organization, or policy is doing anything useful did not require modern methods or even modern ways of thinking. However, the widespread use of social science research methods in evaluation—for example, large surveys, systematic data collection, and statistical analysis—is usually thought to have begun around the 1930s, with the growth of social programs of the New Deal (Rossi *et al.*, 1999). The need to better understand morale and propaganda during the Second World War (1939–1945) was another impetus for evaluation research. By the late 1950s, evaluations were routine for programs in public health, housing, occupational training, and delinquency prevention (Rossi *et al.*, 1999). The greatest spur for evaluation, however, came with the War on Poverty–Great Society programs of the late 1960s and early 1970s, when the federal government devoted considerable economic and political resources to addressing income and racial inequalities. Programs like Head Start, affirmative action efforts, urban renewal, job training, and federal aid to education increasingly required evaluation. Evaluation research, perhaps ironically, was also encouraged by those who sought to end these meliorative programs. Foes of "big government," a term usually used by those opposed to programs for the poor, sought evaluation as a way to justify ending programs feared to be inefficient, wasteful, corrupt, or just unwanted. Thus, evaluation research—which is ideally used to improve programs, help them target their services to the

Journals:
Evaluation Practice, renamed in 1998 to *American Journal of Evaluation*
Evaluation Review: A Journal of Applied Social Research
Sociological Practice: A Journal of Clinical and Applied Sociology
New Directions for Evaluation
Evaluation: The International Journal of Theory, Research and Practice
Evaluation and Program Planning
Journal of Policy Analysis
Administration Science Quarterly

Professional organizations:
American Evaluation Association
Association for Public Policy Analsysis and Management
Sociological Practice Association
American Sociological Association
Society of Applied Sociologists

FIGURE 12.1. Evaluation journals and professional associations.

most appropriate clients, and make them more efficient and effective—was further promoted by those who wanted to extinguish or diminish those programs.

As this is being written, the United States is engaged in reform of the national welfare structure, having shifted from an entitlement program (Aid to Families with Dependent Children), to a system with greater requirements for work outside the home and limitations on the duration of welfare assistance (Temporary Assistance to Needy Families). Evaluation is frequently emphasized with the new system. But meaningful evaluation has been hampered by the difficulties of collecting data on those who appear to have dropped out of the welfare system and by the often temporary nature of work among the poor. But the lack of clear evaluation data does not limit the enthusiasm for reform. It is a safe bet that evaluation reports will continue to be cited by all sides in this national experiment.

The demand for evaluation research is also assisted by the many educational institutions that produce people skilled in evaluation research methods. In addition to sociologists, graduates of sister disciplines, such as anthropology, psychology, and political science, are sometimes taught and often practice the methods used in evaluation research. Schools of education, social work, public administration, public health, and management often include evaluation methods in their curricula. Statisticians and engineers, also, are often evaluators.

We now have tens of thousands of evaluation results based on studies of tens of thousands of interventions. Some evaluation reports have been very influential, the basis for billions of dollars in funding and effort. Some reports have probably never been read by more than a few administrators. In all cases, the numbers never speak for themselves. Data and findings are interpreted through the lens of political and social orientation; the fate of programs is almost always determined in the political arena, not on the desks of statistical experts or of policy analysts.

FOUR MODELS OF EVALUATION RESEARCH

Evaluation research is often presented within the context of four models:

1. Outcomes: object-based; outcome-based and impact assessment
2. Process evaluation or program monitoring

3. Naturalistic investigation
4. Expert

While there are disagreements about this schema, it is a fine way to start. Recognize that there are hundreds of books on each of these methods and some scholars will be upset with any classificatory schema that lumps their favorite method with others. More important, these methods are often combined in practice, improving our understanding of the program or policy being evaluated.

Object-Based Evaluations

Object-based evaluations center on the question: *Has the program or policy achieved its goals?* This question focuses the evaluation on the stated intentions of the program or policy—their formal goals. Thus, the evaluator might assess the number of clients served by a soup kitchen, the number of clients who have fond jobs after completing a job training program, the improvement in interoffice communication following a seminar on office interactions, the number of babies inoculated via a medical outreach effort, the number of communities helped with new traffic plans. Object-based evaluation is not limited to merely counting. One could, if so specified in the program's goals, measure the *quality* of the service or product. For example, the evaluator could measure:

1. The nutritious content of the soup kitchen's soup and/or soup quality ratings by the clients
2. The quality of inoculations via blood tests or subsequent infections, or
3. The change in quality of discussions and interactions following a human relations seminar.

With object-based evaluation, the researcher measures how well the program achieves its goals. It turns out, however, that goals are often not stated, not stated clearly, and not stated in ways that are measurable. Sometimes the goals may be clear but they are so outlandish or outdated that they make no sense.

Consider these goals from a variety of programs:

1. Bring God to the poor of Camden
2. Make the children of Cleveland happy
3. Show that our company cares about the five-county region
4. Improve the quality of life for the elderly of Cherry Hill
5. Make the United States proud again
6. Find a "Big Brother" for each of the boys in our neighborhood
7. Improve the quality of children's nutrition
8. Reduce taxes for honest Chicago businessmen
9. End dependency on welfare and make people work again.

Measuring some of these goals would challenge the most sympathetic and skilled evaluator. Some are so vague that measurement would be meaningless, futile, or silly. But let us take a look at *some* of the goals where we have a chance of measuring something of meaning—perhaps to the program, the participants, or the funders. Keep in mind that the evaluator's first role is often to help the client define goals and to define those goals in ways that can be measured.

1. *Bring God to the poor of Camden.* We are not theologians but we could measure changes in the locals' religious beliefs via surveys; and we could monitor church attendance in poor areas of Camden. Of course we would not know if any changes (assuming we had baseline data) were due to the particular program being evaluated or to larger social trends, or to a local crime wave, or to especially bad weather, or to other "acts of God."

2. *Make the children of Cleveland happy.* We could interview teachers, school counselors, a representative sample of children or parents. We could count smiles of kids at play. We could see if there is a reduction in prescriptions for pediatric antidepressants. Measuring a change in children's happiness is complex and we probably will not know what caused it even if we were happy with our measures.

3. *Show that our company cares about the five-county region.* Here there really are two issues to evaluate: (1) to assess if the company is doing anything that reflects its care for the five-county region and (2) to determine if people in the five-county region have noticed the company's caring acts. The company might claim that employing people from the area is a caring act. Certainly it is beneficial. The company might also point to a recent and generous donation to the local zoo or to its charitable work with the United Way. People in the area, in surveys and in focus groups, may express deep appreciation for the company's employment, good works, and generosity. Alternately, residents may never have heard of the company or think it an exploitative polluter of land, water, and air. Some may be both thankful for the

Example—of evaluating a nutrition education effort—also offers an illustration of the need for sociological sensitivity and imagination. Remember the food pyramid? Do you think it reflects objective science from nutritionists and public health experts? The complaints about the pyramid are many and fascinating, varying from measurement concerns to claims of racism and agism. On the level of measurement, there's how one defines a "serving." The USDA defines a serving as a much smaller unit than most people realized. A bagel, for instance, is not one bread/cereal serving, but three. A USDA "serving" of pasta is ½ cup; but most people eat three or four times that amount. The same applies for typical servings of meat and even vegetables. Then there's the question of what is understood by the admonition to eat fats and sweets "sparingly."

Even more interesting is the perspective that the food pyramid is inherently racist. Only a small proportion (less than 5%) of Caucasians are lactose intolerant. (can't digest milk products or milk). But the proportion of Asians, Latinos, Native Americans Americans and African Americans who are lactose intolerant is much greater—as much as five times higher. Thus the food pyramid's suggestion to eat two-to-three daily servings of milk, yogurt and cheese is seen by some as racist.

Last, the pyramid is criticized for focusing on average adults, and it is not necessarily appropriate for children and the elderly.

Post script on serving size: The suggested size comparison of a serving of meat used to be "about the size of a pack of cigarettes." Clearly, that was not the analogy one wants for a health education effort. They switched the analogy to a pack of cards. I await the anti-gambling forces to press to another analogy. Let's save some time and avoid "about the size of the average revolver handle" as our next metric.

FIGURE 12.2. The "racist" measurement error-prone food pyramid.

employment and the charity but unhappy over the environmental policies. All of this information would address the company's evaluation question (even if some of it were unappealing).

4. *Improve the quality of life for the elderly of Cherry Hill.* We hope the agency or program seeking to improve the lot of the local elderly has some *specific* goals, such as distributing large-print editions of the newspaper, making curb cuts to allow wheelchairs to negotiate sidewalks, or helping people get to the doctor. If they only have vague goals, like "improving the quality of life," it is unclear how they would go about implementing them. This is the same issue we just noted—helping the agency or program with goal specification and clarification.

5. *Find a "Big Brother" for each of the boys in our neighborhood.* Let us assume, for this example, the neighborhood has clear boundaries and we know all the boys in the neighborhood. It is very easy to determine how well the program has done at finding a Big Brother for the neighborhood's boys. (But we would undoubtedly urge that another question be addressed: Did the Big Brothers make a difference in the lives of these boys, e.g., compared with the boys without Big Brothers, did the boys with Big Brothers do better at school?)

6. *Improve the quality of children's nutrition.* There is, in fact, a discipline of nutrition education that routinely evaluates nutrition outreach projects like this. We have been involved in research of this type. Evaluators of these programs watch kids at school lunchrooms and daycare centers, give exhaustive food consumption surveys, monitor what is sold or distributed, and monitor what is thrown away. Again, here, clarification of goals is essential. Do we mean eating "five a day" of vegetables and fruits? Do we mean eating less refined carbohydrates or fats? Perhaps we are concerned about the *balance* of the diet? Are we concerned with vitamin and mineral consumption? All would be used to assess the programs' achievements.

7. *End dependency on welfare and make people work again.* We briefly discussed this earlier but the topic offers many lessons. Let us say that the new policies help former welfare recipients to get jobs and leave welfare. Sounds great: win–win. There have been thousands of these successful cases. However it may be that many of these former welfare recipients take jobs that do not offer benefits, have obliged some to quite promising education and training programs, and offer no opportunities for learning or advancement. Are we still convinced the program is a success? It might be a useful though difficult step toward self-sufficiency. Another difficulty is that many people have left welfare, but officials do not know what these people are now doing. They may be creating wonderful new companies that will employ hundreds of workers, they may be going into crime or drugs, they may be studying economics, social work, or marksmanship. We only know they are no longer collecting welfare from our state. Is this still a success?

The message of these examples should be clear: Meaningful object-based evaluation requires that the program's goals be well articulated and measurable (including a chance to understand causation—who is responsible for achieving the goals?). This does not mean, however, that you just throw up your hands when you are faced with fuzzy or unstated goals. On the contrary, it means, as we suggested earlier, that you have another job: to help the agency or program figure out what they are trying to do. This task often is remarkably helpful, forcing the agency or program to think about its actions, resources, clients, strengths and clients' greatest needs.

It is astonishing to learn how many programs do not have clear goals. How could a program, agency, or policy not have clear goals? How could the goals go unstated? What would prevent one from measuring those goals? Reasons for these seemingly odd gaps are

many. Sometimes programs and interventions are cobbled together in compromise and haste, with insufficient attention to goals and outcomes. In several cases, the goals are confused or contradictory. Sometimes programs are offshoots of other projects or agencies, left as orphans of good intention but poor direction. Some programs exist to funnel money to a legislator's district, such as a project or agency that few expect to be very helpful except as a conduit of money to the district. Sometimes agencies want to help people or communities but have not thought carefully or strategically about what they actually can do, should do, or are doing.

We have evaluated multimillion dollar programs that were created to convince politicians and government regulators that a company was a good corporate citizen. It is true that the company's money has helped the nearby communities in real ways. But the true goal was to improve the political standing of the company. (Remember that we were evaluating what was accomplished for the local communities, not the motivation of the funders.) The multimillion dollar effort had not defined its goals in relation to the good work that it actually did. Our first job was to help the company figure out what it wanted to accomplish. Then we could see if it achieved those goals.

A different type of confusing goal emerged in a drug treatment program we evaluated. The program gave the following statement as proof of its successful results: "all of the 'clients' met with Reverend Bob." We politely asked if they had any other measures of success, such as "clean" urine samples from program graduates, return-to-work data, criminal or arrest data. They were not pleased with our suggestion that a meeting with the Reverend Bob would be insufficient to set the "clients" on the road to recovery. Needless to say, while it is possible that Reverend Bob's influence may have been profoundly effective, few outside funders would be impressed with the program's suggested measure of efficacy.

Sometimes a goal is clearly stated and very measurable but of questionable worth. We evaluated a housing assistance program that listed its own sustained existence as its primary goal. While the stated goal was delightfully honest, one expects an agency of this type to proffer goals such as "help the people of this area buy low-cost homes" or "work with local people and area banks to help first-time, low-income home buyers." The agency felt its work was so important, however, that its first responsibility was to ensure its own survival.

We shall return to the task of evaluating programs' goals in our discussion of outcome-based evaluations and impact assessments.

Outcome-Based Evaluation and Impact Assessment

The previous discussion focused on object-based evaluations—where the evaluation starts from the stated goals of the project or agency. Outcome-based evaluations and impact assessments are similar to the object-based evaluations we have just reviewed but with a notable difference: with outcomes-based evaluations or impact assessments the evaluator—not the program administrator—determines the measures to reflect what the program has accomplished. The evaluator might start with the program's stated goals, but then add other outcomes or "impacts." Some examples:

- The program's stated goals are to improve reading and math scores at a local school. In conducting an evaluation, however, the evaluator realizes it is also important to measure school attendance, scores in other subjects (perhaps subjects dependent on reading and math skills), and overall student satisfaction with school.
- The agency seeks to educate residents about energy costs and usage. The evaluator adds measures of *actual* energy usage and of the types of appliances the residents buy.

Joel Telles, Ph.D., is Vice President, Information Services and Research, at the Delaware Valley Healthcare Council of the Hospital and Healthsystem Association of Pennsylvania. The Healthcare Council is a trade association of hospitals and other healthcare providers in the Philadelphia area. Dr. Telles, who has his doctorate in sociology from Columbia University, has been on the faculty of UCLA and the University of Virginia, and lectures in the Temple University Health Administration Department.

Dr. Telles's work covers a range of activities, including: assembling and reporting large scale databases on healthcare, data analysis to support advocacy initiatives, pure research studies, coordinating data collection and reporting to improve quality of care, project management, direct advocacy and coalition building, data and statistical consultation to members, and internal information systems management. The validity of his studies and data analyses are subject to careful evaluation different from the peer review of academic research. Results must be able to survive scrutiny by opponents in a political discussion. Lawsuits (unsuccessful) brought by groups unhappy with results lead to the need to be sure that the design, conduct and report of the study is not merely intellectually, but legally, defensible. The data and research are used for strategic planning decisions by members who stand to lose or gain large sums of money and affect the healthcare access and quality of many persons. Studies are taken by Council staff to meetings with legislators, regional administrators, state and federal administrators, and the like. For some studies, even information about who participated—without any results—can affect major decisions.

While it is important that reporting is accurate and data not be reported selectively, the implications of the results for the membership must be kept in mind. When possible, the data and analysis reported in a paper are separated from the characterization of the results in advocacy activities. All data and analyses must be reported bearing in mind that others routinely take the data and report it out of context. The changed information is often attributed to the Healthcare Council, since that provides legitimacy to the data. Dr. Telles often receives calls about results he can barely recognize from what he reported. A final difference in the way studies are used is that they often are the basis for subsequent research and change the behavior and processes of those collecting new information. A form Telles created to collect information on non-payment by managed care organizations was adopted by hospitals as the official data log. A survey of the ways in which hospitals do and do not provide easy access to healthcare for Latinos was used by hospitals as a self assessment tool for improving access. A survey of services available for caring for stroke patients currently underway is expected to bring to the attention of the hospitals the guidlines for good stroke care.

FIGURE 12.3. Portraits and stories from real evaluators.

- The organization says that former drug users who meet with Reverend Bob will be better and wiser people, thus leading to less drug use. The evaluator decides that the evaluation should also measure subsequent drug use via arrest records, ER records, interviews with the former drug users, and job records.[3] Sometimes, as we have seen, the evaluator might find the stated goals neither sensible nor measurable.

[3]Note that an outcome-based evaluation or impact assessment may be indistinguishable from an object-based evaluation because the list of stated goals includes the results the evaluator wishes to assess.

In almost any evaluation, the evaluator is seeking to determine if the intervention is making a difference, if that difference is useful, if the cost is reasonable, and if the people being helped are the ones who need the help? These questions are more formally and fully enumerated by Rossi *et al.* (1999), from which the following list is very loosely adapted:

- Do the goals or outcome measures make any sense?
- Can you understand the stated goals?
- Do the goals or outcomes serve the needs of the target population?
- Do the goals meet professional standards?
- Are the goals in line with customary practice and/or norms for other programs?
- Do the goals conform to legal requirements?
- Do the goals or outcomes conform to ethical or moral standards? How do they address issues of social justice and equity?
- How do the goals or outcomes compare with historical data or past performance?
- Were these goals set by administrators, by stakeholders, or by some combination of these groups?
- Were there other stakeholders who should have been involved in goal settings?
- Have experts been involved or consulted in setting these goals?
- Does anyone have, or can we ascertain, preintervention baseline levels for the target population?
- Can the outcome or impact measures be causally related to the program's work?
- What would the conditions be in the absence of this program (counterfactual)?

An Example

In outcomes-based evaluations or impact assessments the evaluator seeks to understand the relationship between the project's efforts and what has happened to the target population and others who might be affected. Thinking about the list above, what would you evaluate if you were evaluating a program that was trying to reduce highway deaths associated with speeding?

Clearly this is a useful goal with wide moral acceptance, seemingly few legal obstacles or professional conflicts. While drivers are obviously the target population, most people would benefit from fewer highway deaths. If the public outreach effort were directed to as many of the states as wanted it, you would limit concerns of inequity. You would have historic data (traffic fatalities per miles driven before the new outreach effort) and you would have similar data from the period after the campaign had been in effect long enough to have some impact. Because you have a consistent numerator and denominator (traffic fatalities per miles driven), you would not have as many worries about changes in driving patterns.

Let us look at that list of evaluation concerns again and consider more of the issues.

Who set the goals and which stakeholders, if any, were involved? It turns out that drivers in the western United States argue that speed limits appropriate for eastern states are foolish in western states, where there are often straight roads, less traffic, and longer distances. Any law that slows down drivers, they say, actually increases accidents.

Who is the target population? As noted, drivers are clearly the target population. But targeting drivers may not be the most effective way of influencing them. It might be more effective to target children, who are impressionable, often in confined spaces where they are forced to listen to people in authority (e.g., school), and are delighted to speak to parents with great moral force. Also, because young children have the most number of years to lose in a fatal car accident, their voices may carry unexpected weight.

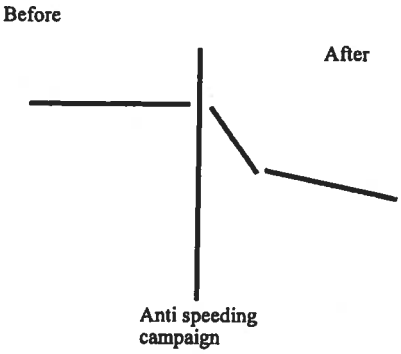

FIGURE 12.4. Traffic fatalities before and after antispeeding campaign.

Can the outcome or impact measures be causally related to the program's work? What would the conditions be in the absence of this program? Figure 12.4 appears to support the efficacy of the antispeeding campaign. The numbers seldom look better than this: steady before the campaign, and clearly going down after the campaign.

But was the reduction in speeding actually caused by the political efforts? It is entirely possible that the campaign *reflected* a public outcry about traffic fatalities. In other words, the antispeeding campaign was the political response to a popular movement, and not a cause of the reduction in foolish driving. Schematically, this relationship would be represented in Figure 12.5, where we see the concern about fatalities "created" both the movement toward safer driving and the antispeeding campaign. Causation is unclear. How important was the antispeeding campaign in reducing the traffic fatalities? Perhaps both change in driving and the campaign were "caused" by responses to fatality increases.

Ideally, an evaluator has comparison groups (control groups) that were not exposed to the intervention (in this case the antispeeding campaign). Of course, in societywide programs, finding a pure control group may be impossible. But this antispeeding effort provides us with states that did not participate in the effort. There will be some leakage, or overlap, from states where newspaper advertisements and TV programs reach some of he citizens in nonparticipating states, but compromises and next-best solutions are often required in evaluation research. Figure 12.6 would represent the "control" group's driving behavior (the states without the antispeeding campaign) in relation to the "experimental" or "treatment" group (the states with the antispeeding campaign). It would provide additional support: a comparison group that was not exposed to the campaign and that does not show a decrease.

Thus, we have the classic research design where we can measure average traffic speeds, traffic accidents, and traffic fatalities before and after the campaign in the treatment states

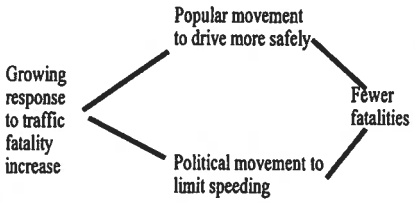

FIGURE 12.5. Traffic fatalities before and after antispeeding campaign.

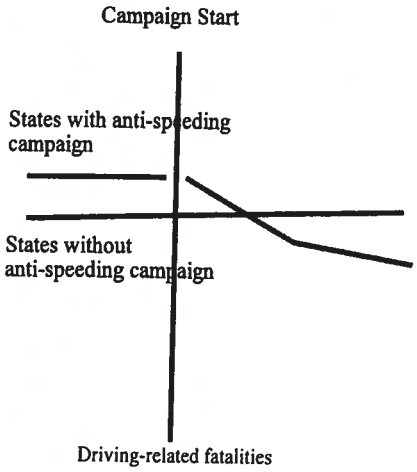

FIGURE 12.6. Traffic fatalities before and after antispeeding campaign.

(those participating in the campaign) and compare those with the control group states (those not participating in the campaign). If the campaign is effective, we should see a reduction in average traffic speeds, traffic accidents, and traffic fatalities in the states where there was an antispeeding campaign that is greater than in the other states.

Summary so far: This example is on a broader scale (societywide) than many evaluations but illustrates many of the issues involved in outcome evaluations or impact assessments. In any evaluation that focuses on the program's results, the evaluator addresses eight basic questions that serve as a useful summary of this subsection:

1. Is the intervention (could also be called policy, program, agency, organization) directed at the appropriate population?
2. Is the intervention reaching the target population?
3. Is it being administered or implemented in the ways designed and intended?
4. What aspects of the implementation are defeating the success of the effort?
5. Does it work?
6. For whom does it work?
7. How much does it cost?
8. What are its costs relative to its effectiveness?

Only one or two of these questions (No. 3 and maybe No. 4) address the *process* by which the program or project achieves its results. But examination of the project's internal actions and decisions—the process—is an area that deserves separate analysis. It is to "process evaluation" that we now turn.

Process Evaluation and Project Monitoring

Evaluation researchers can investigate factors other than outcomes or impacts of programs, agencies, or organizations. We often examine the *processes*—the ongoing actions, events, decisions, and rules *within* the programs, agencies, or organizations. This is called *process evaluation* or *project monitoring*. Examples of this type of evaluation research are:

- Eamining how clients are treated when they come for job training programs (e.g., how long it takes to complete paperwork; how many visits are required to determine skill needs and set up training or courses; friendliness of staff to clients; how convenient are the job referrals for graduates).

  Note that this is *not* an evaluation of how many graduates found jobs or of which type of clients found jobs.
- A nutrition education campaign distributes posters and recipes that promote healthy eating. A process evaluation might consider the locations of the posters and recipe distribution points: Were they where the target population lives or travels or shops? Were the posters and recipes attractive, culturally sensitive, and tasty?

  An outcome or impact evaluation would want to know if the area residents bought healthier foods and if the posters and recipes influenced people to buy and eat healthier foods.
- A school system is operating a reading improvement program for third graders. As part of the process evaluation or program monitoring effort, the evaluator sits in (observes) the reading improvement classes, interviews the students about their levels of interest and enjoyment, logs (keeps a systematic record of) attendance, talks with the teachers about their plans and aspirations, and interviews the program's administrators and designers.

  An outcome or impact assessment would focus on improvement in reading scores among those students who were in the program versus the students not in the program.
- An evaluator of a teen-focused antidrug hotline listens to calls, monitors staff training, and reviews phone logs. He or she considers the interactions between staff and callers, how staff are prepared for their work, how many rings before the line is answered, and the quality of recommendations and referrals.

  An outcome or impact assessment would probably examine the number of local teens arrested, surveys of teen drug use, school reports, police reports, and emergency room records of drug overdoses.

Les Baxter is a senior staff member in the Planning and Evaluation Department of the Pew Charitable Trusts, a large philanthropy based in Philadelphia. The Trusts make extensive use of strategic planning and evaluation activities to strengthen the development of grant programs and inform the decision-making of program staff, senior management and the board. Dr. Baxter and his colleagues are responsible for supporting programmamtic strategic planning and designing and managing assessments of grantmaking for all of the Trusts' program areas. On the planning side, he provides guidance to program staff as they develop grantmaking goals and strategies, track progress toward those goals, and report annually on progress to both the Trusts' management and the board. In evaluation, Dr. Baxter assesses the evaluation needs of major grantmaking areas, designs and develops integrated evaluation approaches to assess the progress and performance of complex multi-year grant portfolios, and then designs and manages all of the evaluation activities required to implement those evaluation approaches. The evaluations use a variety of methods, some of which are at the leading edge of evaluation practice. His recent projects inlcude assessments of the Trusts' grantmaking in climate change, marine protection, and sustainable forestry.

FIGURE 12.7. Portraits and stories from real evaluators.

As must be evident from this discussion, process evaluation or project monitoring focuses on what is happening within the organization, agency, or program. This is not to say, however, that process evaluations ignore the outcomes of the efforts. On the contrary, it is through our understanding of, and improvement of, the workings of the programs and interventions that we can best help them achieve their desired goals. One cannot just "will" better service, more clients served, fewer poor, more literate children. The staff, rules, and actions of the programs are the "agents" of the desired changes. Evaluating and monitoring the programs' processes can help determine *why* the project is succeeding or failing to succeed. Thus, analyses of programs' processes are frequently the most effective way of helping programs accomplish their goals.

Note that process evaluations frequently include structured interviews with administrators and clients. This can be an opportunity to learn more about the program's design, changes, and constraints. It is also an opportunity to provide feedback to the program if that is part of the evaluation plan (see discussion of *summative versus formative* evaluations, below).

Not one or the other. Evaluations often combine research on outcome and on process. Usually, combining both foci is the most effective way of understanding a program and its results. There is no conflict between the two forms of evaluation; they are complementary.

Naturalistic Investigation

To the uninitiated, naturalistic investigation is thought to be just hanging around and watching what goes on. Needless to say, this is a misinterpretation. Naturalistic investigation covers a range of techniques of which seemingly casual observation is one. But this method, if properly conducted, requires exquisite attention to detail, an excellent understanding of the program's history and theoretic underpinnings, and a lot of time. It is closer to ethnography than other forms of evaluation research, where the evaluator often seeks to view the program from the perspective of the client or target audience. The skills for this orientation involve the ability to closely interview participants without the clumsiness usually associated with formal interview protocols. Moreover, because the evaluator discusses his or her observations with the participants, mutual learning is enhanced.

In or own evaluation experiences we observed:

- A nurse fail to check for drug reactions in a semicomatose elderly woman, by saying "You're not allergic to penicillin, are you dear? Take this."
- A social service agency require clients to reapply for the same job training programs seven times in seven days (for no reason we could understand).
- A computer training system for nurses that told them to ignore the built-in software safeguards unless the supervisor was watching.
- A multimillion dollar welfare-to-work evaluation effort where the case workers confused the control and treatment groups, thus rendering evaluation meaningless. The case workers were not sure who was to be subjected to some newly proposed restrictions and who was allowed to continue with the old regulations. Because of the confusion, the state was unable to determine the impact of the new restrictions on clients' behaviors.

In all of these situations, the observations and informal discussions provided insights that would not have been possible from formal interviews or analyses of outcome data.

A few more points about naturalistic investigation:

1. Some students are occasionally drawn to this method because it seldom involves complicated math or statistics. This is a fool's solution. It is often much easier to collect and present statistics than to make penetrating observations and insightful analysis. With the numbers, you have got a recognized method and somewhat predictable measurement concerns. With ethnography, you are alone with the quality of your observations, understanding, and thinking.
2. The amount of time in preparation, and then observation and interviews, can be very demanding.
3. Without an easily explained methodology or formal statistics, the acceptance of naturalistic investigation-based evaluations is dependent on the faith and trust in the evaluator. One does not have the usual guarantees from reanalysis of questionnaires or data, log books, or test scores. The method lacks the classic protection of *replication.*
4. Naturalistic investigation methods have received support from postmodern and feminist scholars who have renewed our basic questions of epistemology and have enhanced our appreciation of multiple research methods.
5. We would add that one evaluation method does not exclude the other. Evaluations are usually best and most useful when they combine differing methods, e.g., process evaluation, outcome evaluation, and naturalistic investigations.

Expert

In our preview to this chapter we talked about such experts as movie critics, restaurant reviewers, and automobile test drivers. Just as we use experts in everyday life, we can use experts in evaluation of programs. Such experts should be very familiar with the type of program being evaluated, perhaps having administered similar programs in their careers. Many complex organizations rely on experts to conduct evaluations. Colleges are evaluated by teams of professors and deans from other schools. They review curricula, hiring practices, faculty publications, student opinions, graduation rates, postgraduation placements, administrators' plans and visions. They may sit in on classes and just walk around. Although this method is called "expert" evaluation, it obviously integrates many of the approaches discussed above.

Hospitals are also evaluated by peer reviewers in a process similar to college reviews. Hospital reviewers walk around but also collect data on infection rates, unexpected illnesses and deaths, number of operations, experience of the clinical teams and so forth. Again, what is called an expert evaluation is really a combination of methods filtered ultimately through the decisions of experts.

Sometimes "expert" evaluations involved competing experts, almost like a courtroom setting. In these cases, often associated with regulatory proceedings, each side has its own experts. An electric company, for example, may have its experts testify about their research on the need for another power plant, while a citizens' group or the regulatory agency's staff may testify on the ability of the region to rely on conservation. Both groups may use economists, survey researchers, demographers, and engineers.

Expert evaluations can be efficient because the evaluators start out with a full understanding of the issues, types of organizations, and constraints. As we have seen, however, they often incorporate many of the evaluation research approaches.

Summative versus Formative. Years ago, evaluation texts spent considerable energy distinguishing between the terms, *summative* and *formative*. *Summative* refers to an evaluation that is intended to give a judgment about the program—to sum up its outcomes or results; a summation of project accomplishments. *Formative* refers to an evaluation that provides information back to the program or policy leaders to help improve, or form, the intervention. We realize now that the dichotomy between summative and formative is usually unnecessary.[4] Increasingly, evaluators seek to incorporate both functions: studying, collecting information, and providing feedback to the project or organization.

More than Yes or No. Evaluation research is dreadfully underused if it only gives a thumbs up or thumbs down judgment on project outcome. Evaluation research can help guide programs and policies with subtle and nuanced insights based on analysis of the findings. Rather than "yes" or "no," evaluators can say, for example, that an intervention is most effective for those with more than an eleventh grade education, for women, for those who completed at least one training course but less than three training courses, for those communities that already have traffic lights, or complex patterns of variables. Evaluators, as we have repeatedly indicated, are also not limited to one methodological approach, but can combine insights and findings from several forms of inquiry.

RESEARCH DESIGNS AND METHODS

In the course of our discussion we have already reviewed several of the research methods and research designs used by evaluators. you will not be surprised to learn that most of the research tools of social science are used in evaluations. Evaluations can employ mail or phone surveys, long face-to-face interviews, observations, logs of transactions, analysis of existing data, and so on. While evaluation researchers often face constraints not encountered by those who can control all aspects of the research context, we have tried to develop new strategies to address, at least partially, these additional constraints.

Below, we provide a short overview of the most common designs. Students who are to conduct evaluations or who wish more information should examine texts devoted to research methods and/or to books on evaluation research. See, for example, Babbie (1998), Rossi *et al.* (1999), Schutt (1999), or Weiss (1998).

Random Assignment to Control and Treatment Groups

The gold standard of research design is the randomized assignment of participants into a control group and a treatment group. This allows researchers to make powerful comparisons linked to the "treatment" (or intervention or program)—where we know that most other factors should be the same. Often, evaluators can use this design in large and small projects. Consider these examples:

- Clients coming to a job training center are randomly assigned to one or the other program.
- People in 25 randomly chosen states receive public health advertisements but the people in the remaining 25 states do not.
- The city institutes new parking regulations for all odd-numbered streets.

[4]Often, legal or financial requirements demand that these functions be separated.

Joan I. Biddle earned her Ph.D. in Sociology at Boston University. She has recently retired from the United States Army Reserve, and holds the rank of Lieutenant Colonel. In addition to her interests in military topics, Dr. Biddle is an adjunct at the New School University where she teaches sociology courses on-line. She also writes about social interaction in cyberspace, distance learning, and family/women. LTC Biddle, Ph.D., is a long-time adjunct at the U.S. Army Command and General Staff College, School of Corresponding Studies, at Fort Leavenworth, Kansas. Dr. Biddle relates this vignette about evaluation in the Army:

Re-enlistment of first-term soldiers, and "lifers," is very important to the survival of the Army. Each unit in the Army has specific numerical goals for the re-enlistment of its soldiers. Meeting these quotas is the ultimate responsibility of the Commander.

A few years ago, Dr. Biddle had an opportunity to do a short study about re-enlistment in one of the Brigades of a large, branch school training post. The Brigade Commander was convinced that there was a "problem" meeting re-enlistment quotes. He had been advised about this problem by the Brigade Command Sergeant Major. At the time, Dr. Biddle was a Major on tour at the post, and the "re-enlistment problem" had been an issue that no one else had time to tackle. The Brigade Executive Officer noted with some interest that Dr. Biddle was a sociologist.

The "problem" presented to Dr. Biddle was from the perspective of the Commander and included his guidance for what he wanted to know. The result was a study comprised mostly of interviews with everyone connected with re-enlistment. Additionally, as we shall see, critically, Dr. Biddle examined: re-enlistment statistics; the projected goals/quotas for re-enlistment in each level of the Brigade; and the Brigade's next higher level for statistics, the Post level, of which the Brigade statistics were a portion.

Her findings differed from what the Brigade Commander had expected: a "problem" meeting re-enlistment quotas. In her written report and discussions she explained that in some units there was sometimes a discrepancy in meeting the quarterly quotas, yet, the annual quotas for the units were always met because losses in one quarter was mostly offset by gains in another quarter, bringing the total numbers around to meet, or even exceed, the annual goal. Her numerical analysis revealed that when all of the numbers for the Company, Batallion, and Brigade levels were examined in aggregate form, there really was no "problem" meeting the re-enlistment goals for the Brigade. There was then, a perception that a problem existed, based upon faulty conclusions reached from an examination of data that had been extracted out of their total context, rather than from an analysis of the data as part of a larger data set.

FIGURE 12.8. Portraits and stories from real evaluators.

You can probably see the difficulty in this arrangement. Many people find it hard to *purposely and arbitrarily* "deny" help to half of those in need or half of those who could benefit just because some eggheaded evaluator says it will provide information on which to make better decisions. Those who believe their public health effort is beneficial, those who think their training program can bring people out of poverty, and those who want all of the city to have better traffic patterns may not want to cooperate with the research design. In addition, program personnel who should randomly assign clients can easily defeat the evaluation research design

for what are viewed as the most noble of reasons. More, the undermining of the design can be subtle, where the mixing (the "contamination") of the "control" and "treatment" groups is not known. For example, a caseworker wants to help those who have had the most problems or a counselor wants to help those who show the most promise for improvement. These are research nightmares. But there are several defenses:

1. Fully explain the design and intention of the research plan to all staff. Explain that we do not know if the new intervention is really effective, and for whom it may be more or less effective, or if it works at all. Only by comparing the new intervention with the old intervention in a structured setting can we determine its value.
2. If this evaluation shows the new intervention is really more effective than the old, the findings from this analysis will help us spread the word and help us receive more funding.
3. We are not denying services to the "control" group; we are just giving them the older, established service or product which may work much better than the new product or service. (This is an important ethical point that is often lost.)
4. The evaluation researchers can help structure the program's operations so that assignment into control or treatment groups is automatic, e.g., a computer program assigns the cases, or you can created "systematic assignment" where every other person is sent to the control group or treatment group. (Note: this latter method is not true randomization but is called *near randomization*, and will often suffice.)
5. In many cases, the assignment to one or the other group can be made where neither the staff nor the client knows which group they are in. This is called *double blind* and is standard methodology in medical and pharmacological research.

Alternative Approaches

With these obstacles to gaining cooperation, evaluators can only occasionally rely on research designs as rigorous as random assignment to control versus treatment groups. Usually, evaluations must make due with alternative approaches. Consider these common research designs:

1. *Constructing control groups by "matching" samples.* The evaluator finds "similar" people in the group that received the "treatment" and in the group that did not (the control). Then the evaluator compares the outcomes of the two groups. Of course, the evaluator cannot be certain that he or she has matched on the correct variables or that his or her measurement of those variables was accurate.
2. *Using statistical controls to "match" groups.* Here the researcher "statistically adjusts" for each of the nontreatment factors that may explain outcome differences. For example, if more people in the control group are older, and if older people do not do as well on the outcome measure, then the researcher "statistically adjusts" the numbers to account for the differences between the two groups.
3. *Using "historical" controls.* The program or agency may have years of experience and records of service before trying the new "treatment" or intervention. The researcher uses the previous data on people served and on outcomes as a basis to compare with the new method. Researchers, obviously, must try to ascertain that the previous groups (historical groups) did not strongly differ on other factors and that the environment is similar. The

research could combine these historical controls with statistical adjustments and "matching" to make the comparison more meaningful.

4. *Time series or interrupted time series.* In the earlier example on reducing traffic fatalities via an antispeeding campaign, we illustrate a time series design. That is, automobile drivers in each state generated data on miles driven, and on fatalities. Then, we "interrupted" the series via our antispeeding campaign in half of the states. By examining the outcome data, we can see the change in the time series and assess the effect of the intervention.

5. *Pre–post design.* For most of life we do not have a control group. If we go on a diet, we seldom have twin siblings who promise to not change their eating patterns. Yet, somehow, we seem to determine the effect of the diet on our bodies. Much evaluation research is similar to the non-control group design, or what is also called the *pre–post design.* We have the "pre," or pretest measure (we weigh ourselves before the diet), and we have the "post," or posttest measure (we weigh ourselves after the diet). The problem with social reality and social programs is that they are much messier to understand than just one person on a diet. Programs and interventions, as we have argued, are inevitably embedded in complex history, culture, and shifting environments. Did the target population change between the pre- and posttests? Did other factors or policies affect the clients? Did they just grow older, wiser, or bored? did the pretest sensitize them to the researchers' concerns and thus bias their responses? These are some of the questions that cannot be answered via the single pre–post design and that haunt researchers.[5] However, just as we can think about several issues in our own diets (e.g., did we eat the same but exercise more; did we eat more salads and less cake?) we can examine factors that we think may have influenced the results. We could, for example, look at the demographic information on the target population for people measured at the pretest stage and the posttest stage. Were they the same? Who dropped out? Can this help explain the findings?

Many years ago we evaluated an adult reading course. The results were extraordinary. Students, on average, raised their reading grade levels three years in nine weeks. These adult students went from an average reading level of third graders to an average of sixth grade reading levels. When we examined the intermediate tests and the characteristics of the dropouts, however, we realized that everyone with problems in the course quite before the posttest. The course did a fine job, but some of that positive result was due to the fact that the slower or less interested students were not included in the final scores. Fortunately, we were able to discern both the excellent work of the adult education program and the effects of the dropouts—a double finding that could be addressed in future efforts (e.g., give special help or encouragement for those who are not doing well).

Natural experiments. Evaluators can use naturally occurring events to examine reactions to changes. A college may change the requirements to demand three years of literature courses for all students; and evaluators could see if this change results in improved writing. A community could get a new road. Evaluators could examine how this affects sales by local stores or access to movies and drugs. About 25 years ago there was a large electric power blackout in New York City. Nine months later the maternity wards were especially busy. Everyone thought the cause in the birth rate spike was due the lack of lights: people were left in the dark with only each other. Interviews with women in the maternity wards revealed a different cause for the increased procreation: no TV. Late night TV, it appeared, serves many functions not anticipated by demographers or, probably, Planned Parenthood.

[5]See the classic work on these and other threats to validity in Campbell and Stanley's 1996 book, *Experimental and Quasi-Experimental Designs for Research.*

Natural experiments also occur in policy and program settings. A state may only have enough training dollars to provide services to half the population. Two similar cities may take different approaches toward informing their citizens about water conservation. One state raises the drinking age from 18 to 21 and the contiguous state leaves it at 18. In cases like these, the evaluator "takes advantage" of the events he or she did not arrange to examine the effects of the changes and differences.

In any evaluation, it is essential that the evaluator create a clear research protocol—a detailed guide to data collection and research procedures. This protocol can be changed as the evaluation proceeds, but it should be available to all interested parties for review and as a written record of plans and actions.

Just an Overview

In this one chapter, we cannot cover all research designs or the dangers to validity and misinterpretation. We wanted, instead, to suggest the extraordinary range of possibilities in designing research and in building on the thoughtfulness and creativity of previous scholars. Research design is like an ever-growing puzzle that challenges the evaluator in the most exciting ways. How can the research be structured to allow the least bias and/or most opportunity to measure whatever bias might be present? What are the sources of "contamination" that may confuse my findings? What prevents me from generalizing from the findings of this evaluation to similar interventions?

When evaluators are involved in the intervention from the start, they can often suggest ways of structuring the program or agency to facilitate the evaluation, provide better research designs, and allow collection of more useful data. sometimes, even when the evaluator is involved from the start, issues of cost, politics, time, ethics, law, or departmental conflict will limit the best evaluation designs. But even knowing these limitations can help the evaluator better understand the intervention. An appreciation of the dynamics of the program and its constraints will help the evaluator explain the evaluation purpose and gain the cooperation of participants.

Programs of course do not exist for the convenience of the evaluator. The best laid plans can and probably will go astray, sometimes because of lack of cooperation and sometimes because of forces outside the program or policy. No social policy or intervention occurs in isolation. History, economics, previous efforts, agency connections, and politics all affect social programs and policies. Evaluation of any social policy or intervention is likewise embedded in this rich context. And it is the context of evaluations to which we now turn.

EVALUATIONS IN CONTEXT

We have suggested that evaluations can be tools of improvement, insight, redirection, refunding, and also of destruction. While most reasons for evaluation are laudatory, one is wise to question the motivation of any request for evaluation. Few projects or policies are evaluated to increase the sum of human knowledge; fewer because program administrators sincerely believe an independent assessment would help them understand and improve their program. Usually, an evaluation is conducted because legislation, regulators, or a funder demand it.

Let us consider what happens when an evaluation is required.

While working for the Internal Revevue Service, Bob Dotzler completed his Ph.D. in sociology at George Washington University specializing in the area of work and organization. At the IRS he worked in various areas including human resource development, performance analysis, work systems design, re-engineering, and Total Quality Management. Evaluation was the skill area that cut across all his work. The strong methodological training he received opened up a wide range of opportunities for participating in key agency evaluation projects. For example, he was asked to provide methodological support to the IRS Commissioner's Study Group on Women and Minorities where he worked on developing and analyzing a large survey sent to 15,000 managers designed to identify internal agency barriers to the advancement of women and minorities. At one point he was assigned responsibility for developing agency guidelines for evaluating all agency training programs.

Throughout his career at the IRS, Dotzler found his evaluation skills to be constantly in demand. Few of the agency analysts with tax backgrounds possessed professional level skills in questionnaire design, measurement, item construction, or data analysis. Similarly, few brought a theoretical perspective to the analysis of organizational problems and organizational change. As a consequence, Dr. Dotzler found himself working on a wide range of evaluation-related projects, e.g., a bi-annual all-employee survey focused on measuring employee morale and job satisfaction, re-engineering procedures, and performance appraisal systems. Bob Dotzler believes the rigor of his sociological training helped advance his professional career in government as well as adding to the quality of agency surveys, procedures, and processes. He is particularly proud of bringing methodological integrity to internal agency research which can and is often challenged by other organizational interests.

FIGURE 12.9. Portraits and stories from real evaluators.

Selecting an Evaluator

If the funder or sponsoring business or government agency selects the evaluator, the issue is generally moot for the program or policy being evaluated. They get whomever is assigned. As we have suggested, however, knowledge of what evaluation research should be will help program leaders understand the needs and aims of an evaluation effort. On the odd chance the data collection or research effort is inappropriate or incomplete, program personnel can discuss this with the sponsor or funder.

More probably, program personnel are involved in selecting the evaluator. Here, obviously, one starts with a search for a combination of experience and expertise. But the real work is in finding an evaluator with whom the program leaders reach a mutual understanding of, and ability to address, these questions:

1. What is the program seeking to accomplish with this evaluation?
   There may well be multiple aims, e.g., justify the program's existence, show its good works, show that it is better than it used to be, impress a local politician.
2. What resources are available for evaluation?
3. If there are needs for immediate feedback from the evaluation, how will this be accomplished?
4. How can disruption of operations be minimized; how much disruption can be tolerated?

5. What biases will be introduced by the evaluation process?
6. How will ethical concerns be addressed? This includes:
   a. Treatment of confidential information
   b. Full presentations to relevant ethical review boards/internal review boards
   c. Sensitive and respectful interactions with clients, staff, and other stakeholders
   d. Long-term care of data collected as part of the evaluation
   e. Care about how information is presented in discussions and in the evaluation report
7. What will be done with the evaluation findings? That is, what is the political and social context of the project within which the evaluation will be a part?

Internal versus External Evaluator

Many funders and government agencies demand outside evaluators (an evaluator not from the agency or program). Often, an outside evaluator is the only way of collecting meaningful information when a powerful director would intimidate, or restrict data access from, an employee of the organization. Outside evaluation reports are generally more respected than reports prepared "in-house." On the other hand, internal evaluators are often very familiar with the agency's operations, know how to efficiently and effectively get the data, have a better idea of "where the bodies are buried," and have a contextual knowledge that requires years of experience.

Evaluation research is both a learning and a teaching experience. The evaluator learns about the program and teaches the program staff about the logic of the evaluation in addition to explaining the specific findings. The program staff help the evaluator better understand her or his own findings by putting them in a richer perspective than she or he would otherwise have. The oft-pictured image of the evaluator as inspector general is almost always wrong. Few evaluations involve just marching in, looking around, examining some data, and marching out while dropping hints about the report to headquarters. The more accurate metaphor is one of a relationship. Both benefit from increased cooperation and understanding. Selecting an evaluator should be a considered and thoughtful process.

The Audiences for Evaluation Reports

1. *Funders or authorizing sponsors.* The first audience is usually the folks who paid for the evaluation: the sponsoring government agency, business, or funder. The funders or authorizing agencies can sometimes end the program, replace it or its leaders, reduce or restructure its functions, or increase its resources. They will probably read the evaluation report.

2. *Program leaders.* The next audience—and the one we tend to think of as most essential—is the leaders of the program being evaluated. These are the people who can most immediately implement the evaluation's insights and recommendations. They can directly influence the rules, allocations, and policies.

The next three audiences are not hierarchically ranked.

3. *Clients or those directly affected by the program or policy.* The clients–those receiving the training, those being helped by the service for shut-ins, those learning how to handle their capital gains, those city commissioners receiving millions of dollars in block grants, and those treated at the homeless shelters—all have good reason to know about the evaluation findings of programs serving them. As one can guess from this list, however, only some of the groups are likely to have access to the evaluation reports.

4. *Staff of the agencies, programs, or interventions.* While leaders of agencies and programs set the rules and hire the staff, it is the staff of these organizations that implement the programs. It is often the staff who are being evaluated. The evaluation report can be a useful tool for improvement and change. It can also be used to foment managerial change. Evaluation reports are seldom distributed to all staff members.

5. *Other stakeholders.* There are others whose lives can be seriously affected by evaluation reports and their recommendations. Is the homeless shelter to be placed on your block? Will the job training program help your business find good workers? How will the new park design affect your daily walk or your real estate? How will new TV ratings affect your kids? Will money you thought would go to your program go to another program instead? Does the agency's board of directors want to reevaluate its executive director? Evaluation reports are often consequential in ways that even evaluators do not realize.

6. *Everyone else.* Earlier, we talked about evaluation research as a way of reducing or eliminating programs and policies that some political groups disliked. That use still exists. Increased use and acceptance of evaluation research gives evaluation reports more prominent roles in political and economic dialogue. Evaluation reports can be the basis for allocations of billions of dollars and determine the actions of hundreds of thousands of people. The use of evaluation reports on needle exchanges, welfare reform, early childhood education, and school choice are just four examples of evaluation as a fulcrum of political, social, and economic debate.

This last issue of evaluation's impact leads us to our final discussion: What happens with evaluation reports and recommendations?

If evaluation research is launched near troublesome social, historical, and political shoals, then evaluation findings are adrift in stormy waters far more ominous and unpredictable. Evaluation reports are written by people who are presumed to be knowledgeable about the programs being evaluated and about evaluation methods. But once a report is released, it may turn up in the hands of those without such knowledge or skills. Its interpretation of findings may be entirely distorted, the context of the research completely twisted. Let us look at what may happen to evaluation reports.

- The recommendations are carefully reviewed and relevant steps are taken. The project is improved in both its processes and its goals.
- Funders gain a better understanding of the goals and work. Funds are reallocated, perhaps with new emphases toward better policies and use of resources.
- Clients and community groups use the report to press for changes that make the agency more responsive to their needs.
- The report is filled with complex mathematical and research terms. Few read it. Some experts who can understand it indicate that the report suggests changes. Absent a broad consensus on what it says, however, few changes are made.
- Conflicting groups within a program or agency both use the report as "proof" that the other side is even more incompetent than they realized. Both groups become more entrenched in their positions.
- Despite a sensitive and well-researched evaluation, including a clear and seemingly convincing report, the program's leaders and the local community are convinced that the report is nonsense because it conflicts with their beliefs. The report is used to support a leaky air conditioner in the director's office.
- Political or social leaders in another region use the report as "proof" that a particular policy or program is dangerous and should be ended. The fact that the methodology and findings are largely ignored is of little consequence.

- The report is held up as proof of the wisdom of a well-supported and well-liked program. Careful reading of the report, however, reveals the program's effect is nonexistent.
- Most of the above—at the same time.

The first lesson from this list is that evaluation reports must be written with great care and clarity. The second lesson is that even if the report is well done, the evaluator (and the program or policy that was evaluated) should be prepared to explain the research and the recommendations. The third lesson is that of political and social wisdom: Documents will be used and abused as others wish. The uses of the information may be the flotsam and jetsam of others' arguments, often long ago tossed from the logic of careful methodology or context.

Why Do It?

Evaluaton research is like Churchill's comment about democracy: It is the worst imaginable system except for all the others. Evaluation research is the best way we have of systematically understanding and improving our efforts at helping people. Only by careful investigation, observation, and data collection can we ascertain if we are intentionally improving or worsening our society and the help we give to our most vulnerable members. Everyone has biases and prejudices. Good research design enables us to assess programs, agencies, and interventions in ways that eliminate or at least make manifest our biases and the measurement biases we introduce. People may differ about even good evaluation research findings, but those findings provide a rational basis for discussion, action, and improvement. We simply do not have a better way of understanding the complexities of social reality.

ACKNOWLEDGMENTS. The author thanks Meg Koppel, Eugene Sapadin, and Joel Telles for their thoughtful comments and suggestions. Any errors and omissions, however, are entirely the responsibility of the author.

REFERENCES

Babbie, E. (1998). *The practice of social research* (8th ed.). Belmont, CA: Wadsworth.

Campbell, D., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Rossi, P. H., Freeman, H. E., & Lipsey, M. W. (1999). *Evaluation: A systematic approach* (6th ed.). Thousand Oaks, CA: Sage.

Schutt, R. K. (1999). *Investigating the social world: The process and practice of research* (2nd ed.). Thousand Oaks, CA: Pine Forge Press.

Weiss, C. H. (1998). *Evaluation* (2nd ed.). Englewood Cliffs, NJ: Prentice–Hall.