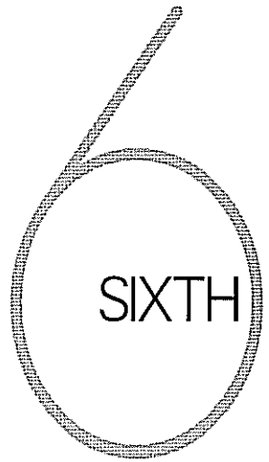
To the memory of *Donald T. Campbell*—
architect of modern evaluation theory and practice;
mentor, directly and indirectly, to all evaluators

PETER H. ROSSI
HOWARD E. FREEMAN
MARK W. LIPSEY

EVALUATION

A SYSTEMATIC APPROACH



SAGE Publications
International Educational and Professional Publisher
Thousand Oaks London New Delhi

For information:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
6 Bonhill Street
London EC2A 4PU
United Kingdom

SAGE Publications India Pvt. Ltd.
M-32 Market
Greater Kailash I
New Delhi 110 048 India

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Rossi, Peter Henry, 1921-
Evaluation: a systematic approach / by Peter H. Rossi, Howard E. Freeman, and Mark W. Lipsey. — 6th ed.
p. cm.
Includes bibliographical references and index.
ISBN 0-7619-0893-5 (acid-free paper)
1. Evaluation research (Social action programs). I. Freeman, Howard E. II. Lipsey, Mark W. III. Title.
H62.R666 1998
361.6'1'072—dc21 98-40244

This book is printed on acid-free paper.

02 03 04 10 9 8 7 6 5

Acquiring Editor: C. Deborah Laughton
Editorial Assistant: Eileen Carr
Production Editor: Astrid Virding
Designer/Typesetter: Janelle LeMaster
Cover Designer: Ravi Balasuriya

CONTENTS

PREFACE

Throughout the six editions of this book, its objectives have remained constant. It provides an introduction to the range of research activities used in appraising the design, implementation, and utility of social programs. That set of research procedures known as evaluation has become solidly incorporated into the routine activities of all levels of government throughout the world, into the operations of nongovernmental organizations, and into the discussions of social issues. We believe that evaluation research has influenced social policies and other efforts to improve the social conditions of the citizens of many communities. It is also an exciting professional role providing opportunities to advance social well-being along with the exercise of technical skills.

Evaluation: A Systematic Approach has strong ambitions to communicate the technical knowledge and collective experiences of practicing evaluators to those who might consider engaging in evaluation and to those who need to know what evaluation is all about. Our intended audiences are students, practitioners, novice social researchers, public officials, sponsors of social programs, social commentators, and the legendary intelligent layperson. Although some very experienced evaluators might find our book too elementary, we hope that reading it will help others at earlier stages of their encounters with evaluation. We also provide references to more advanced discussions of critical topics for those readers who want to pursue some topic in greater depth.

When Howard Freeman died suddenly and prematurely shortly before the fifth edition was published, I was quite sure that there would never be a sixth edition of which I would be a living coauthor. Howard Freeman had been a person of great knowledge, experience, and wit, working with whom was most of the reward of collaboration. Without Howard to work with, the prospect of yet another revision seemed painfully unrewarding.

However, it became increasingly clear after a few years had passed that the fifth edition was rapidly becoming out of date. Our noble Sage senior editor, C. Deborah Laughton, started to urge me to consider a major revision. At first I resisted on the grounds that I had no collaborator with whom to work. Working with great subtlety, stealth, and a bit of benign deception, she told Mark Lipsey that I was interested in having him as coauthor at the same time telling me that Mark was interested in becoming a coauthor. Knowing the high quality of Mark Lipsey's evaluation work, I was quite flattered at his interest. He tells me that he was also pleased that I was interested in working with him. And so a new collaboration was forged. Mark cannot replace Howard: He brings to the collaboration a different background and set of

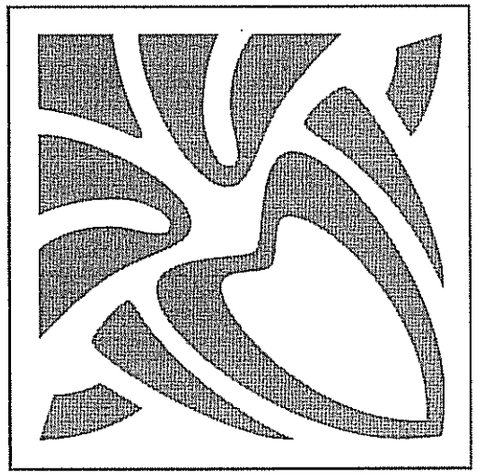
experiences, equal to Howard's in depth and sophistication but different in content in ways that enrich the sixth edition. Working with him has been a great pleasure, greatly enhanced by his sense of proportion and ready humor. I am grateful to C. Deborah Laughton for her skillful matchmaking and for the improvements in this edition that it made possible.

Most of the new material that appears in this revision is Mark Lipsey's contribution. The previous edition covered quite sketchily evaluation diagnostic procedures and how evaluations should be tailored to fit programs and social contexts. The current version has greatly expanded coverage of those topics, adding important detailed material on explicating program theory. The treatment in the first five chapters of this version carries the reader through a sequence that follows more closely the typical steps that the development of evaluations take. Lipsey has also updated the exhibits that appear in this edition, often with examples of evaluations currently under way or very recently completed.

We are grateful to the following reviewers for their comments: Jack McKillip, Ron Andersen, Melissa Jonson-Reid, David MacPhee, and William Shadish. A special acknowledgment is also extended to Kate Peterson for her extensive copyediting efforts.

We have dedicated this edition to the memory of Donald T. Campbell. In our view, there is no one who more deeply influenced the development of evaluation. Not only are his publications cited in all major works on evaluation, but he was also an important mentor. Several generations of his students now are leading figures in the field and their students are rising quickly to prominence. His influence on this volume can be seen in the discussion of impact assessment, which follows rather closely his exposition of research designs laid down in his 1966 publication (Campbell and Stanley, 1966). All evaluators can see farther because we stand on his shoulders: Campbell was a giant who gave us a lofty perspective on how social science can advance the improvement of the human condition.

—P.H.R.



KEY CONCEPTS FOR CHAPTER 1

Social program; social intervention	An organized, planned, and usually ongoing effort designed to ameliorate a social problem or improve social conditions.
Program evaluation	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.
Social research methods	Procedures for studying social behavior devised by social scientists that are based on systematic observation and logical rules for drawing inferences from those observations.
Comprehensive evaluation	An assessment of a social program that covers the need for the program, its design, implementation, impact, and efficiency.
Evaluation sponsor	The person(s), group, or organization that requests or requires the evaluation and provides the resources to conduct it.
Stakeholders	Individuals, groups, or organizations having a significant interest in how well a program functions, for instance, those with decision-making authority over it, funders and sponsors, administrators and personnel, and clients or intended beneficiaries.

CHAPTER 1

PROGRAMS, POLICIES, AND EVALUATIONS

This chapter introduces program evaluation as a robust arena of activity directed at collecting, analyzing, interpreting, and communicating information about the effectiveness of social programs undertaken for the purpose of improving social conditions. Evaluations are conducted for a variety of practical reasons: to aid in decisions concerning whether programs should be continued, improved, expanded, or curtailed; to assess the utility of new programs and initiatives; to increase the effectiveness of program management and administration; and to satisfy the accountability requirements of program sponsors. Evaluations also may contribute to substantive and methodological social science knowledge.

Understanding evaluation in contemporary context requires some appreciation of its history, its distinguishing concepts and purposes, and the inherent tensions and challenges that shape its practice. Program evaluation represents an adaptation of social research methods to the task of studying social intervention in its natural political and organizational circumstances so that sound judgments can be drawn about the need for intervention and the design, implementation, impact, and efficiency of programs that address that need. Individual evaluation studies, and the cumulation of knowledge from many such studies, can make a vital contribution to informed social action aimed at improving the human condition.

The principal purpose of program evaluation, therefore, is to provide valid findings about the effectiveness of social programs to those persons with responsibilities or interests related to their creation, continuation, or improvement.

Long before Sir Thomas More coined the word *utopia* in 1516, many persons had tried to envision a perfect world. That their aspirations, and ours, have not been realized is evident in the social problems and attendant personal problems that confront every country in the world. True, how we define social problems and estimate their scope and which problems are salient to us have changed over time with shifts in values and lifestyles. And it is equally true that communities, societies, and cultures differ widely in the attention they pay to particular problems. But now, as ever, to borrow from Charles Dickens, these are the

best of times for some of us and the worst of times for others.

Since antiquity, organized efforts have been undertaken to describe, understand, and ameliorate the defects in the human condition. This book is rooted in the tradition of scientific study of social problems—a tradition that has aspired to improve the quality of our physical and social environments and enhance our individual and collective well-being through the systematic creation and application of knowledge. Although the term *evaluation research* is a relatively recent invention, the activities that we will consider under this rubric are not. They can be traced to the very beginnings of modern science. Three centuries ago, as Cronbach and colleagues (1980) point out, Thomas Hobbes and his contemporaries endeavored to devise numerical measures to assess social conditions and identify the causes of mortality, morbidity, and social disorganization.

Even social experiments, the most technically challenging form of contemporary evaluation research, are hardly a recent invention. One of the earliest "social experiments" took place in the 1700s when a British ship's captain observed the lack of scurvy among sailors serving on the naval ships of Mediterranean countries. He noticed, too, that citrus fruit was part of their rations. Thereupon he made half his crew consume limes while the other half continued with their regular diet. Despite much grumbling among the crew in the "treatment" group, the experiment was a success—it showed that consuming limes prevented scurvy.

The good captain probably did not know that he was evaluating a demonstration project nor did he likely have an explicit *impact theory* (a term we will discuss later), namely, that scurvy is a consequence of a vitamin C deficiency and that limes are rich in vitamin C. Nevertheless, the intervention worked and British seamen eventually were compelled to consume citrus fruit regularly, a practice that gave rise to the still-popular label *limeys*. Incidentally, it took about 50 years before the captain's "social program" was widely adopted. Then, as now, diffusion and acceptance of evaluation findings did not come easily.

WHAT IS EVALUATION RESEARCH?

Although the broadest definition of evaluation includes all efforts to place value on events, things, processes, or people, we will be concerned here with the evaluation of social programs. For purposes of orientation, we offer a preliminary definition of social program evaluation now and will present and discuss a more complete version later in this chapter: *Program evaluation is the use of social research procedures to systematically investigate the effectiveness of social intervention programs.* More specifically, evaluation researchers (evaluators) use social research methods to study, appraise, and help improve social programs in all their important aspects, including the diagnosis of the social problems they address, their conceptualization and design, their implementation and administration, their outcomes, and their efficiency.

At various times, policymakers, funding organizations, planners, program managers, taxpayers, or program clientele need to distinguish worthwhile programs from ineffective ones and launch new programs or revise existing ones so as to achieve certain desirable results. To do so, they must obtain answers to questions such as the following:

- What are the nature and scope of the problem? Where is it located, whom does it affect, and how does it affect them?
- What is it about the problem or its effects that justifies new, expanded, or modified social programs?
- What feasible interventions are likely to significantly ameliorate the problem?
- What are the appropriate target populations for intervention?
- Is a particular intervention reaching its target population?
- Is the intervention being implemented well? Are the intended services being provided?
- Is the intervention effective in attaining the desired goals or benefits?
- How much does the program cost?
- Is the program cost reasonable in relation to its effectiveness and benefits?

Exhibit 1-A conveys the views of one feisty senator about the need for evaluation evidence on program effectiveness. Answers to questions such as those above are necessary for local or specialized programs, such as job training in a small town, a new mathematics curriculum for elementary schools, or the outpatient services of a community mental health clinic, as well as for broad national or state programs such as health care, family preservation, or educational reform. Providing those answers is the work of persons in the program evaluation field.

Although this text emphasizes the evaluation of social programs, especially human service programs, program evaluation is not restricted to that arena. An excellent example of the broad scope of program evaluation is provided by the work of the Program Evaluation Methodology Division of the U.S. General Accounting Office (GAO). This unit was established in 1980 to foster the application of evaluation research to the program and policy review functions that the GAO performs for Congress. During its history, it has evaluated such endeavors as the procurement and testing of military hardware, quality control for drinking water, the maintenance of major highways, the use of hormones to stimulate growth in beef cattle, and other organized activities far afield from human services.

Indeed, the techniques described in this text have utility to virtually all spheres of activity in which issues of the effectiveness of organized social action are raised. For example, the mass communication and advertising industries use fundamentally the same approaches in developing media programs and marketing products; commercial and industrial corporations evaluate the procedures they use in selecting, training, and promoting employees and organizing their workforces; political candidates develop their campaigns by evaluating the voter appeal of different strategies; consumer products are tested for performance, durability, and safety; and administrators in both the public and private sectors often assess the clerical, fiscal, and personnel practices of their organizations.

The distinctions among these various applications of evaluation lie primarily in the nature and goals of the endeavors being evaluated. Our emphasis in this text is on the evaluation of programs designed to benefit the human condition rather than with such purposes as increasing profits or amassing influence and power. This choice of focus stems not from a sense of righteousness about the proper application of social research methods but from a desire to concentrate on a particularly significant and active area of evaluation combined with a practical need to limit the scope of the book.

EXHIBIT 1-A Veteran Policymaker Wants to See the Evaluation Results

But all the while we were taking on this large—and, as we can now say, hugely successful—effort [deficit reduction], we were constantly besieged by administration officials wanting us to *add* money for this social program or that social program. . . . *My* favorite in this miscellany was something called "family preservation," yet another categorical aid program (there were a dozen in place already) which amounted to a dollop of social services and a press release for some subcommittee chairman. The program was to cost $930 million over five years, starting at $60 million in fiscal year 1994. For three decades I had been watching families come apart in our society; now I was being told by seemingly everyone on the new team that one more program would do the trick. . . . At the risk of indiscretion, let me include in the record at this point a letter I wrote on July 28, 1993, to Dr. Laura D'Andrea Tyson, then the distinguished chairman of the Council of Economic Advisors, regarding the Family Preservation program:

Dear Dr. Tyson:

You will recall that last Thursday when you so kindly joined us at a meeting of the Democratic Policy Committee you and I discussed the President's family preservation proposal. You indicated how much he supports the measure. I assured you I, too, support it, but went on to ask what evidence was there that it would have any effect. You assured me there were such data. Just for fun, I asked for two citations.

The next day we received a fax from Sharon Glied of your staff with a number of citations and a paper, "Evaluating the Results," that appears to have been written by Frank Farrow of the Center for the Study of Social Policy here in Washington and Harold Richman at the Chapin Hall Center at the University of Chicago. The paper is quite direct: "Solid proof that family preservation services can affect a state's overall placement rates is still lacking."

Just yesterday, the same Chapin Hall Center released an "Evaluation of the Illinois Family First Placement Prevention Program: Final Report." This was a large scale study of the Illinois Family First initiative authorized by the Illinois Family Preservation Act of 1987. It was "designed to test effects if this program on out-of-home placement and other outcomes, such as subsequent child maltreatment." Data on case and service characteristics were provided by Family First caseworkers on approximately 4,500 cases: approximately 1,600 families participated in the randomized experiment. The findings are clear enough.

Overall, the Family First placement prevention program results in a slight increase in placement rates (when data from all experimental sites are combined). This effect disappears once case and site variations are taken into account. In other words, there are either negative effects or not effects.

This is nothing new. Here is Peter Rossi's conclusion in his 1992 paper, "Assessing Family Preservation Programs." Evaluations conducted to date "do not form a sufficient basis upon which to firmly decide whether family preservation programs are either effective or not."

May I say to you that there is nothing in the least surprising in either of these findings? From the mid-60s on this has been the repeated, I almost want to say consistent, pattern of evaluation studies. Either few effects or *negative* effects. Thus the negative income tax experiments of the 1970s appeared to produce an *increase* in family breakup.

This pattern of "counterintuitive" findings first appeared in the '60s. Greeley and Rossi, some of my work, and Coleman's. To this day I cannot decide whether we are dealing here with an artifact of methodology or a much larger and more intractable fact of social programs. In any event, by 1978 we had Rossi's Iron Law. To wit: "If there is any empirical law that is emerging from the past decade of widespread evaluation activity, it is that the expected value for any measured effect of a social program is zero."

I write you at such length for what I believe to be an important purpose. In the last six months I have been repeatedly impressed by the number of members of the Clinton administration who have assured me with great vigor that something or other is known in an area of social policy which, to the best of my understanding, is not known at all. This seems to me perilous. It is quite possible to live with uncertainty, with the possibility, even the likelihood that one is wrong. But beware of certainty where none exists. Ideological certainty easily degenerates into an insistence upon ignorance.

The great strength of political conservatives at this time (and for a generation) is that they are open to the thought that matters are complex. Liberals got into a reflexive pattern of denying this. I had hoped twelve years in the wilderness might have changed this; it may be it has only reinforced it. If this is so, current revival of liberalism will be brief and inconsequential.

Respectfully,

Senator Daniel Patrick Moynihan

SOURCE: Adapted, with permission, from D. P. Moynihan, *Miles to Go: A Personal History of Social Policy* (Cambridge, MA: Harvard University Press, 1996), pp. 47-49.

EXHIBIT 1-A Continued

The importance of evaluating social programs—both those currently in effect and those in various stages of design and pilot testing—should not be underestimated. The continuing challenge of devising ways to remedy the deficiencies in the quality of human life, both in industrialized countries and in less developed nations, needs no elaboration here. But along with the need for purposeful, practical, and well-organized efforts to implement new initiatives and improve existing ones comes the need for evaluation to determine if those efforts are worthwhile. Limited resources for social programs in every country, including the United States, make it critical that such investments yield demonstrable and proportionate social benefits. Moreover, experiences of the past several decades have highlighted the barriers to successful implementation of social programs and, correspondingly, the importance of assess-

ing the practicality of program design and the effectiveness of program operations.

To "put some meat on the bones" and make the notion of program evaluation more concrete, we offer below examples of social programs that have been evaluated under the sponsorship of local, state, and federal governmental agencies, international organizations, private foundations and philanthropies, and both nonprofit and for-profit associations and corporations.

• With support from the U.S. Department of Justice, the cities of Houston and Newark instituted community policing on a trial basis. In Houston the police set up neighborhood substations, conducted door-to-door surveys of citizen problems, started newsletters, and held community meetings. In Newark the police established local walking police beats, dispersed groups of loiterers, and conducted random checks of motor vehicles. In both cities the trial neighborhoods showed increases in citizen confidence in the police and slight reductions in crimes when compared with similar areas without community policing.

• In several major cities in the United States, a large private foundation provided the initial operating costs to establish community health centers in low-income areas. The centers were intended as an alternative way for residents to obtain ambulatory patient care otherwise available to them only from hospital outpatient clinics and emergency rooms at great public cost. It was further hoped that by improving access to such care, the clinics might increase timely treatment and thus reduce the need for lengthy and expensive hospital care. Evaluations indicated that some of these centers were cost-effective in comparison with hospital clinics.

• A small number of philanthropic advocates of school vouchers have initiated a privately funded program in New York City for poor families with children in the first three grades of more disadvantaged public schools. In spring 1997, scholarships of $1,400 for a period of three years were offered to eligible families to go toward tuition costs in the private schools of their choice. Some 14,000 scholarship applications were received, and 1,500 successful candidates were chosen by random selection. Taking advantage of this mode of selection, Mathematica Policy Research is regarding the program as a randomized experiment and intends to compare educational outcomes among those selected and attending private schools with outcomes among those who were not selected. The evaluation will be conducted over a three-year period.

• A community mental health center in a medium-sized New England city developed an extensive program using local community members to counsel teenagers and adults about their emotional, sexual, and educational problems. Compared with persons treated by psychiatrists and social workers, the clients of the indigenous counselors did as well in terms of need for hospitalization, maintenance of treatment, and self-reports of satisfaction with the center.

• In the past decade, the federal government has allowed states to modify their welfare programs provided that the changes were evaluated for their effects on clients and costs. Some states instituted strong work and job training requirements, others put time limits on benefits, and a few prohibited increases in benefits for children born while on the welfare rolls. Evaluation research showed that such policies were capable of reducing welfare rolls and increasing employment. Many of the program features studied were incorporated in the federal welfare reforms passed in 1996 (Personal Responsibility and Work Opportunity Reconciliation Act).

• Fully two-thirds of the world's rural children suffer mild to severe malnutrition, with serious consequences for their health, physical growth, and mental development. A major demonstration of the potential for improving children's health status and mental development by providing dietary supplements was undertaken in Central America. Pregnant women, lactating mothers, and children from birth through age 12 were provided with a daily high-protein, high-calorie food supplement. Results showed major gains in physical growth and modest increases in cognitive functioning.

• Over the past two decades, the number of reported cases of child abuse and neglect has more than doubled in the United States. As a consequence more than half a million children are in foster or group care. Concerned that removal from their families might be harmful, many child welfare agencies have provided short-term intensive services to families with abused or neglected children with the goal of preventing removal of the children from their families while keeping them safe from further abuse or neglect. Several evaluation studies have been undertaken in which children at risk of being removed from their homes were randomly assigned to "family preservation" programs or to the usual service. Those assigned to family preservation programs were no less likely to end up being removed from their homes, showing that these programs were ineffective.

• In an effort to increase worker satisfaction and product quality, a large manufacturing company reorganized its employees into independent work teams. Within the teams, workers designated and assigned tasks, recommended productivity quotas to management, and voted on the distribution of bonuses for productivity and quality improvements. Information from an assessment of the program revealed that it reduced days absent from the job, turnover rates, and similar measures of employee inefficiency.

These short examples illustrate the diversity of social interventions that have been systematically evaluated. However, all of them involve one particular evaluation activity: the assessment of program outcomes. As we will discuss later, evaluation may also focus on the need for a program, its design, operation and service delivery, or efficiency. Before moving ahead to more fully describe the nature and range of program evaluation, we provide a brief history of the development of the field to convey a sense of the traditions in which current work is rooted.

A BRIEF HISTORY OF EVALUATION RESEARCH

As we have noted, evaluation research is one facet of the general use of social research for understanding and addressing social problems. However, despite historical roots that extend to the 17th century, systematic evaluation research is a relatively modern development. The application of social research methods to program evaluation coincides with the growth and refinement of the methods themselves as well as with ideological, political, and demographic changes that have occurred during this century. Of key importance were the emergence and increased standing of the social sciences in

EXHIBIT 1-B An Early (1930s) Argument for Studying Social Program

No one can deny the progress in the social sciences. But with all the exacting methods developed, the economists, sociologists, and political scientists have suffered from a lack of large-scale experimental set-ups to match the everyday resources of the scientists in the laboratory.

The current enthusiasm over planning schemes now being devised by the alphabetical corporations of the federal government furnishes some hope that this deficiency may be partially remedied. The blueprints of these agencies and the carrying out of their plans may well be looked upon as the creation of experimental laboratories for the social scientists, and for the social workers, educators, and administrators who may profit from their research.

These laboratories, set up by the planning agencies of the New Deal, permit a more effective use of the experimental method in the research projects of the social scientists. This research, in turn, would not only be an addition to science but would also be a form of social auditing for the planning authorities in noting and accounting for the changes wrought by the programs.

SOURCE: Adapted, with permission, from A. S. Stephan, "Prospects and Possibilities: The New Deal and the New Social Research," *Social Forces*, May 1935, 13:515, 518.

universities and increased support for social research. Social science departments in universities became centers of early work in program evaluation and have continued to occupy an influential place in the field.

Evaluation Research as a Social Science Activity

Commitment to the systematic evaluation of social programs first became commonplace in education and public health. Prior to World War I, the most significant efforts were directed at assessing literacy and occupational training programs and public health initiatives to reduce mortality and morbidity from infectious diseases. By the 1930s, social scientists in various disciplines were advocating the use of rigorous research methods to assess social programs, and systematic evaluations became more frequent (Freeman, 1977). In sociology, for instance, Dodd's study of attempts to introduce water boiling as a public health practice in villages in the Middle East is a landmark in the pre-World War II literature. And it was an Arkansas sociology professor who first pleaded for studies of President Roosevelt's New Deal programs (see Exhibit 1-B). It was also during this period that social experimentation emerged in psychology. Lewin's pioneering "action research" studies and Lippitt and White's work on democratic and authoritarian leadership, for example, were widely influential evaluative studies. The famous Western Electric experiments on worker productivity that contributed the term *Hawthorne effect* to the social science lexicon date from this time as well. (See Bernstein and Freeman, 1975, for a more extended discussion and Bulmer, 1982, Cronbach et al., 1980, and Madaus and Stufflebeam, 1989, for somewhat different historical perspectives.)

From such beginnings, applied social research grew at an accelerating pace, with a particular boost provided by its contributions during World War II. Stouffer and his associates worked with the U.S. Army to develop procedures for monitoring soldier morale and evaluate personnel policies and propaganda techniques, whereas the Office of War Information used sample surveys to monitor civilian morale (Stouffer et al., 1949). A host of smaller studies assessed the efficacy of price controls and media campaigns to modify American eating habits. Similar social science efforts were mounted in Britain and elsewhere.

The Boom Period in Evaluation Research

Following World War II, numerous major programs were launched to meet needs for urban development and housing, technological and cultural education, occupational training, and preventive health activities. It was also during this time that major commitments were made to international programs for family planning, health and nutrition, and rural development. Expenditures were very large and consequently were accompanied by demands for "knowledge of results."

By the end of the 1950s, program evaluation research was commonplace. Social scientists were engaged in assessments of delinquency prevention programs, psychotherapeutic and psychopharmacological treatments, public housing programs, educational activities, community organization initiatives, and numerous other such areas. Studies were undertaken not only in the United States, Europe, and other industrialized countries but also in less developed nations. Increasingly, programs for family planning in Asia, nutrition and health care in Latin America, and agricultural and community development in Africa included evaluation components (Freeman, Rossi, and Wright, 1980; Levine et al., 1981). Expanding knowledge of the methods of social research, including sample surveys and advanced statistical procedures, and increased funding and administrative know-how, made possible even large-scale, multisite evaluation studies.

During the 1960s, the numbers of papers and books about evaluation research grew dramatically. Hayes's (1959) monograph on evaluation research in less developed countries, Suchman's (1967) review of evaluation research methods, and Campbell's (1969) call for social experimentation are a few illustrations. In the United States, a key impetus for the spurt of interest in evaluation research was the federal war on poverty, initiated under Lyndon Johnson's presidency. By the late 1960s, evaluation research had become, in the words of Wall Street, a growth industry.

In the early 1970s, evaluation research emerged as a distinct specialty field in the social sciences. A variety of books appeared, including the first texts (Weiss, 1972), critiques of the methodological quality of evaluation studies (Bernstein and Freeman, 1975), and discussions of the organizational and structural constraints on evaluation research (Riecken and Boruch, 1974). The journal *Evaluation Review* was established in 1976 and continues to be widely read by evaluators. Other journals followed in rapid succession, and today there are about a dozen devoted primarily to evaluation. During this period, special sessions on evaluation studies at the meetings of academic and practitioner groups became commonplace, and professional associations specifically for evaluation researchers were founded (see Exhibit 1-C for a listing of the major journals and professional organizations). By 1980, Cronbach and his associates were able to state, "Evaluation has become the liveliest frontier of American

EXHIBIT 1-C Major Evaluation Journals and Professional Organizations

Journals devoted primarily to program and policy evaluation:

- *Evaluation Review: A Journal of Applied Social Research* (Sage Publications)
- *Evaluation Practice,* renamed (1998) *American Journal of Evaluation* (JAI Press)
- *New Directions for Evaluation* (Jossey-Bass)
- *Evaluation: The International Journal of Theory, Research, and Practice* (Sage Publications Ltd.)
- *Evaluation and Program Planning* (Pergamon)
- *Journal of Policy Analysis and Management* (John Wiley)
- *Canadian Journal of Program Evaluation* (University of Calgary Press)
- *Evaluation Journal of Australasia* (Australasian Evaluation Society)
- *Evaluation & the Health Professions* (Sage Publications)
- *Educational Evaluation and Policy Analysis* (American Educational Research Association)
- *Assessment and Evaluation in Higher Education* (Carfax Publishing Ltd.)

Professional organizations for program and policy evaluators:

- American Evaluation Association (Web page: http://www.eval.org/)
- Association for Public Policy Analysis and Management (Web page: http://qsilver.queensu.ca/appam/)
- Canadian Evaluation Association (Web page: http://www.unites.uqam.ca/ces/ces-sce.html)
- Australasian Evaluation Society (Web page: http://www.parklane.com.au/aes/)
- European Evaluation Society (Web page: http://www.europeanevaluation.org)
- UK Evaluation Society (Web page: http://www.evaluation.org.uk)
- German Evaluation Society (Web page: http://www.fal.de/tissen/geproval.htm)
- Italian Evaluation Society (Web page: http://www.valutazione.it/)

social science" (pp. 12-13). Although the period of rapid growth is over, evaluation continues to be an important specialty area within the social sciences and is widely supported by public and private agencies.

The development of the field of evaluation in the postwar years was made possible to a large extent by advances in research methods and statistics applicable to the study of social problems, social processes, and interpersonal relations. Conversely, the need for sophisticated methods for evaluating social programs stimulated methodological innovation. In particular, two essential inputs contributed to the evolution of the field: improvements in systematic data collection brought about by the refinement of measurement and survey research procedures, and the development of electronic computers that made it possible to analyze large numbers of variables by means of multivariate statistics. The computer revolution was an especially important stimulus to the growth of evaluation research (Nagel, 1986) and has facilitated not only data analysis but data collection as well (Gray, 1988). The close relationship between technological changes and technical developments in applied social research continues today.

But history can obscure as well as illuminate. Although there is continuity in the development of the evaluation field, a qualitative change occurred as it matured. In its early years, evaluation was an endeavor shaped mainly by the interests of social researchers. In later stages, however, the consumers of evaluation research have had a significant influence on the field. Evaluation is now sustained primarily by policymakers, program planners, and administrators who use the findings and believe in the worth of the evaluation enterprise. It is also supported by the interests of the general public and the clients of the programs evaluated. Evaluations may not make front-page headlines, but their findings are often matters of intense concern to informed citizens and those whose lives are affected, directly or indirectly, by the programs at issue. Over the years, these various consumers and sponsors of evaluation have played an increasingly large role in defining the nature of the field.

Incorporation of the consumer perspective into evaluation research has moved the field beyond the study of social programs by applied social researchers. It has also become a political and managerial activity, an input into the complex mosaic from which emerge policy decisions and resources for the planning, design, implementation, and continuance of programs to better the human condition. In this regard, evaluation research must be seen as an integral part of the social policy and public administration movements.

Social Policy and Public Administration Movements

A full treatment of the development of the overlapping social policy and public administration movements would require tracing the remarkable growth of population and industrialization in the United States during the first part of this century. During this period, changing social values resulted in a shift of responsibility for community welfare from family members, church charities, and private philanthropy to government agencies. At least a few highlights are important here.

The Emergence of Government Programs

Social programs and the evaluation activities that accompanied them emerged from the relatively recent transfer of responsibility for the nation's social and environmental conditions, and the quality of life of its citizens, to governmental bodies. As Bremner (1956) has described, before World War I, except for war veterans, the provision of human services was seen primarily as the obligation of individuals and voluntary associations. Poor people, physically and mentally disabled persons, and troubled families were the clients of local charities staffed mainly by volunteers drawn from the ranks of the more fortunate. Our image of these volunteers as wealthy matrons toting baskets of food and hand-me-down clothing to give to the poor and unfortunate is only somewhat exaggerated. Along with civic associations and locally supported charity hospitals, county and

state asylums, locally supported public schools, state normal schools, and sectarian old-age homes, volunteers were the bulwark of our human service "system."

Indeed, government was comparatively small before the 1930s, particularly the federal government. There were few national health, education, and welfare programs and no need for an army of federal employees. The idea of annual federal expenditures of billions of dollars for health research, for instance, would have completely bewildered the government official of the 1920s. The notion of more billions going to purchase medical care for the aged or for poor persons would have been even more mind-boggling. Federal fiscal support of public education was infinitesimal—more dollars for public education currently flow from Washington in six months than were spent in the entire first decade of this century. Moreover, the scope and use of social and economic information mirrored the sparseness of government program operations. Even in the late 1930s, federal expenditures for social science research and statistics were only $40-$50 million, as compared to 40 to 50 times that amount today (Lynn, 1980).

Finally, human services and government operated under different norms than today. Key government officials usually were selected without regard to objective competence criteria; indeed, there were few ways of objectively determining competence. The professional civil service was a fraction of the size it is today, most jobs did not require technical know-how, and formal training programs were not widely available. Moreover, because its activities and influence were comparatively small, there was relatively little interest in what went on in government, at least in terms of human service programs.

All this began to change in the 1930s. Human services grew at a rapid pace with the advent of the Great Depression, and, of course, so did government in general, especially during the period surrounding World War II. In part because of the unwieldiness that accompanied this accelerated growth, there was strong pressure to apply the concepts and techniques of so-called scientific management, which were well regarded in industry, to government programs and activities. These ideas first took hold in the Department of Defense and then diffused to other government organizations, including human service agencies. Concepts and procedures for planning, budgeting, quality control, and accountability, as well as later, more sophisticated notions of cost-benefit analysis and system modeling, became the order of the day in the human resource area.

The Development of Policy and Public Administration Specialists

During this same period, persons with social science training began to apply themselves to understanding the political, organizational, and administrative decision making that took place in executive departments and other governmental agencies. Also, economists were perfecting models for planning and refining theories of macroeconomic social processes (Stokey and Zeckhauser, 1978). In part, the interests of social scientists in government at this time were purely academic. They wanted to know how government worked. However, persons in leadership positions in governmental agencies, groping for ways to deal with their large staffs and full coffers of funds, recognized a critical need for orderly, explicit ways to handle their policy, administrative, program, and planning responsibilities. They became convinced that concepts, techniques, and principles from economics, political science, and sociology could be useful. The study of the public sector thus grew into the largely applied research specialty that is now most commonly called "policy science" or "policy analysis."

Moreover, as the federal government became increasingly complex and technical, its programs could no longer be adequately managed by persons hired as intelligent generalists or because of their connections with political patrons, relatives, or friends. Most midlevel management jobs and many senior executive positions required specific substantive and technical skills, and those who filled them needed either training or extensive experience to do their work competently (see Exhibit 1-D). The state and local counterparts of federal agencies expanded at a similar rate, stimulated in part by federal initiatives and funding, and they too required skilled staffs. In response, university social science departments mobilized to provide trained persons for government positions as well as training researchers. Graduate schools of management, public health, and social work began programs to meet the need for executives and technicians, and special schools, generally with "public administration" in their titles, were established or expanded.

In short, a new army of professionals emerged. Furthermore, the institutionalization of policy analysis and public administration programs in universities has maintained the momentum of the intertwined policy science and public administration movements. Concepts and methods from the social sciences have become the core of the educational programs from which are drawn many of our public officials and program managers as well as the staffs of foundations, public and private human service agencies, and international organizations. In particular, these training programs stress evaluation research, both as an assessment procedure and as a body of knowledge about the effectiveness of programs and practice in the sundry policy areas.

The importance of evaluation is now acknowledged by those in political as well as executive roles. For example, the GAO, Congress's "watchdog," made a major commitment to evaluation in 1980 in response to congressional interest in the assessment of government initiatives. In addition, many federal agencies have their own evaluation units, as do a large number of their state counterparts. Even more commonplace at federal, state, and local levels are procedures for commissioning program evaluation, as the need arises, on a contract basis from university researchers or research firms and consultants.

In short, although evaluation research continues to have an academic side oriented toward training, methodology, theory, and relatively detached study of the nature and effects of social programs, it is a field that now extends well beyond university social science departments. Evaluation is generally practiced in a context of policy making, program management, and client or consumer advocacy. Thus, not only is its history intertwined with the social policy and public administration movements, but its practice typically occurs in the same political and organizational arenas as policy analysis and public administration.

The Great Society and Its Aftermath: Political Ideology and the Evaluation Enterprise

Evaluation activities increased rapidly during the Kennedy and Johnson eras when social

EXHIBIT 1-D The Rise of Policy Analysis

The steady growth in the number, variety, complexity, and social importance of policy issues confronting government is making increasing intellectual demands on public officials and their staffs. What should be done about nuclear safety, teenage pregnancies, urban decline, rising hospital costs, unemployment among black youth, violence toward spouses and children, and the disposal of toxic wastes? Many of these subjects were not on the public agenda 20 years ago. They are priority issues now, and new ones of a similar character emerge virtually every year. For most elected and appointed officials and their staffs, such complicated and controversial questions are outside the scope of their judgment and previous experience. Yet the questions cannot be sidestepped; government executives are expected to deal with them responsibly and effectively.

To aid them in thinking about and deciding on such matters, public officials have been depending to an increasing extent on knowledge derived from research, policy analysis, program evaluations, and statistics to inform or buttress their views. More often than in the past, elected and appointed officials in the various branches and levels of government, from federal judges to town selectmen, are citing studies, official data, and expert opinion in at least partial justification for their actions. Their staffs, which have been increasing in size and responsibility in recent decades, include growing numbers of people trained in or familiar with analytic techniques to gather and evaluate information. Increasing amounts of research, analysis, and data gathering are being done.

Because the power to influence policy is widely shared in our system of government, public officials seeking to influence policy—to play the policy game well—must be persuasive. Because of the changing character of policy issues, it is probably harder to be persuasive than it used to be. Seniority, affability, and clever "wheeling and dealing" may be relatively less influential than being generally knowledgeable and tough-minded, having the ability to offer ideas and solutions that can attract a wide following, or having a reputation as a well-informed critic. Increasingly, officials from the president on down lose influence in policy debates when they cannot get their numbers right or when their ideas and arguments are successfully challenged by opposing experts. Indeed, thorough and detailed command of an issue or problem is often mandatory. Legislatures are requiring executives to be experts in the programs and issues under their jurisdiction. Judges are requiring detailed proof that administrative decisions are not arbitrary and capricious. Budget officials demand positive program evaluations. The public demands accountability. Thus the dynamic processes whereby our political system confronts social problems are perceptibly, if not dramatically, raising the standards of substantive and managerial competence in the performance of public responsibilities.

SOURCE: Adapted, with permission, from Laurence E. Lynn, Jr., *Designing Public Policy* (Santa Monica, CA: Scott, Foresman, 1980).

EXHIBIT 1-E The 1960s Growth in Policy Analysis and Evaluation Research

The year 1965 was an important one in the evolution of "policy analysis and evaluation research" as an independent branch of study. Two developments at the federal government level—the War on Poverty-Great Society initiative and the Executive Order establishing the Planning-Programming-Budgeting (PPB) system—were of signal importance in this regard. Both offered standing, legitimacy, and financial support to scholars who would turn their skills and interests toward examining the efficiency with which public measures allocate resources, their impacts on individual behavior, their effectiveness in attaining the objectives for which they were designed, and their effects on the well-being of rich versus poor, minority versus majority, and North versus South.

The War on Poverty-Great Society developments initiated in 1965 represented a set of social interventions on an unprecedented scale. All impacted by them wanted to know if they were working, and who was being affected by them and how. Those with the skills to answer these questions found both financial support and an interested audience for their efforts. And the social science community responded. The same year saw government-wide adoption of the formal evaluation and analysis methods that had earlier been applied in Robert McNamara's Defense Department in the Planning-Programming-Budgeting system. A presidential Executive Order gave employment and financial support to thousands who wished to apply their analytical skills to such efficiency, effectiveness, and equity questions.

SOURCE: Robert H. Haveman, "Policy Analysis and Evaluation Research After Twenty Years," *Policy Studies Journal*, 1987, 16:191-218.

programs undertaken under the rubrics of the War on Poverty and the Great Society provided extensive resources to deal with unemployment, crime, urban deterioration, access to medical care, and mental health treatment (see Exhibit 1-E). These programs were often hurriedly put into place, and at least a significant portion were poorly conceived, improperly implemented, and ineffectively administered. Findings of limited effectiveness and poor benefit-to-cost ratios for the large-scale federal initiatives of this era prompted widespread reappraisal of the magnitude of effects that can be expected from social programs. Social intervention all too often yields small gains, much to the chagrin of those who advocate them (Weick, 1984). But more realistic expectations for social programs only increase the importance of undertaking evaluation before putting programs into place on a permanent and widespread basis or making significant modifications to them.

Partly as a consequence of the apparent ineffectiveness of many initiatives, the decade of the 1970s was marked by increasing resistance to the continued expansion of government programs (Freeman and Solomon, 1979). The reaction was most clear in referenda such as California's Proposition 13, which limited real estate tax revenue, and in "sunset laws," which required the automatic shutdown of ineffective programs (Adams and Sherman, 1978). Of course, some of the attacks on big

government were simply political campaign rhetoric. And although a number of states and major cities enacted sunset laws, in only a few instances have programs actually been shut down. More often, only superficial and symbolic assessments were undertaken or the programs were given extensions to allow them to develop documentation for their effectiveness.

Nevertheless, it is clear that the rise of fiscal conservatism in the 1970s resulted in a decline in governmental support for new social programs and, to some extent, private support as well. This, in turn, brought about a change in emphasis in the evaluation field. In particular, increased attention has been given to assessing the expenditures of social programs in comparison to their benefits and to demonstrating fiscal accountability and effective management. In the process, many fiscal and political conservatives, often skeptical about social science, have joined the advocates of social action programs in pressing for the information that evaluations provide.

In the 1980s, during both the Reagan and Bush administrations, domestic federal expenditures were curtailed in an attempt to control inflation and reduce the federal deficit. Many of the largest cutbacks were targeted on social programs. A similar posture was manifest in many states and cities; indeed, some of the local and state reactions to their deteriorating economic situations were particularly severe. These developments were partly a consequence of the distrust, hostility, and political actions of community members dismayed with the painful bite of income and property taxes. As we have indicated, however, they were also influenced by disenchantment with the modest effects and poor implementation of many of the programs most ardently championed by public officials, planners, and politicians in the past several decades.

As should be apparent, social programs and, consequently, the evaluation enterprise are shaped by the changing times. Political perspectives during the 1980s, not only in the United States but also in a number of Western European countries, have brought about increased concern with the balance of benefits and costs for social programs, even in social problem areas that receive generous funding. On the intellectual front, both conservative and liberal critique of the Great Society programs have had an impact on the evaluation field. Although these criticisms were sometimes based more on ideology than evidence, they nevertheless have drawn on evaluation results in condemning social programs. For instance, evaluation research has been used to argue that the major federal welfare program, Aid to Families With Dependent Children (AFDC), provides perverse incentives that increase the social problem it was intended to ameliorate (Murray, 1984). The evaluation field has thus been thrust into the middle of contentious debates about the very concept of social intervention and faced with new challenges to demonstrate that any major program initiative can be effective.

Meanwhile, new social problems are continually emerging on the political landscape, accompanied by demands that they receive programmatic attention and that the efforts made to ameliorate them be evaluated. A striking example is the issue of homelessness (Jencks, 1994; Rossi, 1989). At the time the first edition of this text was published (1979), there was little public notice of the homeless, little political activity to initiate and fund public programs to better their lot, and, consequently, little effort to estimate the number of such persons, their characteristics, or the reasons for their condition. Today, views on how to deal with the homeless range from an emphasis on institutionalization to efforts to increase tolerance for street people and make heavier commitments of resources for their medical care, shelter, food, and other necessities. Enumerations of the homeless, diagnoses of their conditions, and demonstration programs with accompanying evaluations are numerous and expanding, despite the budgetary shortfalls at all levels of government.

The Evaluation Field in the 1990s

Fiscal conservatism, the devolution of responsibility to the states, and skepticism about social programs dominate national policy making today. The Clinton Democratic presidency and the Republican majority Congress seem determined to cut federal spending and hand over major social programs to the administration of the states. For instance, welfare reform legislation, passed in 1996 (Personal Responsibility and Work Opportunity Reconciliation Act), that abolished the entitlement status of AFDC required the states to administer it under severe time eligibility restrictions and imposed an emphasis on moving beneficiaries into employment. Such changes have mixed implications for evaluation. On the one hand, these major revisions and reforms in social programs require evaluations if anything is to be learned about their fiscal and social impacts. On the other hand, much of the responsibility for conducting evaluation has devolved to the states along with the programs, and many states do not have the capabilities or the will to undertake the rigorous evaluation needed.

Fundamentally, however, whether there is a "liberal" or "conservative" outlook in government and among the public should not change the role of evaluation research in the social program arena (Freeman, 1983). Rather, these different political conditions raise different sets of evaluation questions corresponding to the shifts in the concerns emphasized by the stakeholders. Regardless of political outlook, two matters are clear about the 1990s. First, restraints on resources will continue to require choosing the social problem areas on which to concentrate and the programs that should be given priority. Second, intensive scrutiny of existing programs will continue because of the pressure to curtail or dismantle those that do not demonstrate that their services are effective and efficient. Moreover, both dissatisfaction with existing programs and shifts in political currents will result in new and modified programs that come forward with promises of being more effective and less costly. All these circumstances will generate a need, and quite likely a demand, for evaluation research.

Major worldwide changes will also affect the evaluation field. The globalization of economic activities may force nations with generous social welfare programs to cut back their expenditures to remain competitive on world markets. Nations emerging from totalitarian socialist regimes, on the other hand, may have to launch new social initiatives. Indeed, in many developing nations throughout the world there is intense pressure to develop *and* evaluate social programs virtually overnight. At the same time, evaluation itself is becoming increasingly international (see Exhibit 1-F).

Perhaps more subtle, but at least as great a source of influence on programs and their evaluations, are shifts in the values and self-interests of community members and organizations. Surveys in the United States and a number of Western European countries, for instance, document a reduced emphasis among workers on earnings and an increased value placed on nonwork activities. As another illus-

EXHIBIT 1-F Evaluation Is Becoming Internationalized

Evaluation is becoming increasingly international, but in ways that go beyond previous conceptions of what *international* means. *International* is no longer used only to describe the efforts of particular evaluators in individual countries around the world—although it certainly is still used in this way. . . . Today, evaluation is also becoming international in the sense of being at the same time more indigenous, more global, and more transnational. By *indigenous,* we mean that evaluators in different countries around the world are developing their own infrastructures to support their endeavors as well as their own preferred theoretical and methodological approaches. By *global,* we mean that developments in one part of the globe frequently affect people, institutions, and programs all around the world. . . . By *transnational,* we mean that the problems and programs that we are called upon to evaluate today often extend beyond the boundaries of any one nation, any one continent, or even one hemisphere. . . . These include problems of pollution, of the economics of developing countries, and of the role of women in society. We cannot say exactly what the best responses to these internationalizing developments will be for evaluators, but we do know that recognizing the developments is the first step toward accommodating to them.

SOURCE: Quoted, with permission, from Eleanor Chelimsky and William R. Shadish, *Evaluation for the 21st Century: A Handbook* (Thousand Oaks, CA: Sage, 1997), pp. xi-xii.

tration, until this decade most large corporations opposed publicly funded national health insurance or governmental health subsidy programs for employed persons. But the extremely high cost of medical care and its impact on worker incomes has led to a decided change in outlook. Trends in values and self-interests quite likely will have much to do with the nature and scope of the social programs that are initiated and those that are continued, with corresponding implications for their assessment.

AN OVERVIEW OF PROGRAM EVALUATION

With the benefit of some historical context, we can attempt a more complete definition of program evaluation, as we wish to use the term, than the preliminary version offered earlier: *Program 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.* Elaborating on the various distinct components of this definition, in turn, will provide an introductory overview of the nature of program evaluation as presented in this book.

Application of Social Research Procedures

The concept of *evaluation* entails, on the one hand, a description of the performance of the entity being evaluated and, on the other, some standards or criteria by which that performance is judged (see Exhibit 1-G). It follows that a central task of the program evaluator is to construct a valid description of those areas of program performance that are at issue in a form that permits incisive comparison with the applicable criteria. This task presents several challenges, some involving identification of the areas of performance at issue and the applicable criteria that we will address later. Here we focus on the problem of constructing a valid description of performance that is sufficiently distinct and precise to permit meaningful assessment.

A valid description of program performance is one that accurately represents what the program actually accomplishes. As should be obvious, it is a serious defect for an evaluation to fail to describe program performance with a reasonable degree of validity. Doing so is a misrepresentation of the facts that may distort a program's accomplishments, deny it credit for its successes, or overlook shortcomings for which it should be accountable. A distinct and precise description of program performance is one that is sufficiently definite and discriminating for meaningful variations in level of performance to be detected. An evaluation that produces an unduly vague or equivocal description of program performance may also fall short by making it impossible to determine with confidence whether program performance actually meets some appropriate standard.

Social research procedures and the accompanying standards of methodological quality have been developed and refined over the years explicitly for the purpose of constructing sound factual descriptions of social phenomena. In particular, contemporary social science techniques of systematic observation, measurement, sampling, research design, and data

EXHIBIT 1-G The Two Arms of Evaluation

Evaluation is the process of determining the merit, worth, and value of things, and evaluations are the products of that process. . . . Evaluation is not the mere accumulation and summarizing of data that are clearly relevant for decision making, although there are still evaluation theorists who take that to be its definition. . . . In all contexts, gathering and analyzing the data that are needed for decision making—difficult though that often is—comprises only one of the two key components in evaluation; absent the other component, and absent a procedure for combining them, we simply lack anything that qualifies as an evaluation. *Consumer Reports* does not just test products and report the test scores; it (i) *rates or ranks* by (ii) *merit or cost-effectiveness.* To get to that kind of conclusion requires an input of something besides data, in the usual sense of that term. The second element is required to get to conclusions about merit or net benefits, and it consists of evaluative premises or standards. . . . A more straightforward approach is just to say that evaluation has two arms, only one of which is engaged in data-gathering. The other arm collects, clarifies, and verifies relevant values and standards.

SOURCE: Quoted, with permission, from Michael Scriven, *Evaluation Thesaurus,* 4th ed. (Newbury Park, CA: Sage, 1991), pp. 1, 4-5.

analysis represent rather highly evolved procedures for producing valid, reliable, and precise characterizations of social behavior. Because social programs are instances of organized social behavior, we take it as virtually self-evident that social research procedures offer the best approach to the task of describing program performance in ways that will be as credible and defensible as possible. Moreover, credibility and defensibility are important characteristics of the evidence evaluators put forward, both because of the practical importance of evaluation in most contexts of application and because of the disputatious reception often given to evaluation results when they do not conform to the expectations of significant stakeholders.

Regardless of the type of social intervention under study, therefore, we advocate evaluations that employ social research procedures for gathering evidence about program performance and analyzing and interpreting that evidence. This commitment to the rules of social research is at the core of our perspective on evaluation and is what we mean by the subtitle of this book, *A Systematic Approach.* This is not to say, however, that we believe evaluation studies must follow some particular social research style or combination of styles, whether quantitative or qualitative, experimental or ethnographic, "positivistic" or "naturalistic." Indeed, one of the principal characteristics of program evaluation is that its methods cover the gamut of prevailing social research paradigms. Nor does this commitment to the methods of social science mean that we think current methods are a finished piece of work beyond improvement. Evaluators must often innovate and improvise as they attempt to find ways to gather credible, defensible evidence about social programs. In fact, evaluators have been, and will likely continue to be, especially productive contributors to furthering methodological development in applied social research.

Finally, our view does not imply that methodological quality is necessarily the most important aspect of an evaluation nor that only the highest technical standards, without compromise, are appropriate for evaluation. As Carol Weiss (1972) once observed, social programs are inherently inhospitable environments for research purposes. The nature of program circumstances, and of the particular issues the evaluator is called on to address, frequently necessitates compromises and adaptations of textbook methodological standards. The challenges to the evaluator, as we see them, are to match the research procedures to the evaluation questions and circumstances as well as possible and, whatever procedures are used, to apply them at the highest possible standard feasible in those circumstances.

The Effectiveness of Social Intervention Programs

Any program evaluation worthy of the name must, of course, evaluate; that is, some assessment must be made of one or more aspects of the program. As indicated above, evaluating something requires that pertinent dimensions of its performance or characteristics be described and then judged against appropriate standards or criteria. Program evaluation generally involves assessment of one or more of five program domains: (a) the need for the program, (b) the design of the program, (c) the program implementation and service delivery, (d) the program impact or outcomes, and (e) program efficiency (cost-effectiveness). In some circumstances, an evaluation of a social program may encompass all these program domains;

evaluations that do so are termed *comprehensive evaluations.*

Evaluation methods can be applied to many kinds of programs, projects, and endeavors, but the domain of program evaluation orients chiefly to social programs and the focus of this book is primarily on that type of program. What we mean by a social program in this context is a planned, organized, and usually ongoing set of activities carried out for the purpose of improving some social condition. A social program thus is directed at ameliorating a social problem or responding to a social need, usually through the provision of some form of human services. As we are using the term, therefore, social programs are defined as entities whose principal reason for existing is to "do good," that is, to produce social benefits and improve social conditions. It follows that they are appropriately held accountable within an evaluative framework on the basis of their contribution to the social good. Most social programs will thus hold themselves accountable for producing positive social effects, at least to the extent of recognizing the legitimacy of that expectation. In addition, of course, many social programs will be held accountable for such results by those parties who invest in them, sponsor them, administer them, or are legally responsible for them, for instance, taxpayers, funders, boards of directors, agency heads, and legislators.

The importance of this issue is that it has critical implications for the question of what criteria or standards should be used to assess programs when conducting a program evaluation. Different values frameworks are appropriate for different types of programs. Many of the evaluation methods described in this book can be applied to programs in the business sector, for instance, but the applicable criteria for assessing performance will generally have more to do with "the bottom line" of profits and productivity than with amelioration of social problems. Similarly, evaluation methods could be applied to assess social clubs, professional organizations, and other such programs whose purposes are to provide certain benefits to their members. The criteria for assessing these programs would largely and appropriately relate only to the interests of the members. The goals of other types of programs are generally quite different from those of social programs and the criteria appropriate for assessing them will also be different.

Our focus on the large and important topic of evaluating social programs, therefore, carries with it a set of assumptions about the general value framework within which appropriate criteria and standards for assessing the various aspects of those programs will be defined. In particular, when we describe evaluation as investigating the *effectiveness* of social programs we are assuming that what effectiveness means for such programs relates ultimately to their contribution to improving social conditions. Of course, there may be ambiguity and dispute about just what contributions a program should be making and the implications for everyday program operations, which an evaluator will have to resolve before appropriate criteria can be defined and an evaluation can be conducted. These matters will be discussed in greater detail in other chapters of this book. Most of our discussion, advice, and illustrations, however, assumes that it is social programs that are being evaluated and that the foundation for judgments about how effective they are is some articulation of the social good they are expected to produce.

Adapting Evaluation to the Political and Organizational Context of the Program

Program evaluation is not a cut-and-dried activity like putting up a prefabricated house or checking a document with a word processor's spelling program. Rather, evaluation is a practice in which the initial evaluation plan must be tailor-made to the particular program circumstances and then typically requires revision and modification during its implementation. The specific form and scope of an evaluation depend primarily on its purposes and audience, the nature of the program being evaluated, and the political and administrative context within which the evaluation is conducted.

The evaluation plan is generally organized around the questions posed about the program by those who request and commission the evaluation (the *evaluation sponsor*) and other pertinent stakeholders. These questions may be stipulated in very specific, fixed terms that allow little flexibility, as in a detailed contract for evaluation services, but typically the evaluator must negotiate with the evaluation sponsors and stakeholders to develop and refine the questions. Although these parties presumably know their own interests and purposes, they will not necessarily formulate their concerns in ways that the evaluator can use to structure an evaluation plan. For instance, the initial questions may be vague, overly general, or phrased in program jargon that must be translated for more general consumption. Occasionally, the evaluation questions put forward are essentially pro forma (e.g., is the program effective?) and have not emerged from careful reflection regarding the relevant issues. In such cases, the evaluator must probe thoroughly to determine what this means to the evaluation sponsor and program stakeholders and why they are concerned.

As important to tailor-making an evaluation plan as the questions to be answered are the reasons why those questions are being asked and the use that will be made of the answers. Social programs consist of, and exist within, a swirl of individual, organizational, and political decisions dealing with a range of issues from the trivia of ordering paper clips to threat of termination. In such a context, an evaluation must deal with the issues that matter, provide information that addresses those issues, develop that information in a way that is timely and meaningful for the decisionmakers, and communicate it in a form that is usable for their purposes. An evaluation might be designed quite differently if it is to provide information about the quality of service as feedback to the program director for purposes of incremental program improvement than if it is to provide such information to an external funder who will use it to decide whether to renew the program's funding. In all cases, however, it must be sensitive to the political context within which it is planned and conducted (see Exhibit 1-H).

As a practical matter, of course, an evaluation must also be tailored to the organizational makeup of the program. The availability of administrative cooperation and support; the ways in which program files and data are kept and access permitted to them; the character of the services provided; the nature, frequency, duration, and location of the contact between program and client; and numerous other such matters must be taken into consideration in the evaluation design. In addition, once an evaluation is launched, it is common for changes and "in-flight" corrections to be required. Modifications, perhaps even compromises, may be necessary in the types, quantity, or quality of the data collected as a result of unanticipated practical or political obstacles.

EXHIBIT 1-H Where Politics and Evaluation Meet

Evaluation is a rational enterprise that takes place in a political context. Political considerations intrude in three major ways, and the evaluator who fails to recognize their presence is in for a series of shocks and frustrations:

First, the policies and programs with which evaluation deals are the creatures of political decisions. They were proposed, defined, debated, enacted, and funded through political processes, and in implementation they remain subject to pressures—both supportive and hostile—that arise out of the play of politics.

Second, because evaluation is undertaken in order to feed into decision making, its reports enter the political arena. There evaluative evidence of program outcomes has to compete for attention with other factors that carry weight in the political process.

Third, and perhaps least recognized, evaluation itself has a political stance. By its very nature, it makes implicit political statements about such issues as the problematic nature of some programs and the unchallengeability of others, the legitimacy of program goals and program strategies, the utility of strategies of incremental reform, and even the appropriate role of the social scientist in policy and program formation.

Knowing that political constraints and resistances exist is not a reason for abandoning evaluation research; rather, it is a precondition for usable evaluation research. Only when the evaluator has insight into the interests and motivations of other actors in the system, into the roles that he himself is consciously or inadvertently playing, the obstacles and opportunities that impinge upon the evaluative effort, and the limitations and possibilities for putting the results of evaluation to work—only with sensitivity to the politics of evaluation research—can the evaluator be as creative and strategically useful as he should be.

SOURCE: Quoted, with permission, from Carol H. Weiss, "Where Politics and Evaluation Research Meet," *Evaluation Practice*, 1993, 14(1):94, where the original 1973 version was reprinted as one of the classics in the evaluation field.

Moreover, adaptations may be required in the basic questions being addressed in response to shifts that occur in the operation of the program or the composition and interests of the stakeholders.

Informing Social Action to Improve Social Conditions

As indicated, this book is about the evaluation of social programs or, more generally, those programs whose mission, whether defined by the program itself or the expectations of the public that supports it, is to intervene in social conditions in ways that make them better. And if the purpose of these programs is in some way to improve the human condition, the purpose of evaluation, in turn, is to improve the programs.

In particular, the role of program evaluation is to provide answers. It answers questions about what the program is doing but, more important, about how well it is being done and whether it is worth doing. It is undertaken on

the assumption that there is an audience with such questions and an interest in the answers. The concept of program evaluation presupposes more than a merely interested audience, however. It is characteristically designed to produce answers that will be useful and will actually be used. An evaluation study, therefore, primarily addresses the audience (or, more accurately, audiences) with the potential to make decisions and take action on the basis of the evaluation results. This point is fundamental to evaluation—its purpose is to inform social action.

In most instances, the main audiences to which an evaluation is directed are the sponsors of the evaluation and other program stakeholders. These are the individuals or groups with rather immediate interests in the particular program being evaluated and includes those with decision-making authority over the program or the capability to influence such decision making. Evaluation findings may assist such persons to make go/no-go decisions about specific program modifications or, perhaps, about initiation or continuation of entire programs. They may bear on political, practical, and resource considerations or make an impression on the views of individuals with influence. They may have direct effects on judgments of a program's value as part of an oversight process that holds the program accountable for results. Or they may have indirect effects in shaping the way program issues are framed and the nature of the debate about them. The evaluation sponsor and other such decisionmakers and stakeholders have a rather obvious primacy in these matters; however, they are not the only audience potentially interested in the evaluation nor are they necessarily the only agents whose actions may be influenced by the evaluation.

Programs, like people, have their unique profiles of characteristics but also share characteristics with others that make for meaningful categories and groupings. What is learned from an evaluation about one specific program, say, a drug use prevention program implemented at a particular high school, also tells us something about the whole category of similar programs. Many of the parties involved with social intervention must make decisions and take action that relates to categories or types of programs rather than individual instances. Policymakers, program planners, and program sponsors and funders, for instance, must often select, promote, or support a particular type of program rather than any one instance. A federal legislative committee may deliberate the merits of compensatory education programs, or a state correctional department may consider instituting boot camps for juvenile offenders, or a philanthropic foundation may decide to promote and underwrite programs that provide visiting nurses to single mothers. The body of evaluation findings for programs of each of these types is very pertinent to decisions and social actions of this sort. Each evaluation study, therefore, not only informs the immediate stakeholders but potentially informs those whose situations require decisions and action about different program concepts.

Indeed, one important form of evaluation research is that which is conducted on demonstration programs, that is, social intervention projects designed and implemented explicitly to test the value of an innovative program concept. In such cases, the findings of the evaluation are significant because of what they reveal about the program concept and are used primarily by those involved in policy making and program development at levels broader than any one program. Another significant evaluation-related activity is the integration of the findings of multiple evaluations of a particular type of program into a synthesis that can inform policy making and program planning. Some evaluation researchers, therefore, have been involved in the activities of systematic research synthesis or meta-analysis (Lipsey and Wilson, 1993).

Evaluations can thus inform social action by providing useful feedback for management and administrative purposes; by supporting the oversight functions of those funders, sponsors, and authorities to which the program is accountable; or by accurately depicting program activities and accomplishments to advocates, adversaries, clients, and other stakeholders. They may also contribute information for planning and policy purposes, indicate if innovative approaches to community problems are worth pursuing, or demonstrate the utility of some principle of professional practice. Evaluation research may even help shape our general understanding of how to bring about planned social change by testing social science hypotheses regarding the effects of certain broad forms of intervention. The common denominator is that evaluation research is intended to be useful and used, either directly and immediately or as an incremental contribution to a cumulative body of practical knowledge.

These assertions, of course, assume that an evaluation would not be undertaken unless there was an audience interested in receiving and, at least potentially, using the findings. Unfortunately, there are instances in which evaluations are commissioned without any intention of using their findings. Evaluations may be conducted only because they are mandated by program funders and then ignored when the findings are presented. Or an evaluation may be carried out because "everyone does it" without expectation of using the results in any significant way. Responsible evaluators try to avoid being drawn into such situations of "ritualistic" evaluation. An early step in planning an evaluation, therefore, is a thorough inquiry into the motivation of the evaluation sponsors, the intended purposes of the evaluation, and the uses to be made of the findings.

EVALUATION RESEARCH IN PRACTICE

We have outlined the general considerations, purposes, and approaches that shape evaluation research and guide its application to any program situation. In actual practice, application of these concepts typically involves something of a balancing act between competing forces. Paramount among these is the inherent conflict between the requirements of systematic inquiry and data collection associated with evaluation research and the organizational imperatives of a social program devoted to delivery of service and maintenance of essential routine activities. The planning phase of evaluation, which is best accomplished in collaboration with program personnel and stakeholders, and, especially, the data collection phase necessarily place unusual and not altogether welcome demands on program personnel and program processes. Data collection, for instance, may require interaction with program files, clients, staff, and facilities that are disruptive of normal program processes and distract from and, in some cases, even compromise the service functions that are the program's primary obligation.

Every evaluation plan, therefore, must negotiate a middle way between optimizing the program circumstances for research purposes and minimizing the disruption caused to nor-

mal program operation. We use the word *negotiate* quite deliberately here, because the best approach to the inherent tension between the requirements of research and those of running a service program is for the evaluator to develop the evaluation plan collaboratively with program personnel. If the needs and purposes of the evaluation are spelled out in detail before the research begins, and those program personnel who will be affected (not just the administrators) are given an opportunity to react, make input, and otherwise help shape the data collection plan, the result is usually a more workable plan and better cooperation from program personnel in the face of the inevitable strains the evaluation will place on them.

In addition to the conflict between evaluation and program functions, there are other inherent tensions in the practice of evaluation that warrant comment. Here we introduce a few of the more notable dilemmas the evaluator must confront: the incompatibility of a fixed evaluation plan with the volatility of social programs; the strain between a press for evaluations to be scientific, on the one hand, and pragmatic, on the other; and the competing approaches to evaluation offered up by a field of great diversity and little consensus.

Evaluation and the Volatility of Social Programs

One of the most challenging aspects of program evaluation is the continually changing decision-making milieu of the social programs that are evaluated. In particular, the resources, priorities, and relative influence of the various sponsors and stakeholders of social programs are dynamic. These changes are frequently associated with the shifts in political context and social trends we noted earlier. For example, the 1996 welfare reform legislation has drastically altered the nature of income support for poor families. A program reconfiguration of this magnitude clearly requires evaluations of family income support programs to be defined differently than in the past with new outcomes and quite different program components at issue.

Priorities and responsibilities more specific to the organizations implementing a program can also change in significant ways. For example, a school system relieved by the courts from forced school busing may lose interest in its programs to increase white students' acceptance of attendance in predominantly minority schools. Or unanticipated problems with the intervention may require modifying the program and, consequently, the evaluation plan as well. For instance, a program to reduce the absence rates of low-income high school students by providing comprehensive medical care might be thwarted if a large proportion of the eligible students refused the services.

Somewhat ironically, preliminary findings from the evaluation itself may stimulate program changes that render the remainder of the evaluation plan obsolete. Consider, for example, a study of the impact of an alcohol treatment program that included six-month and one-year follow-ups of the clients. When the six-month follow-up revealed very high rates of drunkenness among the treatment group, the program staff markedly modified the intervention.

Not all social programs, of course, transform significantly while an evaluation is under way. Nonetheless, the evaluator must attempt to anticipate such changes and prepare for them to the extent possible. More important, perhaps, is to match the form of the evaluation to the program circumstances and prospects at the time the evaluation is planned. It would generally make little sense to design a rigorous impact assessment for a program under consideration for significant revision by relevant decisionmakers. Of equal importance, however, is the flexibility the evaluator brings to the evaluation task. Knowing the dynamic nature of programs, evaluators must be prepared to substantially modify an evaluation if it becomes apparent that the original plan is no longer appropriate to the circumstances. This often involves difficult issues associated with the availability of resources for the evaluation, the time lines for producing results, and the relationships with the program administrators and evaluation sponsors, so it is not to be taken lightly. Social programs are not research laboratories, however, and evaluators must expect to be buffeted about by forces and events outside their control.

The contrast between the image of a research laboratory and the reality of social programs as places to conduct social research leads us directly to another of the inherent tensions in evaluation, that between a scientific and a pragmatic perspective on the process.

Scientific Versus Pragmatic Evaluation Postures

Perhaps the single most influential article in the evaluation field was written by the late Donald Campbell and published in 1969. This article outlined a perspective that Campbell advanced over several decades: Policy and program decisions should emerge from continual social experimentation that tests ways to improve social conditions. Not only did he hold this position in principle, but he contended that the technology of social research made it feasible to actually develop the "experimenting society." Campbell, thus, sought to extend the experimental model, as he learned and practiced it in social psychology, to evaluation research. Although he tempered his position in some of his later writing, it is fair to characterize him as fitting evaluation research into the scientific research paradigm (see Exhibit 1-I).

Campbell's position was challenged by Lee Cronbach, another giant in the evaluation field. While acknowledging that scientific investigation and evaluation may use some of the same logic of inquiry and research procedures, Cronbach argued that the purpose of evaluation sharply differentiates it from scientific research (Cronbach, 1982). In his view, evaluation is more art than science and every evaluation should be tailored to meet the needs of program decisionmakers and stakeholders. Thus, whereas scientific studies strive principally to meet research standards, evaluations should be dedicated to providing maximally useful information for decisionmakers given the political circumstances, program constraints, and available resources (see Exhibit 1-J).

One might be inclined to agree with both these views—that evaluations should meet high standards of scientific research quality *and* be fully dedicated to serving the information needs of program decisionmakers. The problem, of course, is that in practice these two goals often are not especially compatible. In particular, social research at a high scientific standard generally requires resources that exceed what is available in the typical program evaluation context. These resources include time, because high-quality research cannot be done quickly whereas program decisions often have to be made on short notice, and funding proportionate to the expertise, level of effort, and materials required for research to scientific standards. Moreover, research within the scientific framework may require structuring the

EXHIBIT 1-I Reforms as Experiments

The United States and other modern nations should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific social problems, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify, or discard them on the basis of apparent effectiveness on the multiple imperfect criteria available.

SOURCE: Quoted from Donald Campbell, "Reforms as Experiments," *American Psychologist*, April 1969, 24:409.

EXHIBIT 1-J Evaluators as Teachers

An evaluative study of a social program is justified to the extent that it facilitates the work of the polity. It therefore is to be judged primarily by its contribution to public thinking and to the quality of service provided subsequent to the evaluation. . . . An evaluation pays off to the extent that it offers ideas pertinent to pending actions and people think more clearly as a result. To enlighten, it must do more than amass good data. Timely communications—generally not "final" ones—should distribute information to the persons rightfully concerned, and those hearers should take the information into their thinking. To speak broadly, an evaluation ought *to inform and improve the operations of the social system.*

SOURCE: Quoted from Lee J. Cronbach and Associates, *Toward Reform of Program Evaluation* (San Francisco: Jossey-Bass, 1980), pp. 65-66.

inquiry in ways that do not mesh well with the perspectives of those who must make decisions about the program. For example, specifying variables so that they are well defined and measurable under scientific standards may fragment and, in some regards, trivialize what the policymakers see as complex and dynamic facets of the program. Similarly, meeting scientific standards for inferring causality, as when investigating program outcomes, may require such elaborate experimental controls that what is studied is no longer the program's services, but some contrived and constrained version of uncertain relevance to the actual program.

On the other hand, one cannot blithely dismiss scientific concerns in evaluation. Properly understood, what the scientific approach represents is a very considered attempt to produce conclusions that are valid and credible. Even when it falls short of this ideal, which is inevitable, such input makes a very important contribution to a decision-making context that otherwise is rife with self-interested perceptions and assertions, ideological biases, and undocumented claims about the way things are. But this statement, in turn, assumes that those conclusions meaningfully address aspects of the situation of concern to the decision-makers; if not, they may be praiseworthy for their validity and credibility, but still irrelevant.

In practice, therefore, the evaluator must struggle to find a workable balance between the emphasis to be placed on procedures that help ensure the validity of the evaluation findings and those that make the findings timely, meaningful, and useful to the consumers. Where that balance point should be will depend on the purposes of the evaluation, the nature of the program, and the political or decision-making context. In many cases, evaluations will justifiably be undertaken that are "good enough" for answering relevant policy and program questions even though program conditions or available resources prevent them from being the best possible designs from a scientific standpoint. For example, program sponsors concerned about whether a rehabilitation treatment for alcoholics is effective may find six-month follow-up interviews showing that few clients report heavy drinking to be very useful even though the data lack the experimental controls that permit this result to be confidently attributed to the influence of the program.

What further complicates an already difficult situation for evaluation planning is that there is often ambiguity about the identity of the ultimate users of the evaluation and which of the potential users should be given priority in the design. An evaluation generally has various potential audiences, some with very immediate interests in particular aspects of the program under investigation, some with broader interests in the type of intervention the particular program represents, and others falling somewhere in between. Occasionally, the purposes and priority users of an evaluation are defined so clearly and explicitly in advance that the evaluator has relatively little difficulty in balancing scientific and pragmatic considerations. For instance, an evaluation of a demonstration project on needle exchange to prevent AIDS among drug addicts funded by the National Institutes for Health may be clearly stipulated as a contribution to general policy-relevant knowledge about this intervention approach that should meet the highest possible scientific standards. On the other hand, an evaluator retained as a consultant by a program administrator to provide an assessment of a problematic unit in a community mental health center may understand quite clearly that the findings are for the sole purpose of informing certain decisions that administrator must make and, indeed, will not be reported outside the organization.

However, many program evaluation situations are not so clear-cut. Evaluation may be routinely required as part of funding or contract arrangements with the presumption that it will be generally informative to program managers, sponsors, and other interested parties. Or it may evolve from a collaboration between a service agency with a need for information for management purposes and a researcher with broader interests in the type of intervention that particular program provides. Indeed, given the effort and expense required for evaluation, there are probably more instances in which it is expected to be multipurpose than cases where the purpose and user are tightly specified. Unfortunately, the trade-offs between utility for program decisionmakers and scientific rigor are such that it is rarely possible to design an evaluation that serves both interests well. Thus, if evaluators choose to emphasize the

needs of the immediate consumers, they must be prepared for the possibility that the findings will be criticized on methodological grounds by more remote audiences, for example, applied researchers, other evaluators, or sophisticated policymakers or service professionals. This carries with it the risk that the credibility of the evaluation will be undermined, perhaps even in the eyes of the immediate consumers to whom it was directed. But if the evaluation emphasizes methodological quality at the expense of utility, it may satisfy those knowledgeable about research standards but be assailed by program stakeholders as too academic, ivory tower, and even irrelevant to the "real" program issues.

Some evaluation theorists champion utilization as the overriding concern and advocate evaluation that is designed around the specific information needs of individually identified target consumers with whom the evaluator collaborates very closely (e.g., Patton, 1997). The authors of review articles in applied research journals who attempt to synthesize available research on the effectiveness of various interventions, on the other hand, regularly deplore the poor methodological quality of evaluation studies and urge a higher standard. Some commentators want to have it both ways and press the view that evaluations should strive to have utility to program stakeholders *and* contribute to cumulative knowledge about social intervention (Lipsey, 1997). Our outlook, for the didactic purposes of this book, is that all these options are defensible, but not necessarily equally defensible in any given evaluation situation. This, then, presents yet another issue for which the evaluator will be required to make a judgment call and must attempt to tailor the evaluation design to the particular purposes and circumstances presented in each application.

Diversity in Evaluation Outlooks and Approaches

As the preceding discussion illustrates, the field of evaluation is not monolithic in conceptual outlook or methodological approach. On the contrary, it is a contentiously diverse field. The fundamental difference represented historically by Campbell and Cronbach represents but one instance of this diversity. Evaluation practitioners are drawn from a wide range of academic disciplines and professions with different orientations and methods, and this multidisciplinary mix has contributed significantly to the multiplicity of perspectives. Other differences in outlook are related to the motivations of evaluators and the settings in which they work. The solo practitioner who undertakes short-term evaluations on contract with local agencies and the tenured professor with long-term foundation funding will likely have quite divergent views on their evaluation activities.

As the field of evaluation has matured and become institutionalized, interest has developed in explicating the different postures toward evaluation and the methods preferred by leaders in various "camps." In particular, there is a growing interest in identifying congruent elements among different perspectives to advance what is referred to as "evaluation theory" (Shadish, Cook, and Leviton, 1991). Advocates of the evaluation theory movement envision the development of a theory that will serve as the basis for decision making by evaluators as they proceed with their work (see Exhibit 1-K).

Virtually all experienced evaluators see the need for better formulated guidelines as they face the various decision points that come up in any evaluation. Also, there is a need for such guidelines so that training the fledgling evaluator is not so heavily dependent on trial-and-error experience. But not every evaluator subscribes to the view that the field of evaluation has sufficiently clear boundaries to distinguish it conceptually from what goes on generally in the policy sciences or procedurally from the "rules" that guide applied social research. There is probably as much diversity in outlook among evaluators about the utility of evaluation theory as there is about the right way of doing evaluations.

EXHIBIT 1-K The Ideal Evaluation Theory

The ideal (never achievable) evaluation theory would describe and justify why certain evaluation practices lead to particular kinds of results across situations that evaluators confront. It would (a) clarify the activities, processes, and goals of evaluation; (b) explicate relationships among evaluative activities and the processes and goals they facilitate; and (c) empirically test propositions to identify and address those that conflict with research or other critically appraised knowledge about evaluation.

SOURCE: Quoted from William R. Shadish, Thomas D. Cook, and Laura C. Leviton, *Foundations of Program Evaluation: Theories of Practice* (Newbury Park, CA: Sage, 1991), pp. 30-31.

At present, therefore, we must acknowledge that evaluation is at least as much art as science, and perhaps should be and always will be. Inevitably, the evaluator's task is to creatively weave together many competing concerns and objectives into a tapestry in which different viewers can find different messages. We recognize, too, the difficulty of teaching an art form, especially via the written word. Teaching evaluation is analogous to training physicians to be diagnosticians. Any intelligent person can be taught to understand the results from laboratory tests, but a doctor becomes an astute diagnostician only through practice, experience, and attention to the idiosyncrasies of each individual case. In this sense, learning from a text can provide only part of the knowledge needed to become a capable evaluator.

WHO CAN DO EVALUATIONS?

Systematic evaluation is grounded in social science research techniques; hence, most evaluation specialists have had some social research training. But we should be quick to point out that there is great heterogeneity in the disciplinary and professional training of persons doing evaluations (see Exhibit 1-L). Ideally, every evaluator should be familiar with the full repertoire of social research methods. In practice, we can come close to this ideal only by continually broadening and deepening our technical knowledge by means all of us know about: keeping up with the literature, attending workshops and professional conferences, and learning from colleagues. Moreover, it would be deceptive to suggest that this or any textbook can teach someone how to *do* evaluations. There is no substitute for experience. What we do believe is that this book will provide an organized conceptual framework that identifies

Exhibit 1-L Diversity of the Members of the American Evaluation Association (in percentages)

Major Professional Responsibility		*Organizational Setting*		*Primary Discipline*	
Evaluation	28	College or university	40	Education	22
Research	19	Private business	12	Psychology	18
Administration	18	Nonprofit organization	11	Evaluation	14
Teaching	13	Federal government agency	10	Statistical methods	10
Consulting	8	State/local government agency	10	Sociology	6
Student	5	School system	4	Economics and political science	6
Other	9	Other	13	Organizational development	3
				Other	21

SOURCE: Adapted from *Evaluation Practice News* (October 1993); based on 2,045 AEA members as of June 1993.

the important issues and the options for addressing them.

Although knowledge of the concepts and methods instrumental to good program evaluation research is essential for conducting evaluations, it is important to note that a great deal of knowledge about the target problem area (e.g., crime, health, drug abuse) and the nature, range, and results of the interventions that have been used to address that problem are also required. This is necessary not only so the evaluator will understand the issues and context with which the program deals but so that an appropriate evaluation plan can be developed that reflects the reality of the program and existing knowledge relevant to such programs. At the most complex level, evaluation activities can be so technically complicated, sophisticated in conception, costly, and of such long duration that they require the dedicated participation of highly trained specialists at ease with the latest in social science theory, program knowledge, data collection methods, and statistical techniques. Such highly complex evaluations are usually conducted by specialized evaluation staffs. At the other extreme, there are many evaluation tasks that can be understood easily and carried out by persons of modest expertise and experience.

It is the purpose of this book to provide an introduction to the field for those whose current positions, professional interests, or natural curiosity inspire them to want to learn how evaluations are conducted. Studying the book is, of course, only a start along the path to becoming an expert in evaluation. We also aim to provide persons responsible for administering and managing human resource programs with sufficient understanding of evaluation tasks and activities to be able to judge for themselves what kinds of evaluations are appropriate to their programs and projects and to comprehend the results of evaluation studies of their programs. In brief, we have tried to provide a text that is helpful to those who conduct evaluations, those who commission them, those who oversee evaluation staffs, and those who are consumers of evaluation research.

SUMMARY

- Program evaluation is the use of social research methods to systematically investigate the effectiveness of social intervention programs. It draws on the techniques and concepts of social science disciplines and is intended to be useful for improving programs and informing social action aimed at ameliorating social problems.
- Modern evaluation research grew from pioneering efforts in the 1930s and burgeoned in the postwar years as new methodologies were developed that could be applied to the rapidly growing social program arena. The social policy and public administration movements have contributed to the professionalization of the field and to the sophistication of the consumers of evaluation research.
- The need for program evaluation is undiminished in the 1990s and may even be expected to grow. Indeed, contemporary concern over the allocation of scarce resources makes it more essential than ever to evaluate the effectiveness of social interventions.
- Evaluation must be tailored to the political and organizational context of the program to be evaluated. It typically involves assessment of one or more of five program domains: (a) the need for the program, (b) the design of the program, (c) the program implementation and service delivery, (d) the program impact or outcomes, and (e) program efficiency. Evaluation requires an accurate description of the program performance or characteristics at issue and assessment of them against relevant standards or criteria.
- In practice, program evaluation presents many challenges to the evaluator. Program circumstances and activities may change during the course of an evaluation, an appropriate balance must be found between scientific and pragmatic considerations in the evaluation design, and the wide diversity of perspectives and approaches in the evaluation field provide little firm guidance about how best to proceed with an evaluation.
- Most evaluators are trained either in one of the social sciences or in professional schools that offer applied social research courses. Highly specialized, technical, or complex evaluations may require specialized evaluation staffs. A basic knowledge of the evaluation field, however, is relevant not only to those who will perform evaluations but also to the consumers of evaluation research.

KEY CONCEPTS FOR CHAPTER 2

Formative evaluation	Evaluative activities undertaken to furnish information that will guide program improvement.
Summative evaluation	Evaluative activities undertaken to render a summary judgment on certain critical aspects of the program's performance, for instance, to determine if specific goals and objectives were met.
Target	The unit (individual, family, community, etc.) to which a program intervention is directed. All such units within the area served by a program comprise its target population.
Stakeholders	Individuals, groups, or organizations having a significant interest in how well a program functions, for instance, those with decision-making authority over it, funders and sponsors, administrators and personnel, and clients or intended beneficiaries.
Evaluation sponsor	The person(s), group, or organization that requests or requires the evaluation and provides the resources to conduct it.
Independent evaluation	An evaluation in which the evaluator has the primary responsibility for developing the evaluation plan, conducting the evaluation, and disseminating the results.
Participatory or collaborative evaluation	An evaluation organized as a team project in which the evaluator and representatives of one or more stakeholder groups work collaboratively in developing the evaluation plan, conducting the evaluation, or disseminating and using the results.
Empowerment evaluation	A participatory or collaborative evaluation in which the evaluator's role includes consultation and facilitation directed toward the development of the capabilities of the participating stakeholders to conduct evaluation on their own, to use it effectively for advocacy and change, and to have some influence on a program that affects their lives.
Evaluation questions	A set of questions developed by the evaluator, evaluation sponsor, and other stakeholders; the questions define the issues the evaluation will investigate and are stated in terms such that they can be answered using methods available to the evaluator in a way useful to stakeholders.
Needs assessment	An evaluative study that answers questions about the social conditions a program is intended to address and the need for the program.
Assessment of program theory	An evaluative study that answers questions about the conceptualization and design of a program.
Assessment of program process	An evaluative study that answers questions about program operations, implementation, and service delivery. Also known as a process evaluation or an implementation assessment.
Impact assessment	An evaluative study that answers questions about program outcomes and impact on the social conditions it is intended to ameliorate. Also known as an impact evaluation or an outcome evaluation.
Efficiency assessment	An evaluative study that answers questions about program costs in comparison to either the monetary value of its benefits or its effectiveness in terms of the changes brought about in the social conditions it addresses.

CHAPTER 2

TAILORING EVALUATIONS

Every evaluation must be tailored to its program. The tasks that evaluators undertake depend on the purposes of the evaluation, the conceptual and organizational structure of the program, and the resources available. Formulating an evaluation plan therefore requires the evaluator to first explore these aspects of the evaluation situation with the evaluation sponsor and such other stakeholders as policymakers, program personnel, and program participants. Based on this reconnaissance and negotiation with the key stakeholders, the evaluator can then develop a plan that identifies the evaluation questions to be answered, the methods to be used to answer them, and the relationships to be developed with the stakeholders during the course of the evaluation.

No hard-and-fast guidelines direct the process of investigating the evaluation situation and designing an evaluation—it is necessarily a creative and collaborative endeavor. Nonetheless, achieving a good fit between the evaluation plan and the program circumstances usually involves attention to certain critical themes. It is essential, for instance, that the evaluation plan be responsive to the purposes of the evaluation as understood by the evaluation sponsor and other central stakeholders. An evaluation intended to provide feedback to program decisionmakers so that the program can be improved will take a different approach than one intended to help funders determine if a program should be terminated. In addition, the evaluation plan must reflect an understanding of how the program is designed and organized so that the questions asked and the data collection arranged will be appropriate to the circumstances. Finally, any evaluation, of course, will have to be designed within the constraints of available time, personnel, funding, and other such resources.

Although the particulars are diverse, the basic program circumstances for which evaluation is requested typically represent one of a small number of recognizable variations. Consequently, the evaluation designs that result from the tailoring process tend to be adaptations of one or more of a set of familiar evaluation approaches or schemes. In practice, therefore, tailoring an evaluation is often primarily a matter of selecting and adapting these schemes to the specific circumstances of the program to be evaluated. One set of evaluation approaches is defined around the nature of the evaluator-stakeholder interaction. Evaluators may function relatively independently or work quite collaboratively with stakeholders in designing and conducting the evaluation. Another distinct set of evaluation approaches is organized around common combinations of evaluation questions and the usual methods for answering them. Among these are evaluation schemes for assessing social problems and needs, program theory, program process or implementation, program impact or outcome, and program efficiency.

One of the most challenging aspects of evaluation is that there is no "one size fits all" approach. Every evaluation situation has its unique profile of characteristics, and the evaluation design must involve an interplay between the nature of the evaluation situation, on the one side, and the evaluator's repertoire of approaches, techniques, and concepts, on the other. A good evaluation design is one that fits the circumstances while yielding credible and useful answers to the questions that motivate it. This chapter provides an overview of the issues and considerations the evaluator should take into account when tailoring an evaluation plan to accomplish these purposes.

WHAT ASPECTS OF THE EVALUATION PLAN MUST BE TAILORED?

Evaluation designs may be quite simple and direct, perhaps addressing only one narrow question such as whether using a computerized instructional program helps a class of third graders read better. Or they may be prodigiously complex, as in a national evaluation of the operations and effects of a diverse set of programs for reducing substance abuse in multiple urban sites. Fundamentally, however, we can view any evaluation as structured around three issues:

The questions the evaluation is to answer. An endless number of questions might be raised about any social program by a wide range of interested parties. There may be concerns about such matters as the needs of the target population and whether they are being adequately reached and served, the management and operation of the program, the effectiveness of services, whether the program is having its desired impact, and its costs and efficiency. No evaluation can, nor generally should, attempt to address all such concerns. A central feature of an evaluation design, therefore, is a specification of the guiding purpose of the evaluation and the corresponding questions on which it will focus. Later in this chapter, and in more detail in Chapter 3, we discuss the nature of evaluation questions, how they can be derived, and some of the factors that influence the priority they should be given.

The methods and procedures the evaluation will use to answer the questions. An important aspect of the evaluator's distinctive expertise is knowing how to obtain useful, timely, and credible information about the various dimensions of program performance that are to be evaluated. A large repertoire of social research techniques and conceptual tools are available for this task. An evaluation design must identify the methods that will be used to answer each of the questions at issue and organize them into a feasible work plan. Moreover, the methods selected must not only be capable of providing meaningful answers to the questions but also must be practical while still providing the degree of scientific rigor appropriate to the evaluation circumstances. Most of the rest of this book (Chapters 4-11) is devoted to consideration of evaluation methods and the circumstances in which they are applicable.

The nature of the evaluator-stakeholder relationship. One of the most important lessons from the first several decades of experience with systematic evaluation is that there is nothing automatic about the assimilation and use of evaluation findings by the stakeholders presumed interested in them. Part of an evaluation design, therefore, is a plan for effectively interacting with program stakeholders to identify and clarify the issues, conduct the evaluation, and make effective use of the evaluation findings. This interaction may be highly collaborative, with the evaluator serving as a consultant or facilitator to a group of stakeholders who take primary responsibility for planning, conducting, and using the evaluation. Or the evaluator may take that responsibility but seek essential guidance and information from the stakeholders. In addition, an evaluation plan should indicate which audiences are to receive which information at what times, what the nature and schedule of written reports and oral briefings will be, and how broadly findings are to be disseminated beyond the evaluation sponsor. The evaluator-stakeholder relationship is discussed later in this chapter and in Chapter 12.

WHAT CONSIDERATIONS SHOULD GUIDE EVALUATION PLANNING?

Many aspects of the program and the circumstances of the evaluation will necessarily shape the evaluation design. Some of these involve general considerations of almost universal relevance to evaluation planning, but others will be specific to the particular situation of each evaluation. Development of the evaluation plan, therefore, must be guided by a careful analysis of the evaluation context. The more significant considerations for that analysis can be organized into three categories, having to do with (a) the purposes of the evaluation, (b) the program structure and circumstances, and (c) the resources available for the evaluation. All these topics will receive later attention in the course of discussion about the specific aspects of the evaluation plan they most influence; an overview is provided here.

The Purposes of the Evaluation

Evaluations are initiated for many reasons and may have quite different purposes from one situation to another. They may be intended to help management improve a program; support advocacy by supporters or critics; gain knowledge about program effects; provide input to decisions about program funding, structure, or administration; respond to political pressures; or have any of a number of such purposes individually or in combination. One of the first determinations the evaluator must make, therefore, is just what those purposes are. This is not always a simple matter. Some statement of the purposes generally accompanies the initial request for an evaluation, but these announced purposes rarely tell the whole story and sometimes are only rhetorical. Furthermore, evaluations may be routinely required in a program situation or sought simply because it is presumed to be a good idea without any distinct articulation of its purposes or the sponsor's intent (see Exhibit 2-A).

The prospective evaluator determines the purposes of the evaluation by attempting to establish as firmly as possible who wants the evaluation, what they want, and why they want it. There is no cut-and-dried method for doing this, but it is usually best approached the way a journalist would try to dig out a story. That is, source documents should be examined, key informants with different vantage points on the situation should be interviewed, and pertinent history and background should be uncovered. Although the details will vary greatly, evaluations are generally done for one or more of the following broad reasons (Chelimsky, 1997):

EXHIBIT 2-A Does Anybody Want This Evaluation?

Our initial meetings with the Bureau of Community Services administrators produced only vague statements about the reasons for the evaluation. They said they wanted some information about the cost-effectiveness of both New Dawn and Pegasus and also how well each program was being implemented. . . . It gradually became clear that the person most interested in the evaluation was an administrator in charge of contracts for the Department of Corrections, but we were unable to obtain specific information concerning where or how the evaluation would be used. We could only discern that an evaluation of state-run facilities had been mandated, but it was not clear by whom.

SOURCE: Quoted from Dennis J. Palumbo and Michael A. Hallett, "Conflict Versus Consensus Models in Policy Evaluation and Implementation," *Evaluation and Program Planning*, 1993, 16(1):11-23.

program improvement, accountability, knowledge generation, and political ruses or public relations.

Program improvement. The evaluation findings may be intended to furnish information that will guide program improvement. Such evaluation is often called *formative evaluation* (Scriven, 1991) because its purpose is to help form or shape the program to perform better (for an example, see Exhibit 2-B). The audiences for the findings of formative evaluation typically are the program planners (in the case of programs in the planning stage) or program administrators, oversight boards, or funders with an interest in optimizing program effectiveness. The information desired by these persons may relate to the need for the program, the program concept and design, its implementation, its impact, or its efficiency. Typically, the evaluator in this situation will work closely with program management and other stakeholders in designing, conducting, and reporting the evaluation. Evaluation for program improvement characteristically emphasizes findings that are timely, concrete, and immediately useful. Correspondingly, the communication between the evaluator and the respective audiences about the findings may occur regularly throughout the evaluation and be relatively informal.

Accountability. The use of social resources such as taxpayer dollars by human service programs is justified on the grounds that these programs make beneficial contributions to society. It follows that persons with significant responsibilities for such social investments will expect programs to manage resources effectively and efficiently and actually produce the intended benefits. Evaluation may be conducted, therefore, to determine if these expectations are met. Such evaluation is often called *summative evaluation* (Scriven, 1991) because its purpose is to render a summary judgment on certain critical aspects of the program's performance (Exhibit 2-C provides an example). The findings of summative evaluation are usually intended for decisionmakers with major roles in program oversight, for example, a

EXHIBIT 2-B A Stop-Smoking Telephone Help Line That Nobody Called

Formative evaluation procedures were used to help design a "stop smoking" hotline for 2,148 adult smokers in a cancer control project sponsored by a health maintenance organization (HMO). Phone scripts for use by the hotline counselors and other aspects of the planned services were discussed with focus groups of smokers and reviewed in telephone interviews with a representative sample of HMO members who smoked. Feedback from these informants led to refinement of the scripts, hours of operation arranged around the times participants said they were most likely to call, and advertising of the service through newsletters and "quit kits" routinely distributed to all project participants. Despite these efforts, an average of less than three calls per month was made during the 33 months the hotline was in operation, about a 2.4% use rate by the target population. To further assess this disappointing response, comparisons were made with similar services around the country. This revealed that 1%-2% use rates were typical but the other hotlines served much larger populations and therefore received many more calls. The program sponsors concluded that to be successful, the smoker's hotline would have to be offered to a larger population and be intensively publicized.

SOURCE: Adapted from Russell E. Glasgow, H. Landow, J. Hollis, S. G. McRae, and P. A. La Chance, "A Stop-Smoking Telephone Help Line That Nobody Called," *American Journal of Public Health*, February 1993, 83:252-253.

EXHIBIT 2-C U.S. General Accounting Office Assesses Early Effects of the Mammography Quality Standards Act

The Mammography Quality Standards Act of 1992 required the Food and Drug Administration (FDA) to administer a code of uniform standards for mammogram-screening procedures in all the states. When the act was passed, Congress was concerned that access to mammography services might decrease because providers would choose to drop them rather than upgrade operations to comply with the new standards. The U.S. General Accounting Office (GAO) was asked to assess the early effects of implementing the act and report back to Congress. They found that the FDA had taken a gradual approach to implementing the act's requirements, which had helped to minimize adverse effects on access. The FDA inspectors had not closed many facilities that failed to meet certification standards; rather, they had given them additional time to correct the problems found during inspections and to meet the new quality assurance requirements. Only a relatively small number of facilities had terminated their mammography services and those were generally small-volume providers located within 25 miles of another certified facility. The GAO concluded that the Mammography Quality Standards Act was having a positive effect on the quality of mammography services, as Congress had intended.

SOURCE: Adapted from U.S. General Accounting Office, *Mammography Services: Initial Impact of New Federal Law Has Been Positive*. Report 10/27/95, GAO/HEHS-96-17 (Washington, DC: General Accounting Office, 1995).

funding agency, governing board, legislative committee, political decisionmaker, or upper management, but may also be of interest to critics, constituents, and concerned citizens outside the formal decision-making channels. Summative evaluation may influence such significant decisions as program continuation, allocation of resources, restructuring, or legal action. For this reason, such evaluation often requires information of sufficient credibility under scientific standards to provide a confident basis for action and to withstand criticism aimed at discrediting it. The evaluator may be expected to function relatively independently in planning, conducting, and reporting the evaluation with input from, but no direct decision-making participation by, stakeholders. Similarly, it may be important to avoid premature or careless conclusions and, therefore, communication of the evaluation findings to the respective audiences may be relatively formal, rely chiefly on written reports, and occur primarily at the end of the evaluation.

Knowledge generation. Some evaluations are not intended to directly inform decisions related to specific programs in place or contemplated but, rather, mainly describe the nature and effects of an intervention for broader purposes and audiences. The intervention at issue, for instance, might be a demonstration configured expressly to try out a promising concept such as integrated services for children with mental health problems or monthly visits by nurses to pregnant women at risk of premature births (see Exhibit 2-D for another example). A similar situation occurs when an academic researcher arranges to study an intervention with interesting characteristics, for example, an innovative science curriculum, to contribute to knowledge about that particular form of intervention. Because evaluations of this sort are intended to make contributions to the social science knowledge base, they are usually conducted in a scientific framework using the most rigorous methods feasible. The audience for the resulting findings may include the research sponsors in cases of demonstration projects or externally funded research. Beyond that, however, the audience is generally quite diffuse—all those interested in the particular type of program or, perhaps, the particular methods used to study it. Dissemination of the evaluation findings in these situations is most likely through scholarly journals, conferences, and other such professional outlets. These knowledge generation studies may turn out to be useful for the development of new public programs as program developers draw on social science research for program ideas.

Political ruses or public relations. Sometimes, the true purpose of the evaluation, at least for those who initiate it, has little to do with actually obtaining information about program performance. It is not unusual, for instance, for program administrators or boards to launch an evaluation because they believe it will be good public relations and might impress funders or political decisionmakers. Occasionally, an evaluation is commissioned to provide a public context for a decision that has already been made behind the scenes to terminate a program, fire an administrator, or the like. Or the evaluation may be a delaying tactic to appease critics and defer difficult decisions, rather like appointing a committee to study a problem rather than acting on the problem.

Virtually all evaluations have some elements of political maneuvering and public relations among their instigating motives, but when these are the principal purposes, the prospective evaluator is presented with a difficult dilemma. The evaluation must either be guided by the political or public relations purposes, which may compromise its integrity, or focus on program performance issues that are of no real interest to those commissioning the evaluation and may even be threatening. In either case, the evaluator would be well advised to try to avoid such situations. If a lack of serious intent becomes evident during the initial exploration of the evaluation context, proceeding with an evaluation plan would not generally be wise. Instead, the prospective evaluator may wish to assume an "evaluation consultant" role and assist the relevant parties to clarify the nature of evaluation, identify appropriate and realistic expectations, and redirect the effort toward more appropriate uses.

EXHIBIT 2-D Testing an Innovative Treatment Concept for Pathological Gambling

Pathological gambling is characterized by a loss of control over gambling impulses, lies about the extent of gambling, family and job disruption, stealing money, and chasing losses with additional gambling. Though recent increases in the availability of gambling have led to corresponding increases in the prevalence of pathological gambling, few treatment programs have been developed to help the victims of this disorder. Research on the psychology of gambling has shown that problem gamblers develop an illusion of control such that they believe they can employ strategies that will increase their winnings despite the inherent randomness of games of chance. A team of clinical researchers in Canada hypothesized that a treatment based on "cognitive correction" of these erroneous beliefs would be an effective therapy. Because excessive gambling leads to financial problems and interpersonal difficulties, they combined their cognitive intervention with problem-solving and social skills training.

To test their treatment concept, the researchers used media advertisements and referrals from health providers to recruit 40 pathological gamblers willing to accept treatment. These were randomly assigned to the treatment or control group and measures of pathological gambling, perception of control, desire to gamble, self-efficacy perception, and frequency of gambling were taken at various intervals before and after the treatment period. The results showed significant changes in the treatment group on all outcome measures with maintenance of the gains at 6- and 12-month follow-up. However, the results may have been compromised by high attrition—8 of the 20 gamblers who began treatment and 3 of the 20 in the control group dropped out, a common occurrence during intervention for addictive problems. Despite this limitation, the researchers concluded that their results were strong enough to demonstrate the effectiveness of their treatment concept.

SOURCE: Adapted from Caroline Sylvain, Robert Ladouceur, and Jean-Marie Boisvert, "Cognitive and Behavioral Treatment of Pathological Gambling: A Controlled Study," *Journal of Consulting and Clinical Psychology*, 1997, 65(5):727-732.

The Program Structure and Circumstances

No two programs are identical in their organizational structure and environmental,

social, and political circumstances, even when they ostensibly provide the "same" service. The particulars of a program's structure and circumstances constitute major features of the evaluation situation to which the evaluation plan must be tailored. Although there is a myriad of such particulars, three broad categories are especially important to evaluators because of their pervasive influence on evaluation design and implementation:

- The stage of program development—whether the program being evaluated is new or innovative, established but still developing or undergoing restructuring, or established and presumed stable.
- The administrative and political context of the program, in particular, the degree of consensus, conflict, or confusion among stakeholders about the values or principles the program embodies, its mission and goals, or its social significance.
- The structure of the program, including both its conceptual and organizational makeup. This involves the nature of the program rationale; the diversity, scope, and character of the services provided and of the target populations for those services; location of service sites and facilities; administrative arrangements; record-keeping procedures; and so forth.

The influence of these factors will be considered in relation to various specific aspects of evaluation design discussed throughout the remainder of this book; we provide a brief orientation here.

The Stage of Program Development

The life of a social program can be thought of as a developmental progression in which different questions are at issue at different stages and, therefore, different evaluation approaches must be applied to answer those questions (see Exhibit 2-E). Assessment of a program still in the early stages of planning will be a distinctly different endeavor than assessment of a well-established program. Similarly, assessment of an established program for which restructuring is contemplated or under way will raise different concerns than a program presumed stable in its basic operations and functions.

When new programs are initiated, especially innovative ones, evaluation is often requested to examine the social needs the program should address, the program design and objectives, the definition of its target population, the expected outcomes, and the means by which it assumes those outcomes can be attained. These issues are especially relevant during the planning phase when the basic design is being formulated and changes can be made relatively easily. The evaluator, therefore, may function as a planning consultant before the program is launched by helping to assess and improve the program design as it is developed. Assessment of the program conceptualization may also be the focal point of an evaluation after the planning phase when the program is in the early stage of implementation. Decision-makers associated with young programs are often open to some amount of reformulation of the program model and may wish to have it assessed to ensure that their approach is as good as it can be.

The following examples illustrate the role of evaluators in the early stages of program development:

• A small New England city wanted to establish an emergency shelter for homeless persons. To determine how many beds should be provided, the city funded a nighttime survey, attempting to count all persons sleeping in public places such as bus stations, parks, or store entrance ways.

• A program to increase public awareness of risk factors in cardiovascular diseases and encourage exercise and proper diet attempted to organize discussion groups among employees of local firms. The evaluators found that this approach was largely unsuccessful because workers were reluctant to form discussion groups. The evaluators suggested a more successful approach using existing organized groups such as churches, fraternal organizations, and clubs.

Sometimes evaluations of new programs are expected to address questions of impact and efficiency, but the unsettled nature of the programs in their beginning years most often makes those issues premature. It can easily take a year or more for a new program to establish facilities, acquire and train staff, make contact with the target population, and develop its services to the desired level. During this period, it may not be realistic to expect much impact on the social conditions toward which the program is directed. Formative evaluation aimed at clarifying target population needs, improving program operations, and enhancing the quality of service delivery, using approaches

EXHIBIT 2-E Stages of Program Development and Related Evaluation Functions

Stage of Program Development	*Question to Be Asked*	*Evaluation Function*
1. Assessment of social problems and needs	To what extent are community needs and standards met?	Needs assessment; problem description
2. Determination of goals	What must be done to meet those needs and standards?	Needs assessment; service needs
3. Design of program alternatives	What services could be used to produce the desired changes?	Assessment of program logic or theory
4. Selection of alternative	Which of the possible program approaches is best?	Feasibility study; formative evaluation
5. Program implementation	How should the program be put into operation?	Implementation assessment
6. Program operation	Is the program operating as planned?	Process evaluation; program monitoring
7. Program outcomes	Is the program having the desired effects?	Outcome evaluation
8. Program efficiency	Are program effects attained at a reasonable cost?	Cost-benefit analysis; cost-effectiveness analysis

SOURCE: Adapted from S. Mark Pancer and Anne Westhues, "A Developmental Stage Approach to Program Planning and Evaluation," *Evaluation Review*, 1989, 13(1):56-77.

such as those discussed later in Chapters 4-6, is likely to be more apt in these cases.

Although the evaluation of new programs represents an important activity for the field, by far the greater effort goes into assessing established programs. Evaluating these programs requires first understanding their social and political history. Most well-established social programs have sprung from long-standing ameliorative efforts, and unless some crisis necessitates consideration of fundamental change, they are constrained to their traditional forms and approaches. Often there is considerable opposition from some stakeholders to any questioning of their fundamental assumptions or the ways they have been put into place. The value of such well-entrenched programs as Social Security pensions, guidance counselors in schools, vocational programs for disabled persons, parole supervision for released convicts, and community health education for the prevention of diseases is taken for granted.

Evaluation of established, stable programs, therefore, rarely focuses on assessing the underlying program conceptualization. It is more likely that attention will be directed toward such issues as coverage, effective service delivery, and the impact and efficiency of those services. However, if the program is very large and well established, it can be difficult to evaluate impact and efficiency, especially if it is a full-coverage program that provides services to virtually the entire eligible population. In such cases, the evaluator has limited ability to develop credible depictions of what things would be like in the absence of the program as a baseline for assessing its impact. Often, evaluation of such programs is directed toward assessing the extent to which the program objectives are explicit and relevant to the interests of program sponsors, staff, and other stakeholders, whether the program is conforming to program plans, and whether it is reaching the appropriate target population. For example, the U.S. Department of Agriculture conducts periodic studies of participation in the food stamps program to measure the extent to which eligible households are enrolled and to guide outreach efforts to increase participation (Trippe, 1995).

Sometimes, however, evaluation is sought for established programs primarily because the program status quo has been called into question. This may be due to external pressures such as political attack, competition, mounting program costs, or dramatic changes in the target population served. Or it may occur because program sponsors and staff are dissatisfied with the effectiveness of their interventions and wish to bring about improvement. In either event, some restructuring may be considered, and evaluation is sought to guide that change. In circumstances such as these, the evaluation may focus on any and all aspects of the program. Questions might be raised about the need for the program, its conceptualization and design, its operations and implementation, and its impact and efficiency.

The federal food stamps program mentioned above, for instance, has been a national program for more than two decades. It is intended to increase the quantity and quality of food consumed by poor households by providing them with food stamps redeemable only by purchasing approved foods at grocery stores. The Department of Agriculture contemplated abandoning food stamps and issuing checks instead, thereby eliminating the high costs of printing, distributing, and redeeming an earmarked currency. To test the effects of cashing out food stamps, it started four experiments comparing the food consumption in households receiving food stamps with the food consumption of households receiving the same dollar amount of benefits in the form of checks

(Fraker, Martini, and Ohls, 1995). Significant differences were found: Households receiving checks purchased less food than those that received food stamps. The Department of Agriculture therefore decided to retain food stamps.

The Administrative and Political Context of the Program

Except possibly for academic researchers who conduct an evaluation study on their own initiative for knowledge generation purposes, evaluators are not free to establish their own definitions of what the program is about, its goals and objectives, and what evaluation questions should be addressed. The evaluator interacts with the evaluation sponsor, program personnel, and other program stakeholders to develop this essential background. Somewhat different perspectives from these various groups are to be expected and, in most instances, the evaluator will attempt to develop an evaluation plan that reflects all significant views and concerns or, at least, is compatible with the prevailing views among the major parties.

If significant stakeholders are in substantial conflict about the mission, goals, probity, procedures, or presenting issues for the program, it presents an immense difficulty for evaluation design (see Exhibit 2-F). The evaluator can attempt to incorporate the conflicting perspectives into the design, but this may not be an easy task. The evaluation sponsors may be unwilling to embrace the inclusion of issues and perspectives from groups they view as adversaries. Furthermore, the issues and perspectives may be so different that it is difficult to incorporate them in a single evaluation plan or to do so may require more time and resources than are available. For instance, it would be challenging to design an evaluation that simultaneously addressed effectiveness questions generated by stakeholders who asserted that the purpose of a program for dysfunctional families was to protect the children from abuse and, therefore, should readily support removing them from the home if there is suspicion of abuse, and those generated by stakeholders insisting that the purpose was to keep families intact and help them work out their problems. Each of these perspectives entails different objectives, and procedures to attain those objectives, that have correspondingly different conceptions of program effectiveness associated with them. Although the evaluator might attempt to design an evaluation that would encompass these different perspectives and thus inform both sets of stakeholders, such an effort would require a careful balance in determining what data to collect and what criteria to apply to interpret them.

Alternatively, the evaluator could plan the evaluation from the perspective of one of the stakeholders, typically the evaluation sponsor or some other stakeholder designated by the evaluation sponsor. This, of course, will not be greeted with enthusiasm by stakeholders with conflicting perspectives and they may well oppose the evaluation and criticize the evaluator. The challenge to the evaluator is to be clear and straightforward about the perspective represented in the evaluation and the reasons for it, despite the objections. It is important to recognize that it is not necessarily wrong to plan and conduct an evaluation from the perspective of one stakeholder without giving strong representation to conflicting views. Nonetheless, evaluators generally solicit input from all the major stakeholders and attempt to incorporate their concerns so that the evaluation plan will be as comprehensive as possible and the results as useful as possible. Where there are conflict-

EXHIBIT 2-F Stakeholder Conflict Over Home Arrest Program

In an evaluation of a home arrest program using electronic monitoring for offenders on parole, the evaluators made the following comments about stakeholder views:

> There were numerous conflicting goals that were considered important by different agencies, including lowering costs and prison diversion, control and public safety, intermediate punishment and increased options for corrections, and treatment and rehabilitation. Different stakeholders emphasized different goals. Some legislators stressed reduced costs, others emphasized public safety, and still others were mainly concerned with diverting offenders from prison. Some implementors stressed the need for control and discipline for certain "dysfunctional" individuals, whereas others focused on rehabilitation and helping offenders become reintegrated into society. Thus, there was no common ground for enabling "key policymakers, managers, and staff" to come to an agreement about which goals should have priority or about what might constitute program improvement.

SOURCE: Dennis J. Palumbo and Michael A. Hallett, "Conflict Versus Consensus Models in Policy Evaluation and Implementation," *Evaluation and Program Planning*, 1993, 16(1):11-23.

ing perspectives, however, it is not inappropriate for an evaluation sponsor to seek information relevant to its perspective or the evaluator to conduct such an evaluation even if some stakeholder views are given little or no influence.

Suppose, for instance, that the funding sponsors for a program to provide job training to the hard-core unemployed have concerns about whether a program is "creaming" the cases that are easy to work with, providing services that are more weighted toward vocational counseling than job skill training, and are inefficiently organized. They might quite appropriately commission an evaluation to examine these questions. Program managers and their advocates, on the other hand, may have a sharply conflicting perspective that justifies their selection of clients, training program, and management practices. A conscientious evaluator will listen to the managers' perspective and encourage their input so that the evaluation design can be as sensitive as possible to the realities of the program and the legitimate concerns of management about misrepresentation of what they are doing and why. But the evaluation design should, nonetheless, be developed primarily from the perspective of the evaluation sponsors and the issues that concern them. The evaluator's primary responsibilities are simply to be clear about the perspective the evaluation takes, so there is no misunderstanding, and to treat the program personnel fairly and honestly.

Another approach to situations of stakeholder conflict is for the evaluator to attempt to design an evaluation that facilitates better understanding among the conflicting parties about the aspects of the program at issue. This might be done, for instance, by efforts to clarify the different concerns, assumptions, and perspectives of the parties; some portion of such

conflicts often involves matters that the evaluator can examine and report on in ways that inform all parties. For instance, parents of special education children may believe that their children are stigmatized and discriminated against when mainstreamed in regular classrooms. Teachers may feel equally strongly that this is not true. A careful observational study of the interaction of regular and special education children conducted by the evaluator may reveal that there is a problem, but that it occurs outside the classroom on the playground and during other informal interactions among the children.

Where stakeholder conflict is deep and hostile, it may be based on such profound differences in political values or ideology that no matter how comprehensive and ecumenical, an evaluation cannot conjoin them. One school of thought in the evaluation field holds that all program situations are of this sort and that it is the central feature to which the evaluator must attend. In this view, the social problems that programs address, the programs themselves, and the meaning and importance of those programs are all social constructions that will inevitably differ for different individuals and groups. Thus, rather than focus on program objectives, decisions, outcomes, and the like, evaluators are advised to engage directly the diverse claims, concerns, issues, and values put forth by the various stakeholders.

Guba and Lincoln (1989), the leading proponents of this particular construction of the evaluation enterprise, have argued that the proper role of the evaluator is to facilitate interpretive dialogue among the program stakeholders. Correspondingly, the primary purpose of the evaluation is to facilitate a negotiation among the stakeholders from which a more shared construction of the value and social significance of the program can emerge that still respects the pluralism of ideologies and concerns represented by the different stakeholders. Additional discussions of this perspective can be found in Guba and Lincoln (1987, 1989, 1994). It may have particular appeal for the evaluator working in contexts where the evaluation is highly politicized or stakeholder values and views about the program are strongly divergent.

Finally, the evaluator must realize that despite best efforts to communicate effectively and develop an appropriate, responsive evaluation plan, program stakeholders owe primary allegiance to their own positions and political alignments. This means that sponsors of evaluation and other stakeholders may turn on the evaluator and harshly criticize the evaluation if the results contradict the policies and perspectives they advocate. Thus, even those evaluators who do a superb job of working with stakeholders and incorporating their views and concerns in the evaluation plan should not expect to be acclaimed as heroes when the results are in. The multiplicity of stakeholder perspectives makes it likely that no matter how the results come out, someone will be unhappy. Evaluators work in a political environment, and it is the nature of such environments that stakeholders will often react to evidence contrary to their interests with a vigorous attempt to discredit it and those who produced it. It may matter little that everyone agreed in advance on the evaluation questions and the plan for answering them or that each stakeholder group understood that honest results might not favor its position. Nonetheless, it is highly advisable for the evaluator to give early attention to the identification of stakeholders, working out a strategy for minimizing discord in the evaluation due to their different perspectives, and conditioning their expectations for the nature and significance of the evaluation results.

The Conceptual and Organizational Structure of the Program

It is a simple truism that if authoritative stakeholders do not have a clear idea about what a program is supposed to be doing, it will be difficult to evaluate how well it is doing it. One factor that shapes the evaluation design, therefore, is the nature of the program conceptualization, that is, the distinctness and explicitness of its plan of operation, the logic that connects its activities to the intended outcomes, and the rationale provided for why it does what it does. This conceptual structure or *program theory* can itself be a focus of evaluation.

If there is significant uncertainty about whether the program conceptualization is appropriate for the social problem the program addresses, it may make little sense for the evaluation design to focus on how well those concepts are implemented. In such cases, the evaluation activities may be more usefully devoted to assessing and better developing the program plan.

In the planning stages of a new program, establishing the program design is a major activity and its nature and details are usually easily identified and articulated. The participation of an evaluator, however, often helps sharpen and shape the conceptualization to make it both more explicit and more useful for identifying key issues of program performance. After the planning stage, especially for well-established programs, program personnel or sponsors generally find little need or opportunity for identifying and reviewing basic assumptions and expectations in any systematic manner. Everyday practice and routine operating procedures tend to dominate, and personnel may find it difficult to articulate the underlying program rationale or agree on any single version of it. For instance, the personnel in a counseling agency under contract to the school district to work with children who are having academic problems may be quite articulate in describing their counseling theories, goals for clients, and therapeutic techniques. But they may have difficulty expressing and agreeing on a view of how their focus on improving family communication is supposed to translate into better grades. However, the evaluator needs some understanding of their assumptions on this matter to plan any assessment of the program's overt performance or outcomes. Correspondingly, the more explicit and cogent the program conceptualization, the less specific assessment it may require in the evaluation plan and the easier it will be to identify the program functions and effects on which the evaluation should focus.

At a more concrete level, the organizational structure of the program must also be taken into consideration when planning an evaluation. Such program characteristics as multiple services or target populations, distributed service sites or facilities, or extensive programmatic collaboration with other organizational entities have powerful implications for the nature and range of evaluation questions to be covered, data collection procedures, resources required for the evaluation, and stakeholder groups to involve. Organizational structures that are larger, more complex, more decentralized, and more geographically dispersed will present greater practical difficulties than their simpler counterparts. In such cases, a team of evaluators is often needed, with resources and time proportionate to the size and complexity of the program. The challenges of evaluation for complex, multisite programs are sufficiently daunting that planning and conducting

EXHIBIT 2-G Multisite Evaluations in Criminal Justice: Structural Obstacles to Success

Besides the usual methodological considerations involved in conducting credible evaluations, the structural features of criminal justice settings impose social, political, and organizational constraints that make multisite evaluations difficult and risky. To begin, the system is extremely decentralized. Police departments, for example, can operate within the province of municipalities, counties, campuses, public housing, mass transit, and the states. The criminal justice system is also highly fragmented. Cities administer police departments and jails; counties administer sheriffs' and prosecutors' offices, jails, and probation agencies; state governments run the prisons. Agencies are embedded in disparate political settings, each with its own priorities for taxing and spending. In addition, criminal justice agencies foster a subculture of secrecy concerning their work that has serious consequences for evaluators, who are readily seen as "snoops" for management, the courts, or individuals with political agendas. Line staff easily adopt an "us against them" mentality toward outside evaluators. Also, criminal justice agencies generally exist in highly charged political environments. They are the most visible components of local government, as well as the most expensive, and their actions are frequently monitored by the media, who historically have assumed a watchdog or adversarial posture toward the system. Finally, the criminal justice system operates within a context of individual rights—legal constraint in procedural issues, an unwillingness to risk injustice in individual cases, and a stated (though not actually delivered) commitment to providing individualized treatment. This translates, for example, into a general aversion to the concept of random or unbiased assignment, the hallmark of the best designs for yielding interpretable information about program effects.

SOURCE: Adapted from Wesley G. Skogan and Arthur J. Lurigio, *Multisite Evaluations in Criminal Justice Settings: Structural Obstacles to Success*, New Directions for Evaluation, no. 50 (San Francisco: Jossey-Bass, summer 1991), pp. 83-96.

them are distinct topics of discussion in the evaluation literature (see Exhibit 2-G; Turpin and Sinacore, 1991).

Equally important is the nature and structure of the particular intervention or service the program provides. The easiest interventions to evaluate are those that are discrete, one-shot events (e.g., serving meals to homeless persons) expected to have relatively immediate observable effects (they are not hungry). The organizational activities and delivery systems for such interventions are usually relatively straightforward (soup kitchen), the service itself is uncomplicated (hand out meals), and the outcomes are direct (people eat). These features greatly simplify the evaluation questions likely to be raised, the data collection required to address them, and the interpretation of the findings.

The most difficult interventions to evaluate are those that are diffuse in nature (community organizing), extend over long time periods (an elementary school math curriculum),

vary widely across applications (eclectic psychotherapy), or have expected outcomes that are long term (preschool compensatory education) or indistinct (improved quality of life). For interventions of this sort, many evaluation questions dealing with program process and outcome can arise because of the differentiated nature of the services and their potential effects. Furthermore, the evaluator may have difficulty developing measures that cleanly capture critical aspects of program implementation and outcome when they are complex or diffuse. Actual data collection, too, may be challenging if it must take place over extended time periods or involve many different variables and observations. All these factors have implications for the particulars of the evaluation plan and, especially, for the effort and resources that will be required to complete the plan.

The Resources Available for the Evaluation

It requires resources to conduct a program evaluation. Some number of person-hours devoted to the evaluation activities and materials, equipment, and facilities to support data collection, analysis, and reporting must be available whether drawn from the existing resources of the program or evaluation sponsor or separately funded. An important aspect of planning an evaluation, therefore, is to break down the tasks and timelines so that a detailed estimate can be made of the personnel, materials, and expenses associated with completion of the steps essential to the plan. The sum total of the resources required must then, of course, fit within what is available or some changes in either the plan or the resources must be made. Useful advice on the practicalities of resource planning, budgeting, and determining timelines can be found in Hedrick, Bickman, and Rog (1992), Card, Greeno, and Peterson (1992), and Fink (1995, chap. 9).

Although the available funding is, of course, one of the critical resource issues around which the evaluation must be planned, it is important to recognize that the dollar amount of that funding is not the only resource that will concern the evaluator. Evaluation is a specialized form of inquiry that takes place largely within the operating environment of the program being evaluated. This means, for instance, that pertinent technical expertise must be available if the evaluation is to be done well. In a large evaluation project, a number of proficient evaluators, data collectors, data managers, analysts, and assistants may be required to do a quality job. Even with generous funding, it will not always be easy to obtain the services of sufficient persons with the requisite expertise. This is why large, complex evaluation projects are often done through contracts with research firms with appropriate personnel on hand.

Another critical resource for an evaluation is support from program management, staff, and other closely related stakeholders. For instance, the degree of cooperation from program personnel on certain aspects of data collection such as opportunity to observe key program activities can have considerable influence on how much an evaluation can accomplish. Although these factors cannot be easily represented as dollar values, they are valuable resources for an evaluation. Barriers to access and lack of cooperation from the program, or worse, active resistance, are very expensive to the evaluation effort. It can take a considerable amount of time and effort to overcome these obstacles sufficiently to complete the evaluation, not to mention the associated stress for all concerned. In the most severe cases, such resistance may compromise the scope or validity of the evaluation or even make it impossible to complete.

An especially important interaction with the program involves access to and use of program records, documents, and other such internal data sources. It is a rare evaluation design that does not require at least some information from program records and many are based heavily on those records. Such records may be necessary to identify the number and characteristics of clients served, the type and amount of services they received, and the cost of providing those services. The scope, completeness, and quality of program records, as well as access to those records, thus are frequently major resource issues for an evaluation. Information that can be confidently obtained from program records need not be sought in a separate, and almost certainly more expensive, data collection administered by the evaluator. In some cases, program personnel may be provided to compile such data with, of course, appropriate monitoring from the evaluator to ensure their integrity. Indeed, evaluations conducted by evaluators internal to an organization, and therefore already on the payroll, that rely primarily on program records may be undertaken with very little direct funding. Evaluation sponsors and program managers are often quite unaware of what will be required of them in the course of an evaluation. They may, for instance, be surprised by an evaluator's request to have "hands on" access to records or to have program staff undertake activities in support of the evaluation.

Program records vary in how easy it is to use them. Records kept in writing are often difficult to use without considerable amounts of processing. In contrast, a management information system (MIS) consisting of records kept in machine-readable databases is usually easier to process. Increasingly, agency records are kept on computers, with duplicate copies obtainable for analysis. Of course, machine databases may contain missing information that reduces their utility, but in most cases, MIS data are valuable in evaluations.

The crucial point here is that the evaluator must view cooperation from program personnel, access to program materials, and the nature, quality, and availability of data from program records as major resource issues when planning an evaluation. The potential for misunderstanding and resistance can be lowered considerably if early discussions with evaluation sponsors, program personnel, and other relevant stakeholders spell out the resources and support needed for the various aspects of the evaluation. It follows that an important step in the planning process is to canvass such resources as thoroughly as possible so that realistic assumptions can be made about them in the evaluation design (Hatry, 1994). As early as possible during planning, therefore, the evaluator should meet with a range of program personnel and discuss their role in the evaluation and issues of access to staff, records, clients, and other pertinent information sources. It is also wise to determine what program records are kept, where and how they are kept, and what access will be permitted. It is advisable to actually inspect a sample of actual program records both to try out the procedures for accessing and working with them and to determine their completeness and quality. Because of the heavy workload demands often put on program personnel, record keeping is not always a high priority. A program may, therefore, have a very complete set of forms and procedures, but examining their files will reveal that they are used inconsistently or, perhaps, hardly at all.

Alongside adequate funding and cooperation from program personnel, experienced evaluators know that one of the most precious resources is time. The period of time allotted for completion of the evaluation and the flexibility of those time parameters are essential considerations in evaluation planning but are rarely determined by the evaluator's preferences. The decisions about the program that the evaluation is expected to inform follow the scheduling imperatives of the policy process. Evaluation results often have to be in the hands of certain decisionmakers by a certain date to have any chance of playing a role in a decision; after that they may be relatively useless. Such constraints often set very tight timelines for conducting an evaluation. Further complicating the situation is a pervasive underestimation among evaluation sponsors and decisionmakers of how long it takes to complete an evaluation. It is not uncommon for evaluation sponsors to request an evaluation that encompasses an imposing range of issues and requires considerable effort and then expect results in a matter of a few months.

The trade-offs here are quite significant. An evaluation can have breadth, depth, and rigor but will require proportionate funding and time. Or it can be cheap and quick but will, of necessity, either deal with a very narrow issue or be relatively superficial (or both). All but the most sophisticated evaluation sponsors usually want evaluations that have breadth, depth, and rigor *and* are cheap and quick. The result is all too often both overburdened evaluators working frantically against deadlines with inadequate resources and frustrated evaluation sponsors perturbed about shortfalls and delays in receiving the product they have paid for. An especially direct relationship exists between the time and technical expertise available for the evaluation and the methods and procedures that can be realistically planned. With few exceptions, the higher the scientific standard to be met by the evaluation findings, the greater the time, expertise, effort, and program cooperation that is required. Evaluations to which very limited resources are allocated of necessity must either focus on a circumscribed issue or rely on relatively informal procedures for obtaining pertinent information.

It takes careful planning to get the scope of work for the evaluation in balance with the funding, program cooperation, time, and other essential resources allocated to the project. Evaluation sponsors who insist on more work than available resources adequately support, or evaluators who overpromise what they can accomplish with those resources, are creating a situation that will likely result in shoddy work, unfulfilled promises, or both. It is generally better for an evaluation to answer a limited number of important questions well than a larger number poorly. Because evaluation sponsors and other stakeholders often do not have realistic notions of the amount of effort and expertise required to conduct quality evaluation, it is very easy for misunderstandings and conflicts to develop. The best way to prevent this is to negotiate very explicitly with the evaluation sponsor about the resources to be made available to the evaluation and the trade-offs associated with the inevitable constraints on resources.

THE NATURE OF THE EVALUATOR-STAKEHOLDER RELATIONSHIP

One of the matters requiring early attention in the planning of an evaluation is the nature of the relationship between the evaluator and the primary stakeholders. Every program is necessarily a social structure in which various individuals and groups engage in the roles and activities that constitute the program: Program managers administer, staff provide service, participants receive service, and so forth. In addition, every program is a nexus in a set of political and social relationships among those with an association or interest in the program, such as relevant policymakers, competing programs, and advocacy groups. These parties are typically involved in, affected by, or interested in the evaluation, and interaction with them must be anticipated as part of the evaluation. Who are the parties typically involved in, or affected by, evaluations? Listed below are some of the stakeholder groups that often either participate directly or become interested in the evaluation process and its results:

- *Policymakers and decisionmakers:* Persons responsible for deciding whether the program is to be started, continued, discontinued, expanded, restructured, or curtailed.
- *Program sponsors:* Organizations that initiate and fund the program. They may also overlap with policymakers and decisionmakers.
- *Evaluation sponsors:* Organizations that initiate and fund the evaluation (sometimes the evaluation sponsors and the program sponsors are the same).
- *Target participants:* Persons, households, or other units who receive the intervention or services being evaluated.
- *Program managers:* The personnel responsible for overseeing and administering the intervention program.
- *Program staff:* Personnel responsible for delivering the program services or in supporting roles.
- *Program competitors:* Organizations or groups who compete with the program for available resources. For instance, an educational program providing alternative schools will attract the attention of the public schools because they see the new schools as competitors.
- *Contextual stakeholders:* Organizations, groups, individuals, and other social units in the immediate environment of a program with interests in what the program is doing or what happens to it (e.g., other agencies or programs, public officials, or citizens' groups in the jurisdiction in which the program operates).
- *Evaluation and research community:* Evaluation professionals who read evaluations and pass judgment on their technical quality and credibility and academic and other researchers who work in areas related to a program.

Although other parties might be involved in the "politics of evaluation," this list represents the stakeholders who most often pay attention to an evaluation and with whom the evaluator may interact while the evaluation is being conducted or when its findings are reported. We emphasize *may:* these stakeholders are *potential* participants in one way or another in the evaluation process or *potential* audiences for the evaluation. In any given case, all these groups or only a few may be involved. But whatever the assortment of individuals and groups with significant interests in the evaluation, the evaluator must plan to interact with them in some fashion and be aware of their concerns. Consideration of the appropriate form of interaction for at least the major stakeholders thus should be part of evaluation planning (see Exhibit 2-H for one point of view on involving stakeholders).

EXHIBIT 2-H Stakeholder Involvement in Evaluation: Suggestions for Practice

Based on experience working with school district staff, one evaluator offers the following advice for bolstering evaluation use through stakeholder involvement:

- *Identify stakeholders:* At the outset, define the specific stakeholders who will be involved with emphasis on those closest to the program and who hold high stakes in it.
- *Involve stakeholders early:* Engage stakeholders in the evaluation process as soon as they have been identified because many critical decisions that affect the evaluation occur early in the process.
- *Involve stakeholders continuously:* The input of key stakeholders should be part of virtually all phases of the evaluation; if possible, schedule regular group meetings.
- *Involve stakeholders actively:* The essential element of stakeholder involvement is that it be active; stakeholders should be asked to address design issues, help draft survey questions, provide input into the final report, and deliberate about all important aspects of the project.
- *Establish a structure:* Develop and use a conceptual framework based in content familiar to stakeholders that can help keep dialogue focused. This framework should highlight key issues within the local setting as topics for discussion so that stakeholders can share concerns and ideas, identify information needs, and interpret evaluation results.

SOURCE: Adapted from Robert A. Reineke, "Stakeholder Involvement in Evaluation: Suggestions for Practice," *Evaluation Practice*, 1991, 12(1):39-44.

The process of considering the relationships with stakeholders necessarily starts with the evaluation sponsor. The sponsor is the agent who initiates the evaluation, usually provides the funding, and makes the decisions about how and when it will be done and who should do it. Various relationships with the evaluation sponsor are possible, and their particular form will largely depend on the sponsor's preferences and whatever negotiation takes place with the evaluator. A common situation is one in which the sponsor expects the evaluator to function as an independent professional practitioner who will receive guidance from the sponsor, especially at the beginning, but otherwise take full responsibility for planning, conducting, and reporting the evaluation. For instance, government agencies and other program funders often commission evaluations by publishing a request for proposals (RFP) or request for applications (RFA) to which evaluators respond with statements of their capability, proposed design, budget, and time line, as requested. The evaluation sponsor then selects an evaluator from among those responding and establishes a contractual arrangement for the agreed-on work.

Other situations are configured so that the evaluator works more collaboratively with the evaluation sponsors. For instance, the sponsors may want to be involved in an ongoing way with the planning, implementation, and analysis of results, either to react step by step as the evaluator develops the project or to actually participate with the evaluator in each step. Variations on this form of relationship are typical for internal evaluators who are part of the organization whose program is being evaluated. In such cases, the evaluator generally works closely with management in planning and conducting the evaluation, whether management of the evaluation unit, the program being evaluated, someone higher up in the organization, or some combination. Or an evaluator from outside the organization may be retained as an evaluation consultant to assist the evaluation sponsors in planning and conducting the evaluation but not take the primary role in doing that work.

In some instances, the evaluation sponsor will ask that the evaluator work collaboratively but stipulate that the collaboration be with a stakeholder group other than the evaluation sponsors themselves. For instance, private foundations that fund social programs often want an evaluation to be developed in close interaction with the local stakeholders of the program. An especially interesting variant of this approach is when the evaluation sponsor requires that the evaluation be a collaborative venture in which the recipients of program services take the primary role in planning, setting priorities, collecting information, and interpreting the results. Part of the philosophy of the W. K. Kellogg Foundation, for instance, is to avoid being prescriptive in its approach to evaluation of the programs it funds. In the words of the foundation's director of evaluation, "We believe that people in the communities and institutions we serve are in the best position to make decisions, to implement the programs that are the best suited for their circumstances at a given time, and to evaluate the lessons learned" (Millett, 1996, p. 68).

The evaluator's relationship to the evaluation sponsor or another stakeholder designated by the evaluation sponsor is so central to the evaluation context and planning process that a somewhat specialized, and not altogether systematic, vocabulary has arisen in the evaluation profession to describe various circumstances. Some of the major forms of evaluator-stakeholder relationships recognized in this vocabulary are as follows:

Independent evaluation. The evaluator takes the primary responsibility for developing the evaluation plan, conducting the evaluation, and disseminating the results. The evaluator may initiate and direct the evaluation quite autonomously, as when a social scientist undertakes an evaluation of an interesting program for purposes of knowledge generation under the researcher's own sponsorship or with research funding that leaves the particulars to the researcher's discretion. More often, the independent evaluator is commissioned by a sponsoring agency that stipulates the purposes and nature of the evaluation but leaves it to the evaluator to do the detailed planning and conduct the evaluation. In such cases, however, the evaluator generally confers with a range of stakeholders to give them some influence in shaping the evaluation.

Participatory or collaborative evaluation. This form of evaluation is organized as a team project with the evaluator and representatives of one or more stakeholder groups constituting the team (Greene, 1988; Mark and Shotland, 1985). The participating stakeholders are directly involved in planning, conducting, and analyzing the evaluation in collaboration with the evaluator whose function might range from

team leader or consultant to that of a resource person to be called on only as needed. One particularly well-known form of participatory evaluation is Patton's (1986, 1997) "utilization-focused evaluation." Patton's approach emphasizes close collaboration with those specific individuals who will use the evaluation findings to ensure that the evaluation is responsive to their needs and produces information that they can and will actually use.

Empowerment evaluation. Various proponents have articulated a concept of evaluator-stakeholder interaction that emphasizes the initiative, advocacy, and self-determination of the stakeholder group (Fetterman, Kaftarian, and Wandersman, 1996). In this form of evaluation, the evaluator-stakeholder relationship is participatory and collaborative, as described above. In addition, however, the evaluator's role includes consultation and facilitation directed toward development of the capabilities of the participating stakeholders to conduct evaluation on their own, to use it effectively for advocacy and change, and to experience some sense of control over a program that affects their lives. The evaluation process, therefore, is not only directed at producing informative and useful findings but also at enhancing the self-development and political influence of the participants. As these themes imply, empowerment evaluation most appropriately involves those stakeholders who otherwise have little power in the program context, often the program recipients or intended beneficiaries.

A significant contribution of the participatory and empowerment perspectives is to call into question what might otherwise be a routine presumption that an independent evaluation is appropriate. There are, of course, many situations in which the evaluation sponsor explicitly wants an independent evaluation that will pursue that sponsor's concerns under expert guidance from a professional evaluator. This process assures that the perspective of the evaluation sponsor will have priority and, given a competent evaluator, that the results will have a certain credibility stemming from the evaluator's expertise and a decision-making process that filters the influence of the self-interests of the stakeholders.

There are other situations, however, where the advantages of independent evaluation are not relevant or are outweighed by the benefits of a more participatory process. Direct participation by the evaluation sponsors or one or more other stakeholder groups can ensure that the evaluation results will address their concerns and be useful and usable for them. Moreover, it can create a sense of ownership in the evaluation that amplifies the significance of its findings and reduces its potential to engender resistance. And as the empowerment theorists point out, when stakeholder groups with little formal power are able to conduct and use an evaluation, it can alter the balance of power in a program context by enhancing their influence and sense of efficacy. It is thus appropriate for the evaluation sponsors and the evaluator to give explicit consideration to the question of how the evaluation responsibilities are to be assigned and the arrangements for organizing the evaluator-stakeholder interactions. Where such deliberation has not already taken place or an arrangement stipulated, it may be constructive for the evaluator to raise the issue and suggest that it be a matter of discussion during the earliest phase of evaluation planning. Exhibit 2-I illustrates what might be in store for stakeholders who opt to participate in a collaborative evaluation.

Whether the evaluation is planned and conducted by an independent evaluator or by a team of stakeholders has considerable effect on the nature of the decision making, the evaluator's role, and, most likely, the focus and character of the evaluation. The resulting project, nonetheless, should represent an application of recognizable evaluation concepts and methods to a particular program. We thus distinguish the process of working with stakeholders, whether as an independent evaluator, col-

EXHIBIT 2-1 Blueprint for a Participatory Evaluation Process

Two Canadian evaluators who were involved in a participatory evaluation of programs sponsored by a community economic development organization offered the following "blueprint" for the evaluation process they used:

Initiation phase:

- Presentation to the board of administrators for their approval.
- Identification of interested stakeholders.

Selecting the topics and questions to be addressed:

- Interested stakeholders meet several times in small groups, each centering on one of the services offered by the organization, to brainstorm ideas for questions; guidelines in Patton (1986) are followed to explain tasks such as focusing evaluation questions.
- Questions are rephrased clearly, regrouped for each program, and collated into one document.
- A general meeting of stakeholders is called to prioritize questions according to their potential utility and to plan how the evaluation results will be utilized once they are available.

Instrument design and data collection:

- Small groups are reconvened to decide on the final wording and format of questions retained at the general meeting.
- Data are collected by the program evaluator and other interested participants who are given appropriate training.

Data analysis and reporting:

- Data analysis proceeds in small groups, with the evaluator participating in all groups.
- Individual reports are drafted for each program; the evaluator is responsible for writing reports in consultation with stakeholders.

Strategic planning:

- A series of strategic planning meetings is convened to study the evaluation reports and decide on follow-up steps.
- The evaluation questions are revised for future use on an ongoing basis; program workers are expected to coordinate future evaluation efforts.

SOURCE: Adapted from Danielle Papineau and Margaret C. Kiely, "Participatory Evaluation in a Community Organization: Fostering Stakeholder Empowerment and Utilization," *Evaluation and Program Planning*, 1996, 19(1):79-93.

laborator, facilitator, or resource person, from the evaluation plan that results from that process. That plan may be developed and largely settled early in the process or may emerge piecemeal as the process develops and evolves, but the features of a good plan for the evaluation context and the program at issue can be considered separate from the process through which the planning and implementation is done. In the remainder of this chapter, therefore, we will discuss general planning issues and, when reference is made to the evaluator's role, assume that can mean either an independent evaluator or a collaborative team.

First, however, one other aspect of evaluator-stakeholder interaction must be addressed: the communication and dissemination of the evaluation findings. Even in the most participatory evaluation, there will be stakeholders who have not been directly involved who will want to know the results. For an independent evaluation, of course, the evaluation sponsor may be among those stakeholders. For evaluation to be useful, a necessary step is that its findings be communicated to those with interest in the program, especially to those with responsibility for making important decisions about the program. It is difficult to communicate evaluation findings in fine detail, and additionally, there is often inherent uncertainty about what information will be of most interest to stakeholders at the time the evaluation is completed. It is usually best, therefore, to discuss this issue with the major stakeholders and develop an organized communication and dissemination plan from the beginning.

As a general framework, the communication and dissemination plan should indicate what information from the evaluation is to be communicated to which stakeholders in what form and at which time. Different information may be relevant for different stakeholder groups, depending on their interest in the program and their decision-making roles. Moreover, that information might be communicated as soon as available or later when all phases of the evaluation are complete, and it might be communicated in writing or verbally and formally or informally. Typically, the evaluator should consider multiple communication events ranging from informal oral briefings to formal written reports. The objective of the communication and dissemination plan should be to report the findings of most interest to each stakeholder as soon as possible and proper and in forms that are easy to understand and use (see Exhibit 2-J).

Evaluation conducted for purposes of program improvement, for instance, might include regular briefings for the evaluation sponsor and program managers conducted as soon as each distinct data collection reaches a point where tentative analysis and interpretation are possible. These might be relatively informal briefings in which the evaluator presents a verbal summary with supporting handouts and encourages discussion. Other stakeholders might receive similar interim briefings in written or verbal format at less frequent intervals. At the conclusion of the evaluation, a written report might be prepared for the record and for the more peripheral stakeholders or might not if no use for it was apparent. Evaluation conducted for purposes of accountability, on the other hand, might properly involve a more formal communication and dissemination process throughout. Information might be released only after it was thoroughly verified and analyzed and primarily in carefully worded written form with verbal briefings used only as a supplement. This higher level of caution and formality would be justified if the stakes were high or the evaluation results were expected to

EXHIBIT 2-J Successful Communication With Stakeholders

Torres, Preskill, and Piontek (1996) surveyed and interviewed members of the American Evaluation Association about their experiences communicating with stakeholders and reporting evaluation findings. The respondents identified the following elements of effective communication:

- Ongoing, collaborative communication processes were the most successful. Periodic meetings and informal conversations can be used to maintain close contact throughout the evaluation, and interim memos and draft reports can be used to convey findings as they develop.
- It is important to use varied formats for communication. These might include short reports and summaries, verbal presentations, and opportunities for informal interaction.
- The content of the communication should be tailored to the audience and be easy for them to understand. Communication should use clear language, graphs and charts, and vivid, concrete illustrations. It should present contextual information about the program and the evaluation, cover both positive and negative findings, and be specific about recommendations.

SOURCE: Adapted from Rosalie T. Torres, Hallie S. Preskill, and Mary E. Piontek, *Evaluation Strategies for Communicating and Reporting: Enhancing Learning in Organizations* (Thousand Oaks, CA: Sage, 1996), pp. 4-6.

be especially controversial. It should be noted that this approach does not necessarily mean being highly secretive about the evaluation findings until the final report is delivered. The process of verifying and analyzing the information in such formal circumstances might quite appropriately include soliciting the reactions of potential critics and other important stakeholders to the initial summaries of the major findings and incorporating their feedback in the final report.

Whatever the schedule, form, and audiences for evaluation findings, it is wise to include some consideration of the communication media and materials in the planning. Evaluation findings, like the programs they describe, are rarely simple and easily understood. Communication will often be most effective if it makes good use of graphical and pictorial displays, uses engaging audiovisual materials, and personalizes portions of the story as much as possible through well-chosen anecdotes and case examples. To have such material available when needed, the evaluator must plan for its development during the course of the evaluation. For instance, it may be appropriate to make audio or video recordings of certain events or situations, systematically collect anecdotes and case examples from which to select representative instances, and make other such preparations for effective communication. Useful advice for planning effective communication and dissemination activities is found in Torres, Preskill, and Piontek (1996; also see Exhibit 2-J).

For many evaluations, it is also appropriate to allow stakeholders access to the database on which the evaluation was based, at the same

time safeguarding the privacy of sources from whom the data were obtained. Making a dataset public signals to stakeholders that the evaluators have nothing to hide and allows them to try alternative modes of analysis to verify that the evaluator's findings were properly drawn. We believe that making data publicly available should be a routine procedure in evaluations of large-scale or very important programs.

EVALUATION QUESTIONS AND EVALUATION METHODS

A program evaluation is essentially an information-gathering and -interpreting endeavor that attempts to answer a specified set of questions about a program's performance and effectiveness. An important step in designing an evaluation, therefore, is determining the questions the evaluation must answer. This is sometimes done in a very perfunctory manner, but we advocate that it be given studious and detailed attention. A carefully constructed set of evaluation questions gives structure to the evaluation, leads to appropriate and thoughtful planning, and serves as a basis for informative discussions about who is interested in the answers and how they will be used. Indeed, constructing such questions and planning how to answer them is the primary way in which an evaluation is tailored to the unique circumstances associated with each program that comes under scrutiny.

Generally, the evaluation sponsor puts forward some initial evaluation questions when proposing or commissioning an evaluation or, in the case of a competition to select an evaluator, as part of the RFP or RFA that goes out to prospective evaluators. Those initial declarations are the obvious starting point for defining the questions around which the evaluation will be designed but usually should not be taken as final for purposes of evaluation planning. Often the questions presented at this stage are too general or abstract to function well as a basis for evaluation planning. Or the questions, as worded, may be beyond the capability of the evaluator to answer within the operative constraints on time, resources, available information, and organizational or political arrangements.

Any initial description of what the evaluation sponsors have in mind, therefore, must usually be further explored, refined, and augmented to obtain a set of meaningful, appropriate evaluation questions around which the evaluation can actually be planned. In addition, it is usually useful for the evaluator to analyze the program independently and derive evaluation questions that may not otherwise arise so that they too can be considered during the planning process. However accomplished, a thorough effort must be made to generate a set of candidate evaluation questions that covers all the issues of potential relevance to the concerns of the evaluation sponsor, the decisionmakers who will use the findings, and other significant stakeholders. This approach allows the evaluation design to be responsive to the needs of decisionmakers and offers the potential to involve stakeholders as collaborators in the evaluation process. It is relatively easy to generate questions, however, so the initial set resulting from a diligent effort will likely be too large for the evaluation to answer them all. The evaluator, evaluation sponsor, and other key stakeholders must therefore impose priorities to select a workable number dealing with the most important issues.

Because the evaluation questions to be addressed are so pivotal to evaluation planning, Chapter 3 is devoted entirely to discussing the form they should take, how they are generated, and how they are winnowed, organized, and integrated to provide the structure for the evaluation design. For present purposes, we will assume that an appropriate set of evaluation questions has been identified and consider some of the broader implications of their character for tailoring and planning the evaluation.

In particular, the evaluation questions to be answered for a given program will, of necessity, be very specific to the idiosyncratic nature of that program. They will ask such things as "How many of the households that fall below the federal poverty line in the Fairview School District need afterschool care for school-aged children between 3 and 7 p.m. on weekdays?" "What proportion of the juveniles on probation have at least three contacts with their probation officer per month for the full six-month probationary period?" and "What nutritional benefits does the meals-on-wheels program at the senior citizens' center have for the housebound frail elderly in its catchment area?" Beyond the specifics, however, evaluation questions fall into recognizable types according to the program issues they address. Five such types are readily distinguished:

- Questions about the need for program services
- Questions about program conceptualization or design
- Questions about program operations and service delivery
- Questions about program outcomes
- Questions about program cost and efficiency

Evaluators have developed relatively distinct conceptual frameworks and associated methods to address each type of evaluation question. Evaluators use these schemes to organize their thinking about how to approach different program evaluation situations. For planning purposes, an evaluator will typically select the general evaluation approach that corresponds to the types of questions to be answered in an evaluation, then tailor the particulars to the specifics of the questions and the program situation. To complete our discussion of tailoring evaluations, therefore, we must introduce the common evaluation approaches or schemes and review the circumstances in which they are most applicable.

The common conceptual and methodological frameworks in evaluation correspond to the types of frequent evaluation questions, as follows:

- *Needs assessment:* answers questions about the social conditions a program is intended to address and the need for the program.
- *Assessment of program theory:* answers questions about program conceptualization and design.
- *Assessment of program process (or process evaluation):* answers questions about program operations, implementation, and service delivery.
- *Impact assessment (impact evaluation or outcome evaluation):* answers questions about program outcomes and impact.
- *Efficiency assessment*: answers questions about program cost and cost-effectiveness.

These forms of evaluation are discussed in detail in later chapters of this volume (Chapters 4-11). Here we will only provide some guidance regarding the circumstances for which each is most appropriate.

Needs Assessment

The primary rationale for initiating or maintaining a social program is a presenting or incipient social problem—by which we mean socially recognized deficiencies in the social conditions—that legitimate social agents endeavor to remedy. The impetus for a new program to increase literacy, for example, is likely to be recognition that a significant number of persons in a given population are deficient in reading skills. Similarly, an ongoing program may be justified by the persistence of a social problem: Driver education in high schools receives public support because of the continuing high rates of automobile accidents among adolescent drivers.

If there is no significant problem or no perceived need for intervention, there is generally no basis for affirming the value of a program that purports to address this nonproblem. One important form of evaluation, therefore, assesses the nature, magnitude, and distribution of a social problem; the extent to which there is a need for intervention to address it; and the implications of these circumstances for the conceptualization and design of the intervention. These diagnostic activities are often referred to as *needs assessment* in the evaluation field but overlap what is called social epidemiology and social indicators research in other fields (McKillip, 1987; Reviere et al., 1996; Soriano, 1995; Witkin and Altschuld, 1995). Needs assessment is often used as a first step in designing and planning a new program or restructuring an established program to provide information about what services are needed and how they might best be delivered to those who need them. Needs assessment is also often important for established, stable programs to examine whether they are responsive to the actual needs of the target participants and to provide guidance for improvement.

Needs assessments may take the form of finding out the needs of a potential target population as they perceive them. For example, homeless persons may be queried about the kinds of services for which they feel the greatest need (e.g., see Exhibit 2-K). Alternatively, the objective of needs assessment may be to describe conditions in such a way that the services needed to alleviate them can be inferred.

Needs assessments may be conducted through surveys of knowledgeable informants, such as personnel of service agencies or potential service recipients, that focus on perceived problems and needs, services desired, and shortcomings of existing services. They may also analyze demographic and social indicator data from such sources as the U.S. Census or data from local agencies that describe the availability of services and patterns of current use. The resulting descriptions of social problems, service utilization, and perceived needs must then be assessed against some set of social norms or some view of desired conditions held by social agents or those experiencing the problems to evaluate their magnitude, seriousness, and actionable implications. Chapter 4 of this book discusses the various aspects of needs assessment in detail.

Assessment of Program Theory

Given a recognized problem and need for intervention, it does not follow that any program, willy-nilly, will be appropriate for the job. The conceptualization and design of the program must reflect valid assumptions about the nature of the target problem and represent a well-founded and feasible approach to resolving it. Put another way, every social program is based on some plan or blueprint that represents the way it is "supposed to work" according to those who understand its history, purposes, and activities the best. This plan is rarely written out in complete detail, and may not be written out at all, but exists nonetheless as a shared conceptualization among the principal stakeholders. Because this program plan consists essentially of a set of assumptions and expectations about how the program should conduct its business and attain its goals, we will refer to it as the program theory (discussed more fully in the next chapter). If this theory is faulty, the intervention will fail no matter how elegantly it is conceived or how well it is implemented (Chen, 1990; Weiss, 1972).

Assessment of the program theory involves, first, representing it in explicit and detailed written or graphical form. Then, various approaches can be used to examine how reasonable, feasible, ethical, and otherwise appropriate it is. Assessment of program theory is most essential to programs during their early stages, for example, when they are new or in pilot testing or even earlier when they are in the planning stage. However, it is also applicable to established programs, especially when questions arise about how well matched their services are to the social needs they are attempting to meet. The sponsors of this form of evaluation are generally those attempting to launch a new program, such as the funding agency or administrators, or those seeking assurance that the conceptualization and design of a program are appropriate for its purposes. Exhibit 2-L, for example, describes an examination of the conceptual foundation for family preservation programs, which indicated that they had little prospect for success.

Evaluation of program theory rests on the presumption that the need for the program and

EXHIBIT 2-K Needs for Help Among Homeless Men and Women

A representative sample of 1,260 homeless men and women were interviewed in New York City's municipal shelters for single adults to determine their perception of their needs. The interview covered 20 items, each indicating need for help in a particular area. Most respondents identified multiple needs, averaging 6.3. The need for help in finding a place to live and having a steady income were the most commonly cited needs overall, closely followed by the need for help in finding a job and improving job skills. Compared to women, men more often reported needs for help with drinking problems, drug problems, learning how to handle money, getting veterans benefits, problems with the police, getting along better with other people, and finding a place to live. Women more frequently reported needs for help with health and medical problems and learning self-protection skills. The evaluators pointed out that for programs to be truly responsive to these multiple needs, they must have the capacity to deliver or broker access to a comprehensive range of services.

SOURCE: Adapted by permission from Daniel B. Herman, Elmer L. Struening, and Susan M. Barrow, "Self-Reported Needs for Help Among Homeless Men and Women," *Evaluation and Program Planning*, 1994, 17(3):249-256. Copyright © 1998, John Wiley & Sons, Inc.

EXHIBIT 2-L A Flaw in the Design of Family Preservation Programs

As part of an evaluability assessment (see Chapter 5), evaluators working under contract to the U.S. Department of Health and Human Services reviewed the design of family preservation programs (FPPs). FPPs are time-limited, intensive home-based services to families in crisis that are intended to prevent the placement of children in foster care. The evaluators held discussions with the staff of federal and national private sector agencies about the definition of FPPs, reviewed available literature, obtained descriptions of state and local programs, and made site visits to four programs. From this information they developed "models" of how the programs were supposed to operate and then obtained the views of policymakers, program managers, and operating-level staff on four key dimensions: (a) program goals, (b) aspects of the child welfare system that affect the programs, (c) the target population, and (d) the characteristics that distinguish FPPs from other home-based services. Based on their own analysis and discussions with an expert advisory committee, the evaluators concluded that as currently designed, family preservation programs could not achieve the policymakers' primary goal of preventing placement in foster care. The major flaw found in the program design was the practical difficulty of identifying children at "imminent risk" of placement; this meant that programs could not consistently target families with children truly at risk of placement.

SOURCE: Adapted from Joseph S. Wholey, "Assessing the Feasibility and Likely Usefulness of Evaluation," in *Handbook of Practical Program Evaluation*, ed. J. S. Wholey, H. P. Hatry, and K. E. Newcomer (San Francisco: Jossey-Bass, 1994), pp. 29-31. Wholey's account, in turn, is based on Kaye and Bell (1993).

adequate diagnosis of the problem the program is to address has already been established or can confidently be assumed. Also, because analysis of program design requires a close collaboration among the evaluator, program designers and managers, and other key stakeholders, it is most readily accomplished when all parties are willing to be fully engaged in the process and can establish constructive working relationships. Somewhat paradoxically, however, many of the techniques associated with the evaluation of program theory are applicable in situations of stakeholder conflict, for example, disagreement about program goals and objectives, appropriate priorities, and the nature of program activities. Assessment of program theory involves making that theory explicit so there will be little uncertainty about significant aspects of the program concept and intended implementation. Stakeholder disagreement over such matters raises uncertainty and may also create a politically fluid situation in which fundamental program changes are possible. These conditions are sufficiently similar to those of a new program still under design to permit methods for assessing program theory to be potentially helpful. In this case, however, the context of application may be distinguished more by hostility among stakeholder groups than cooperative working arrangements, posing special challenges (and hazards) to the evaluator.

Assessment of Program Process

Given a plausible theory about how to favorably intervene to ameliorate accurately diagnosed social problems, a program must still be implemented well to have a reasonable prospect of actually affecting the target problem. Many programs are not implemented and executed according to their intended design. A program may simply be poorly managed or be compromised by political interference. Sometimes personnel are not available or facilities are in disrepair; sometimes project staff cannot carry out the program due to lack of motivation or expertise. Often the program design is not well structured, leaving much room for interpretation, or the original program plan may not be transmitted well to staff so that program activities drift over time. Possibly, the intended program participants do not exist in the numbers required, cannot be identified precisely, or are not cooperative. For example, some programs to serve children with congenital heart conditions found it so difficult and costly to locate potential clients that insufficient funds remained for providing the intended treatments.

A central and widely used form of evaluation, therefore, assesses the fidelity and effectiveness of program implementation. Implementation assessment evaluates program process, the activities and operations of the program. For this reason, it is commonly called *process evaluation* or, when it is an ongoing function, *program monitoring*. Process evaluation addresses questions related to how well the program is functioning. It might include assessment of how congruous the services are with the goals of the program, whether services are delivered as intended to appropriate recipients, how well service delivery is organized, the effectiveness of program management, the use of program resources, and other such matters (Exhibit 2-M provides an example).

In a typical process evaluation, criteria are developed for the program functions viewed as critical in two ways. These may be configured in the form of a "blueprint" of the intended program design that depicts the functions, activities, and service transactions that the program is supposed to accomplish in, perhaps, flowchart form (a version of program theory; see Chapter 3). Or the criteria may be stated in the form of specific administrative or service objectives, for example, to enroll 20 new clients each week, to provide a minimum of ten sessions of service to each client within the first three months, to have 90% of the clients receive the full term of service, and to make educational presentations to one community group each week. Such criteria can be developed in various ways: They may simply be stipulated by program administrators, they may be mandated by program funders, they may be derived from studies and reports of other programs or follow the specification of some model program, they may result from a process of reflection and goal setting by program personnel or other stakeholders, or they may be drawn from accepted principles of organizational effectiveness or professional practice.

With the critical program functions and corresponding performance criteria identified, the other component of process evaluation is the definition and operationalization of performance measures that describe program accomplishments in relation to the respective criteria. Thus, data collection procedures might be put in place to determine the number of new patients enrolled each week, the percentage who complete a full term of service, the number of presentations to community groups, and the like. Program performance can then be assessed by comparing what is found on these

EXHIBIT 2-M Failure on the Front Lines: Implementing Welfare Reform

Work Pays is a state-level welfare reform demonstration program in California designed to establish incentives to work and disincentives for staying on the AFDC welfare program. The program administrators recognized that to realize the policymakers' intent, the workers in local welfare offices would have to inform their clients about the new policy and present this information in a positive, individualized way that would reinforce clients' understanding of their obligations and choices about work and welfare. An implementation assessment was therefore conducted in which researchers interviewed welfare workers about the Work Pays program and observed a number of meetings with clients. This information revealed that the type of transaction expected between welfare workers and their clients under the new policy was exceedingly rare. In more than 80% of their interviews with clients, workers did not provide and interpret information about the new policy. Most workers continued their routine patterns of collecting and verifying eligibility information and providing scripted recitations of welfare rules. However, the evaluators also found that the workers had been given only minimal information about the Work Pays program and no additional time or resources for educating their large caseloads about the changes. These findings demonstrated that welfare reform was not fully implemented at the street level in California and revealed some of the reasons why it was not.

SOURCE: Adapted from Marcia K. Meyers, Bonnie Glaser, and Karin MacDonald, "On the Front Lines of Welfare Delivery: Are Workers Implementing Policy Reforms?" *Journal of Policy Analysis and Management*, 1998, 17(1):1-22.

measures with the criterion for desired performance on that program function.

Although process evaluation is often done as a one-shot evaluation study, say, for one cohort of program clients, it should be apparent that similar procedures can be used routinely as a management tool. When set up to provide periodic performance data for key program functions on a continuous basis, this form of assessment is generally known as program monitoring or performance monitoring. There are many good reasons for programs to institute monitoring schemes. For instance, monitoring provides a way for program managers to ensure that the day-to-day operations of a program are conducted appropriately and efficiently and thus helps them properly administer the program. Managers who develop reputations for wasting funds, using staff resources inappropriately, and being inefficient in other regards frequently jeopardize not only their own positions but the futures of their programs.

Also, program monitoring information systems give program administrators a powerful tool for documenting for program sponsors and stakeholders the operational effectiveness of the organization, justifying the ways staff are deployed, requesting further support, and defending the program's performance compared with its competitors. Routinely collecting and reporting program performance information, therefore, makes the program accountable and provides evidence to funders and sponsors that

what was paid for and deemed desirable was actually accomplished.

Process evaluation of some variant is the assessment approach most frequently applied to social programs. It is used both as a freestanding evaluation and in conjunction with impact evaluation as part of a more comprehensive evaluation. As a freestanding evaluation, it yields quality assurance information. That is, it assesses the extent to which a program is implemented as intended and operating up to the standards established for it. When the program model employed is one of established effectiveness, a demonstration that the model is well implemented is presumptive evidence that the expected outcomes are produced as well. When the program is new, a process evaluation provides invaluable feedback to administrators and other stakeholders about the progress that has been made operationalizing the program theory. From a management perspective, process evaluation provides the feedback that allows a program to be managed for high performance (Wholey and Hatry, 1992), and the associated data collection and reporting of key indicators may be institutionalized in the form of an MIS to provide routine, ongoing performance feedback.

In its other common application, process evaluation is an indispensable adjunct to impact evaluation. The information about program outcomes that impact evaluation provides is incomplete and ambiguous without knowledge of the program activities and services that produced those outcomes. When no impact is found, process evaluation has significant diagnostic value by indicating whether this result occurred because of *implementation failure*, that is, the intended services were not provided hence the expected benefits could not have occurred, or *theory failure*, that is, the program was implemented as intended but failed to produce the expected effects. On the other hand, when program effects are found, process evaluation helps confirm that they resulted from program activities, rather than spurious sources, and identify those aspects of service most instrumental to producing the effects so that program managers know where to concentrate their efforts.

As a general evaluation approach, process evaluation is widely applicable to social programs. For stable programs that have established operating procedures, personnel, and facilities, process evaluation may provide summative information relevant to both program accountability and knowledge generation. For new programs or those in flux, process evaluation may constitute a formative evaluation that provides useful feedback for program managers attempting to improve program operations. In either case, however, process evaluation requires a well-defined, consensual program theory that stipulates the "program as intended." If program managers and other pertinent stakeholders cannot delineate the program model that is supposed to be implemented, or cannot agree on the intended clientele, services, and procedures, then there is no defined process for the evaluator to observe and assess. In this case, the evaluator may adopt the techniques for assessing program theory and work with program managers, evaluation sponsors, and other stakeholders to better define the program conceptualization.

Process evaluation and its variants are described in greater depth in Chapter 6 of this volume. It is important to note here, however, that although it is widely used as a freestanding evaluation approach to give evaluation sponsors, program managers, and other stakeholders an assessment of how well the program is implemented, it is not a substitute for impact evaluation. Process evaluation does not address

the question of whether a program produces the intended outcomes and benefits for its recipients. Even in cases where the program model is known to be effective in other applications and process evaluation demonstrates that it is well implemented, there remains a possibility that circumstances are sufficiently different in the program at issue to keep it from being effective there.

Impact Assessment

An impact assessment, sometimes called an impact evaluation or an outcome evaluation, gauges the extent to which a program produces the intended improvements in the social conditions it addresses. The evaluation questions around which impact assessment is organized relate to such matters as whether the desired program outcomes were attained, whether the program was effective in producing change in the social conditions targeted, and whether program impact included unintended side effects. These questions assume a set of operationally defined objectives and criteria of success. The objectives may be social-behavioral ones, such as lowering functional illiteracy or nutritional deficiencies among children; they may be community related, such as reducing the frequency of certain crimes; or they may be physical, such as decreasing water pollution or the amount of litter on city streets. Impact assessments are essential when there is an interest in determining if a program is effective in its efforts to ameliorate a target problem, comparing the effectiveness of different programs, or in testing the utility of new efforts to address a particular community problem.

Impact assessment has the basic aim of producing an estimate of the net effects of an intervention—that is, an estimate of the impact of the intervention uncontaminated by the influence of the other processes and events that also affect the conditions the program attempts to change. To conduct an impact assessment, the evaluator needs a plan for collecting data that will permit a persuasive demonstration that observed changes are a function of the intervention and cannot readily be accounted for in other ways. This requires a careful specification of the outcome variables on which program effects may occur, development of measures for those variables, and a research design that not only establishes the status of program recipients on those measures but also estimates what their status would be had they not received the intervention. Much of the complexity of impact assessment is associated with obtaining a valid estimate of the latter status, known as the *counterfactual* because it describes a condition contrary to what actually happened to program recipients. Specific impact assessment designs vary considerably. Sometimes it is possible to use classic experimental designs in which control and experimental groups are constructed by random assignment and receive different interventions. For practical reasons, however, it is often necessary to employ statistical approaches to isolating program effects rather than true experiments. Thus, nonrandomized quasi-experiments and other nonexperimental methods are commonly employed in impact assessments. With proper safeguards and appropriate qualifications, such nonexperimental designs may provide reasonable estimates of effects (Exhibit 2-N describes such a situation).

As mentioned above, impact assessment is often combined with process evaluation so that a linkage can be made between program implementation and the program outcomes observed. When an impact assessment is conducted without any semblance of a process evaluation, it is often referred to as *black box evaluation* because the evaluator may learn what the program effects are but does not know anything about the program processes that produced those effects—the program is a black box into which the evaluation cannot (or does not) see.

Determining when an impact assessment is appropriate, and what evaluation design to use when it is, present considerable challenge to the evaluator. On the one hand, evaluation sponsors often believe that they need an impact evaluation and, indeed, it is the only way to determine if the program is having the intended effects. On the other hand, impact assessment is characteristically very demanding of expertise, time, and resources and is often very difficult to set up properly within the constraints of routine program operation. For these reasons, a full-blown impact assessment is not to be undertaken lightly. It is generally appropriate only when there is an important purpose to be served by learning about program effects. This may be because the program concept is innovative and promising or in circumstances where identifiable decisionmakers have a high likelihood of actually using evidence about program impact as a basis for

EXHIBIT 2-N No Impact on Garbage

Taiwan is a high-density island country with a garbage problem. Garbage accumulation has increased exponentially in recent years, 26 rivers are polluted by garbage, and the number of landfill sites is increasingly limited. Consequently, in 1993 a demonstration garbage reduction program (GRD) was launched in Nei-fu, a suburb of Taipei, and evaluated for its impact on the amount of waste produced. Garbage is collected daily in Taiwan and the plan of the GRD was to disrupt this routine by suspending Tuesday collections. The theory was that requiring residents to store garbage one day a week in their homes, which are ill equipped for that function, would create sufficient inconvenience and unpleasantness to raise awareness of the garbage problem. As a result, it was expected that residents would make efforts to reduce the volume of garbage they produced. A process evaluation established that the program was implemented as planned.

The impact assessment was conducted by obtaining records of the daily volume of garbage for Nei-fu and the similar, adjacent suburb of Nan-kan for a period beginning four months prior to the program onset and continuing four months after. Analysis showed no reduction in the volume of garbage collected in Nei-fu during the program period relative to the preprogram volume or that in the comparison community. The evidence indicated that residents simply saved their customary volume of Tuesday garbage and disposed of it on Wednesday, with no carryover effects on the volume for the remainder of each week. Interviews with residents revealed that the program theory was wrong—they did not report the inconvenience or unpleasantness expected to be associated with storing garbage in their homes.

SOURCE: Adapted from Huey-Tsyh Chen, Juju C. S. Wang, and Lung-Ho Lin, "Evaluating the Process and Outcome of a Garbage Reduction Program in Taiwan," *Evaluation Review*, 1997, 21(1):27-42.

significant action. Such conditions may occur, for instance, with a demonstration project set up to test a program concept that, if effective, will be disseminated to other sites. Or impact assessment may be called for in high-accountability contexts where, for instance, continued funding for a program depends on its ability to demonstrate impact.

If the need for outcome information is sufficient to justify the expense and effort of an impact assessment, there is still a question of whether the program circumstances are suitable for such an evaluation. For instance, it makes little sense to establish the impact of a program that is not well structured or cannot be adequately described. Even if positive effects are found under such circumstances, ambiguity remains about what program features caused them or how they would be replicated elsewhere. Impact assessment, therefore, is most appropriate for mature, stable programs with a well-defined program model and a clear use for the results that justifies the effort required to conduct this form of evaluation. The most useful impact assessment results are for well-structured, well-documented programs believed to have important effects that must be established to support important decisions about the program or the program model.

Chapters 7-10 of this volume discuss impact assessment and the various ways in which it can be designed and conducted. Although all such designs are demanding, some are easier to implement than others in typical program circumstances. For impact assessment, therefore, much of the tailoring that must be done for application to a particular program is determining just which design to use, how to configure it, and what problems are associated with any compromises.

Efficiency Assessment

Unless programs have a demonstrable impact, it is hard to defend implementing or maintaining them—hence the need for impact assessments. But knowledge of impact alone is often insufficient; program results must also be judged against their costs. This is especially true in the present political climate as the resources for supporting social programs are curtailed and competition among programs for funds is intensified. Some programs may not be supportable because of high costs relative to their impact. In the face of budget problems, for example, some universities have terminated their student counseling programs because costs are high and benefits slight. Other initiatives may be expanded, retained, or terminated on the basis of their comparative costs. For instance, findings about the impact of institutional versus community care for adolescent offenders suggest that community programs are preferable because of their markedly lower costs.

Taking account of the relationship between costs and effectiveness requires efficiency assessments (Exhibit 2-O provides an example). This form of evaluation builds on process and impact assessment. If it is established that a program is well implemented and produces the desired outcomes, questions related to efficiency become relevant. Typical questions of this sort include "Is a program producing sufficient benefits in relation to the costs incurred?" and "Does it produce a particular benefit at a lower cost per unit of outcome than other interventions or delivery systems designed to achieve the same goal?" The techniques for answering these types of questions are found in two closely related approaches: cost-benefit and cost-effectiveness analyses. Cost-benefit analysis studies the relationship between program costs and outcomes, with both costs and outcomes expressed in monetary terms. Cost-effectiveness analysis examines the relationship between program costs and outcomes in terms of the costs per unit of outcome achieved. Efficiency assessment can be tricky and arguable because it requires making assumptions about the dollar value of program-related activities and, sometimes, imputing monetary value to program benefits. Nevertheless, such estimates are often essential for decisions about the allocation of resources to programs, identifying the program models that produce the strongest results with a given amount of funding, and determining the degree of political support stakeholders provide to a program.

Like impact assessment, efficiency assessment is most appropriate for mature, stable programs with a well-structured and well-documented program model. In addition, as mentioned above, efficiency analysis builds on both process and impact assessment, so it is important that the nature and magnitude of program effects be determined in advance of, or parallel to, efficiency assessment. Given the specialized expertise required to conduct efficiency assess-

EXHIBIT 2-O The Cost-Effectiveness of Community Treatment for Persons With Mental Disabilities

If provided with supportive services, persons with mental disabilities can often be maintained in community settings rather than state mental hospitals. But is such community treatment more costly than residential hospital care? A team of researchers in Ohio compared the costs of a community program that provides housing subsidies and case management for state-certified severely mentally disabled clients with the costs of residential patients at the regional psychiatric hospital. Program clients were interviewed monthly for more than two years to determine their consumption of mental health services, medical and dental services, housing services, and other personal consumption. Information on the cost of those services was obtained from the respective service providers and combined with the direct cost of the community program itself. Costs for wards where patients resided 90 or more days were gathered from the Ohio Department of Mental Health budget data and subdivided into categories that corresponded as closely as possible to those tabulated for the community program participants. Mental health care comprised the largest component of service cost for both program and hospital clients. Overall, however, the total cost for all services was estimated at $1,730 per month for the most intensive version of community program services and about $6,250 per month for residential hospital care. Community care, therefore, was much less costly than hospital care, not more costly.

SOURCE: Adapted from George C. Galster, Timothy F. Champney, and Yolonda Williams, "Costs of Caring for Persons With Long-Term Mental Illness in Alternative Residential Settings," *Evaluation and Program Planning*, 1994, 17(3):239-348.

ments, it is also apparent that it should be undertaken only when there is a clear need and an identified user for the information. With the high level of concern about program costs in many contexts, however, this may not be an unusual circumstance.

The procedures used for efficiency assessment are not as demanding of resources and program cooperation as are those of impact evaluation, but they are quite technical and require a high level of expertise. Chapter 11 discusses efficiency assessment methods in more detail.

STITCHING IT ALL TOGETHER

This chapter has reviewed the major considerations involved in tailoring an evaluation so that there is an appropriate fit between the evaluation plan and the circumstances of the program to be evaluated. However, it supplies few directly prescriptive injunctions that tell the prospective evaluator just what approach to take, what options to select, and how to go about putting the elements of an evaluation plan together. Designing an evaluation is not a mechanical activity that can be accomplished by applying a set of rules. A good evaluation plan is heavily contextualized by the political situation, the nature of the program, the interests of the stakeholders, and many other such specific features of the program landscape. It thus requires the evaluator to make many judgment calls based on a careful reconnaissance of that landscape. Moreover, experienced evaluators will disagree among themselves on what "call" should be made for many aspects of an evaluation plan for a given program situation. What we have tried to do in this chapter, therefore, is identify the major issues the evaluator must engage, the main alternatives that might be contemplated, and the primary considerations involved in making choices that tailor an evaluation plan to the program circumstances (see Exhibit 2-P for a similar perspective from "down under").

It should be evident that there is a certain logic in the relationships among the various program issues and the evaluation approaches surveyed in this chapter. In that logic, the social conditions, target population, and associated service needs a program is intended to address must first be identified and assessed. With that assessment in hand, the evaluator can ask if the basic conceptualization of the program (the program as intended) represents a reasonable means for addressing those needs. Deficiencies in this domain must be remedied by reconceptualizing the program. If the program theory is reasonable, the next question in this evaluation logic is whether the program is actually implemented and operationalized as intended, that is, as stipulated by the theory. Shortcomings in implementation generally must be solved by managerial initiatives. If the program is implemented as intended, then it is meaningful to ask if it has the intended effects—this is impact evaluation. A program that does not have the intended impact is either not implemented as intended or the program theory on which the program's operational plan is based must be faulty. If the intended effects are produced, it is then especially germane for the evaluation to inquire into the costs associated with attaining those effects and, especially, the efficiency with which they were attained. Programs with costs judged to be disproportionately large relative to their benefits, or to alternate ways of attaining those benefits, may need to find more cost-effective approaches or be replaced with more cost-effective alternatives.

EXHIBIT 2-P An Australian Team's Ten-Step Approach to Program Evaluation

A systematic approach to evaluation planning developed by the Research and Evaluation Support Services Unit of the New South Wales Department of Education is organized around the following ten questions:

1. *What is the program to be evaluated?* A common problem for the evaluator is defining what constitutes the "program" for evaluation. Some educational programs consist of a set of initiatives that may not be closely integrated. In general, a program is defined as "the set of operations, actions, or activities designed to produce certain desired effects or outcomes."
2. *Why is the program being evaluated?* Evaluation may focus on information needs related to what should be done, what can be done, what is being done, or what has been done.
3. *How are people to be prepared for the evaluation?* Thought should be given to who is likely to feel threatened by the evaluation, whose acceptance of the evaluation is essential, and what might be done to provide reassurance and gain acceptance.
4. *What are the main issues/questions with which the evaluation is to deal?* In this step the evaluator expands on the decisions made at Step 2 and develops a list of major issues or questions that define the evaluation's focus.
5. *Who will do what?* The responsibilities of participants in the evaluation should be understood and agreed on before the project is started.
6. *What are the resources for the evaluation?* In addition to workers, the evaluation may need other resources including money, material, facilities, transportation, and the like.
7. *What data need to be collected?* Data collection activities should be planned in relation to the major issues/questions identified in Step 4. They should be specific with regard to from whom the data are to be collected, how they will be collected, and what information must be covered.
8. *How will the data be analyzed?* Before data are collected, consideration should be given to what ultimately will be done with them. The answers will influence decisions both about the information to be collected and the form in which it will be collected.
9. *What will be the reporting procedure?* In this step, decisions should be made about to whom the reports will be provided and the appropriate ways to do this (e.g., written report, group discussion, newspapers). A further consideration is who should be asked to respond to the evaluation report. This step is important both for maintaining support for the evaluation and for ensuring that the report is communicating effectively to the groups who need to know the results.
10. *How will the report be implemented?* Attention should be given to identifying who is to be responsible for making recommendations on the basis of the evaluation results and who is responsible for implementing the recommendations. It should not be automatically assumed that the evaluation team will be the ones formulating recommendations; the evaluation may inform relevant decisionmakers who assume that task.

SOURCE: Adapted from Linda J. Lee and John F. Sampson, "A Practical Approach to Program Evaluation," *Evaluation and Program Planning*, 1990, 13(2):157-164.

SUMMARY

- Every evaluation must be tailored to the circumstances of the program being evaluated so that the evaluation design will be capable of yielding credible and useful answers to the specific questions at issue while still being sufficiently practical to actually implement within the resources available.
- One important influence on an evaluation plan is the purpose the evaluation is intended to serve. Evaluations generally are initiated to either provide feedback for program improvement to program managers and sponsors, establish accountability to decisionmakers with responsibility to ensure that the program is effective, or contribute to knowledge about some form of social intervention. The overall purpose of the evaluation necessarily shapes its focus, scope, and construction.
- Another important factor in planning an evaluation is the nature of the program structure and circumstances. The evaluation design must be responsive to how new or open to change the program is, the degree of consensus or conflict among stakeholders about the nature and mission of the program, the values and concepts inherent in the program rationale and design, and the way in which the program is organized and administered.
- Evaluation planning must also accommodate to the inevitable limitations on the resources available for the evaluation effort. The critical resources include not only funding but also the amount of time allowed for completion of the work, pertinent technical expertise, program and stakeholder cooperation, and access to important records and program material. A balance must generally be found between what is most desirable from an evaluation standpoint and what is feasible in terms of available resources.
- The evaluation design itself can be structured around three issues: (a) the questions the evaluation is to answer, (b) the methods and procedures to be used to answer those questions, and (c) the nature of the evaluator-stakeholder interactions during the course of the evaluation.
- Deciding on the appropriate relationship between the evaluator and the evaluation sponsor, as well as other major stakeholders, is an often neglected, but critical aspect of an evaluation plan. An independent evaluation, in which the evaluator takes primary responsibility for designing and conducting the evaluation, is often expected. In some circumstances, however, a more participatory or collaborative interaction with stakeholders may be desirable, with the evaluation conducted as a team project. In the latter case, the evaluation may be designed to help develop the capabilities of the participating stakeholders in ways that enhance their skills or political influence.

- The evaluation questions that are identified during planning, and the methods for answering them, generally fall into one or more recognizable categories having to do with (a) the need for services, (b) program conceptualization and design, (c) program implementation, (d) program outcomes, or (e) program efficiency. Evaluators have developed relatively distinct conceptual and methodological approaches for each of these different categories of issues that are referred to by such terms as *needs assessment*, *process evaluation*, and *impact assessment*. In practice, much of evaluation planning consists of identifying the evaluation approach corresponding to the type of questions to be answered in an evaluation, then tailoring the specifics to the program situation.

KEY CONCEPTS FOR CHAPTER 3

Evaluation questions	A set of questions developed by the evaluator, evaluation sponsor, and other stakeholders; the questions define the issues the evaluation will investigate and are stated in terms such that they can be answered in a way useful to stakeholders using methods available to the evaluator.
Program goal	A statement, usually general and abstract, of a desired state toward which a program is directed. Compare with **program objectives.**
Program objectives	Specific, operationalized statements detailing the desired accomplishments of a program.
Performance criterion	The standard against which a dimension of program performance is compared so that it can be evaluated.
Utilization of evaluation	The use of the concepts and findings of an evaluation by decisionmakers and other stakeholders whether at the day-to-day management level or at broader funding or policy levels.
Program theory	The set of assumptions about the manner in which the program relates to the social benefits it is expected to produce and the strategy and tactics the program has adopted to achieve its goals and objectives.
Impact theory	The beliefs, assumptions, and expectations inherent in a program about the nature of the change brought about by program action and how it results in the intended improvement in social conditions. Program impact theory is causal theory: It describes a cause-and-effect sequence in which certain program activities are the instigating causes and certain social benefits are the effects they eventually produce.
Service utilization plan	The assumptions and expectations about how the target population will make initial contact with the program and be engaged with it through the completion of the intended services. In its simplest form, a service utilization plan describes the sequence of events through which the intended clients are expected to interact with the intended services.
Organizational plan	The assumptions and expectations about what the program must do to bring about the transactions between the target population and the program that will produce the intended changes in social conditions. The program's organizational plan is articulated from the perspective of program management and encompasses both the functions and activities the program is expected to perform and the human, financial, and physical resources required for that performance.
Program process theory	The combination of the program's organizational plan and its service utilization plan into an overall description of the assumptions and expectations about how the program is supposed to operate.
Implementation failure	The program does not adequately perform the activities specified in the program design that are assumed to be necessary for bringing about the intended social improvements. It includes situations in which no service, not enough service, or the wrong service is delivered, or the service varies excessively across the target population.
Theory failure	The program is implemented as planned but its services do not produce the immediate effects on the participants that are expected or the ultimate social benefits intended or both.

CHAPTER 3

IDENTIFYING ISSUES AND FORMULATING QUESTIONS

The previous chapter presented an overview of the many considerations that go into tailoring an evaluation. Although all those matters are important to evaluation design, the essence of the evaluation enterprise is generating credible answers to questions about the performance of a social program. Good evaluation questions must address program issues that are meaningful in relation to the nature of the program and also of concern to key stakeholders. They must be answerable with the research techniques available to the evaluator, and they must be formulated so that the criteria by which the corresponding program performance will be judged are explicit or can be determined in a straightforward way.

A set of appropriate evaluation questions, therefore, is the hub around which evaluation revolves. It follows that a careful, explicit formulation of those questions greatly facilitates both the design of the evaluation and the subsequent use of its findings. Evaluation questions may take various forms, some of which are more useful and meaningful than others for stakeholders and program decisionmakers. Furthermore, some forms of the evaluation questions are more amenable to the evaluator's task of providing credible answers, and some address critical program effectiveness issues more directly than others. Careful development and formulation of the questions that the evaluation will be designed to answer are, therefore, crucial steps in conducting a program evaluation.

This chapter discusses practical ways in which effective evaluation questions can be fashioned from input by stakeholders and analysis by the evaluator. An essential procedure for this purpose is identification of the decisionmakers who will use the evaluation results, what information they need, and how they expect to use it. The evaluator's own analysis of the program is also important. One approach that is particularly useful for this purpose is articulation of the program theory, a detailed account of how and why the program is supposed to work. Consideration of program theory focuses attention on critical events and premises that may be candidates for inquiry in the evaluation.

Program evaluation is fundamentally an endeavor that gathers and interprets information about program performance to answer questions relevant to decision making or, at least, of appreciable interest to one or more program stakeholders. A critical phase in an evaluation, therefore, is the identification and formulation of the questions the evaluation is to address. One might assume that this step would be very straightforward, indeed, that the questions would be stipulated routinely as part of the process of commissioning the evaluation. As described in Chapter 2, however, it is rare for final, workable evaluation questions to be specified by the evaluation sponsor at the beginning of an evaluation. Nor can the evaluator usually step up and define the focal questions unilaterally on the basis of his or her professional expertise. That maneuver would increase the risk that the evaluation would not be responsive to stakeholder concerns, would not be useful or used, and would be attacked as irrelevant or inappropriate.

To ensure that the evaluation will attend to the matters of greatest concern to the pertinent decisionmakers and stakeholders, therefore, the initial evaluation questions are best formulated through interaction and negotiation with those decisionmakers and stakeholders. This, of course, helps direct the evaluation toward the most relevant practical issues. Of equal importance, however, is that such interaction engages key stakeholders in the process of defining the evaluation in a detailed and personal way that increases the likelihood that they will understand, appreciate, and make effective use of the findings when they become available.

Although stakeholder input is critical, the evaluator should not depend only on the decisionmakers and stakeholders to identify the issues the evaluation will address. Sometimes the evaluation sponsors are very knowledgeable about evaluation and will already have done the necessary background work and formulated a complete and workable set of questions to which the evaluation should attend. Such situations are not typical, however. More often, the evaluation sponsors and program stakeholders are not especially expert at evaluation or, if so, have not done all the groundwork needed to focus the evaluation. This means that the evaluator will rarely be presented at the outset with a finished list of every issue the evaluation should address for the results to be useful, interpretable, and complete. Nor will the issues and questions that are put forward generally be formulated in a manner that permits ready translation into research design.

Thus, although the specification of the evaluation questions must involve input from stakeholders, the evaluator's role is also crucial. The stakeholders will be the experts on the practical and political issues facing the program, but generally the evaluator will know the most about how to analyze a program and focus an evaluation. The evaluator, therefore, must be prepared to raise issues for consideration that otherwise might be overlooked, identify aspects of program operations and outcomes that might warrant inquiry, and work with stakeholders to translate their concerns into questions of a form that evaluation research can attempt to answer.

In all but the most routine situations, it is generally wise for the evaluator to construct a written statement of the specific questions that will guide the evaluation design. This provides a reference to consult while designing the evaluation and selecting research procedures that can be very useful. Perhaps more important, this written statement can be discussed with the evaluation sponsor and key stakeholders to ensure that it encompasses their concerns and defines a focus for the evaluation

EXHIBIT 3-A What It Means to Evaluate Something

There are different kinds of inquiry across practice areas, such as that which is found in law, medicine, and science. Common to each kind of inquiry is a general pattern of reasoning or basic logic that guides and informs the practice. . . . Evaluation is one kind of inquiry, and it, too, has a basic logic or general pattern of reasoning [that has been put forth by Michael Scriven]. . . . This general logic of evaluation is as follows:

1. *Establishing criteria of merit.* On what dimensions must the evaluand [thing being evaluated] do well?
2. *Constructing standards.* How well should the evaluand perform?
3. *Measuring performance and comparing with standards.* How well did the evaluand perform?
4. *Synthesizing and integrating data into a judgment of merit or worth.* What is the merit or worth of the evaluand?

. . . To evaluate anything means to assess the merit or worth of something against criteria and standards. The basic logic explicated by Scriven reflects what it means when we use the term *to evaluate.*

SOURCE: Quoted from Deborah M. Fournier, *Establishing Evaluative Conclusions: A Distinction Between General and Working Logic,* New Directions for Evaluation, no. 68 (San Francisco: Jossey-Bass, 1995), p. 16.

acceptable to them. Such a procedure also can safeguard against later misunderstanding of what the evaluation was supposed to accomplish.

The remainder of this chapter examines the two most important topics related to specifying the issues and questions that will guide an evaluation: (a) how to formulate evaluation questions in such a way that they can be addressed using the research procedures available to the evaluator, and (b) how to determine the specific questions on which the evaluation should focus.

WHAT MAKES A GOOD EVALUATION QUESTION?

The form evaluation questions should take is shaped by the functions they must perform. Their principal role is to focus the evaluation on those areas of program performance at issue for key decisionmakers and stakeholders and to facilitate development of a design for data collection that will provide meaningful information about how well the program is performing. A good evaluation question, therefore, must identify a distinct dimension of program performance that is at issue and do so in such a way that the quality of the performance can be credibly assessed. Such assessment, in turn, requires an accurate description of the nature of the performance and some standard by which it can be evaluated (see Exhibit 3-A). Thus a good evaluation question must specify some measurable or observable dimension of program performance in reference to the criterion by which that performance is to be judged. Each of these different aspects warrants further discussion.

Dimensions of Program Performance

Good evaluation questions will identify aspects of performance dimensions that are relevant to the expectations stakeholders hold for the program and represent domains in which the program can realistically hope to have accomplishments, but where effectiveness cannot necessarily be taken for granted. It would hardly be fair to ask if a low-income housing weatherization program reduced the prevalence of drug dealing in a neighborhood. Nor would it generally be useful to ask if the program got a good deal on its purchase of file cabinets for the office. Furthermore, the evaluation questions must involve performance dimensions that are sufficiently specific, concrete, and practical that meaningful information can be obtained about their status. An evaluator would have great difficulty determining if an adult literacy program improved a community's competitiveness in the global economy or if the counselors in a drug prevention program were sufficiently caring in their relations with clients.

Evaluation Questions Must Be Reasonable and Appropriate

Program advocates sometimes put forward grandiose goals (e.g., improve the quality of life for children), expect unrealistically large effects, or believe the program has accomplishments that are disproportionate to its actual capabilities. Good evaluation questions deal with performance dimensions that are appropriate and realistic for the program. This means that the evaluator must often work with relevant stakeholders to scale down and focus the evaluation questions. The manager of a community health program, for instance, might initially ask, "Are our education and outreach services successful in informing the public about the risk of AIDS?" In practice, however, those services may consist of little more than occasional presentations by program staff at Rotary Clubs and health fairs. With this rather modest level of activity, it may not be realistic to expect the public at large to receive much AIDS information through this channel, much less for that information to lower the risk of AIDS in the community. If a question about this service is deemed important for the evaluation, a better version might be something such as "Do our education and outreach services raise awareness of AIDS issues among the audiences addressed?" and "Do those audiences represent community leaders who are likely to influence the opinions of significant others?"

There are two complementary ways for an evaluator, in collaboration with pertinent stakeholders, to assess how appropriate and realistic a candidate evaluation question is. The first is to examine the question in the context of the actual program activities related to it. In the example above, for instance, the low-key nature of the education and outreach services were clearly not up to the task of "informing the public about the risk of AIDS" and there would be little point in having the evaluation attempt to determine if this was accomplished. The evaluator and relevant stakeholders should identify and scrutinize the program components, activities, and personnel assignments that relate to program performance and formulate the evaluation question in a way that is reasonable given those characteristics.

Another form of review for candidate evaluation questions is to analyze them in relationship to the experience and findings reported in applicable social science and social service literature. Questions involving certain program performance dimensions would be assessed as more appropriate if they were consistent with experience in similar programs or studies of those programs. For instance, the sponsor of an evaluation of a program for juvenile delinquents might initially ask if the program increases the self-esteem of the delinquents, in the belief that increased self-esteem is a problem for these juveniles and improvements in self-esteem will lead to better behavior. Examination of the applicable social science research, however, will reveal that juvenile delinquents do not generally have problems with self-esteem and, moreover, that increases in self-esteem are not generally associated with reductions in delinquency. In light of this information, the evaluator and the evaluation sponsor may well agree that the question of the program's impact on self-esteem is not appropriate after all.

The foundation for formulating appropriate and realistic evaluation questions is detailed and complete program description. Early in the process, the evaluator should become thoroughly acquainted with the program—how it is structured, what activities take place, the roles and tasks of the various personnel, the nature of the participants, and the assumptions inherent in its principal functions. The stakeholder groups with whom the evaluator collaborates (especially program managers and staff) will also, of course, have knowledge about the program. Evaluation questions that are inspired by close consideration of actual program activities and assumptions will almost automatically be appropriate and realistic. And a clear understanding of the program operation and its rationale is a necessary prerequisite for entering the social science and social service literature to find relevant concepts and analogous situations. The investigation and articulation of the various components of program theory, described later in this chapter, generally provide a very effective approach to developing this detailed understanding of the program.

Evaluation Questions Must Be Answerable

It is rather obvious that the evaluation questions around which an evaluation plan is developed should be answerable. Questions that cannot be answered may be intriguing to philosophers but serve poorly the needs of evaluators and the decisionmakers who intend to use the evaluation results. What is not so obvious, perhaps, is how easy it is to formulate an unanswerable evaluation question without realizing it. This may occur because the terms used in the question, although seemingly commonsensical, are actually ambiguous or vague when the time comes for a definitive interpretation ("Does this program enhance family values?"). Or sensible-sounding questions may invoke issues for which there are so few observable indicators that little can be learned about them ("Are the case managers sensitive to the social circumstances of their clients?"). Also, some questions lack sufficient indication of the relevant criteria to be answered ("Is this program successful?"). Finally, some questions may be answerable but to do so would require more expertise, data, or resources than are available to the evaluation ("Do the prenatal services this program provides to high-risk women increase the chances that their children will complete college?").

For an evaluation question to be answerable, it must be possible to identify in advance some evidence or "observables" that can realistically be obtained and will be credible as the basis for an answer. This generally means developing questions that involve measurable

performance dimensions, that are sufficiently unambiguous so that explicit, noncontroversial definitions can be given for each of their terms, and for which the relevant standards or criteria are specified or obvious. The best way for an evaluator to test whether a candidate question is answerable in these terms is to determine whether a realistically attainable, specific evaluation finding can be imagined such that the relevant decisionmakers and stakeholders agree that it constitutes a meaningful answer.

Suppose, for instance, that a proposed evaluation question for a compensatory education program like Head Start is, "Are we reaching the children most in need of this program?" To affirm that this is an answerable question, the evaluator should be able to

a. Define the group of children at issue (e.g., those in census tract such and such, age five years or less, living in households with annual income under 150% of the federal poverty level);
b. Identify the specific measurable characteristics and cutoff values that represent the greatest need (e.g., annual income below the federal poverty level, single parent in the household with educational attainment of less than high school);
c. Give an example of the evaluation finding that might result (e.g., 60% of the children currently served fall in the high-need category; 75% of the high-need children in the catchment area are not enrolled in the program);
d. Stipulate the evaluative criteria (e.g., to be satisfactory, at least 90% of the children in the program should be high need and at least 50% of the high-need children in the catchment area should be in the program); and
e. Have the evaluation sponsors and other pertinent stakeholders (who should be involved in the whole process) agree that this would, indeed, answer the question.

If such conditions can be met and, in addition, the resources are available to collect, analyze, and report the applicable data, then the evaluation question can be considered answerable.

Criteria for Program Performance

Beginning a study with a reasonable, answerable question or set of questions, of course, is standard in the social sciences (although often framed as hypotheses). What distinguishes *evaluation* questions is that they have to do with performance and are associated, at least implicitly, with some criteria by which that performance can be judged. Identifying the relevant criteria was mentioned above as part of what makes an evaluation question answerable. However, this is such an important and distinctive aspect of evaluation questions that it warrants separate discussion.

When program managers or evaluation sponsors ask such things as "Are we targeting the right client population?" or "Do our services benefit the recipients?" they are not only asking for a description of the program's performance with regard to serving appropriate clients and providing services that yield benefits. They are also asking if that performance is good enough according to some standard or judgment. There is likely little doubt that at least a few of the "right client population" receive services or that some recipients receive some benefit from services. But is it enough? Some criterion level must be set by which the

EXHIBIT 3-B Many Criteria May Be Relevant to Program Performance

The standards by which program performance may be judged in an evaluation include:

- The needs or wants of the target population
- Stated program goals and objectives
- Professional standards
- Customary practice; norms for other programs
- Legal requirements
- Ethical or moral values; social justice, equity
- Past performance; historical data
- Targets set by program managers
- Expert opinion
- Preintervention baseline levels for the target population
- Conditions expected in the absence of the program (counterfactual)
- Cost or relative cost

numbers and amounts can be evaluated on those performance dimensions.

One implication of this distinctive feature of evaluation is that good evaluation questions will, when possible, convey the performance standard that is applicable as well as the performance dimension that is at issue. Thus, evaluation questions should be much like this: "Is at least 75% of the program clientele appropriate for services?" (by some explicit definition of *appropriate*) or "Do the majority of those who receive the employment services get jobs within 30 days of the conclusion of training that they keep at least three months?" In addition, the performance standards represented in these questions should have some defensible, though possibly indirect, relationship to the social needs the program addresses. There must be some reason why attaining that standard is meaningful, and the strongest rationale is that it represents a level of performance sufficient for that program function to contribute effectively to the overall program purpose of improving the target social conditions.

A considerable complication for the evaluator is that there are many forms in which the applicable performance criteria may appear for various dimensions of program performance (see Exhibit 3-B), and indeed, it is not always possible to establish an explicit, consensual performance standard in advance of collecting data and reporting results. Nonetheless, to the extent that the formulation of the initial evaluation questions includes explicit criteria on which key stakeholders agree, evaluation planning is made easier and the potential for disagreement over the interpretation of the evaluation results is reduced. It is worth noting that the criterion issue cannot be avoided. An evaluation that only describes program performance, and does not attempt to assess it, is not truly an evaluation (by definition; see Exhibit 3-A) and, at most, only pushes the issue of setting criteria and judging performance onto the consumer of the information.

With these considerations in mind, we turn attention to the various kinds of performance criteria that appear in evaluation studies

and may be relevant to formulation of useful evaluation questions. Perhaps the most common criteria are those based on program goals and objectives. In this case, certain desirable accomplishments are identified as the program aims by program officials and sponsors. Often these statements of goals and objectives are not very specific with regard to the nature or level of program performance they represent. One of the goals of a shelter for battered women, for instance, might be to "empower them to take control of their own lives." Although reflecting commendable values, this statement may leave the evaluator uncertain of the tangible manifestations of such empowerment or what level of empowerment constitutes attainment of this goal. Considerable discussion with stakeholders may be necessary to translate such statements into mutually acceptable terminology that describes the intended outcomes more concretely, identifies the observable indicators that correspond to those outcomes, and specifies the level of accomplishment on each that would be considered a success in accomplishing this goal.

Some program objectives, on the other hand, may be very specific. These often come in the form of administrative objectives adopted as targets for routine program functions. The target levels may be set according to past experience or the experience of comparable programs, judgment of what is reasonable and desirable, or maybe only on a "best guess" basis. Examples of administrative objectives may be to establish intake for 90% of the referrals within 30 days, to have 75% of the clients complete the full term of service, to have 85% "good" or "outstanding" ratings on the client satisfaction questionnaire, to provide at least three appropriate services to each person under case management, and the like. There is typically a certain amount of arbitrariness in these criterion levels, but if they are administratively stipulated or can be established through stakeholder consensus and are reasonable, they are quite serviceable in the formulation of evaluation questions and the interpretation of the subsequent findings. However, it is not generally wise for the evaluator to press for such specific statements of target performance levels if the program does not have them or cannot readily and confidently develop them. Setting such targets on a highly arbitrary basis only creates a situation in which they are arbitrarily revised when the evaluation results are in.

In some, albeit rare, instances there are established professional standards that can be invoked as program performance criteria. This is particularly likely in medical and health programs where various practice guidelines and managed care standards have developed and may be relevant for setting desirable performance levels for programs. Much more common, however, is the situation where there are no established criteria or even arbitrary administrative objectives to invoke. A typical situation is one in which the performance dimension itself is clearly recognized, but there is ambiguity about the criterion for good performance on that dimension. For instance, relevant stakeholders may agree that the program should have a low drop-out rate, a high proportion completing service, a high level of client satisfaction, and the like, but only nebulous ideas as to what level constitutes "low" or "high" on the respective dimensions. Sometimes the evaluator can make use of prior experience or find information in the evaluation and program literature that provides a reasonable basis for setting a criterion level. Another approach is to collect judgment ratings from relevant stakeholders to establish the criterion levels or, perhaps more appropriate in such circumstances, to identify broad criterion ranges that can be accepted to distinguish, say, high, medium, and low performance.

When the performance dimension in an evaluation question involves outcome or impact issues, establishing a criterion level can be particularly difficult. Program stakeholders and evaluators alike may have little idea about how much change on a given outcome variable (e.g., a scale of attitude toward drug use) is large and how much is small. By default, these judgments are often made on the basis of statistical criteria. For instance, any statistically significant improvement in an outcome dimension may be viewed as an indication of program success. This is a poor practice for reasons that will be more fully examined later in this volume when impact evaluation is discussed. Statistical criteria have no intrinsic relationship to the practical significance of a change on an outcome dimension and can be misleading. Thus, as much as possible, the evaluator should attempt to determine and specify in practical terms what "success" level is appropriate for judging the nature and magnitude of the program effects.

Typical Evaluation Questions

As should be evident from the discussions above, well-formulated evaluation questions are very concrete and specific to the program at issue and the circumstances of the prospective evaluation. It follows that the variety of questions that might be relevant to some social program or another is enormous. As noted in Chapter 2, however, evaluation questions typically deal with one of five general program issues. Some of the more common questions in each category, stated in summary form, are as follows.

Questions about the need for program services:

- What are the nature and magnitude of the problem to be addressed?
- What are the characteristics of the population in need?
- What are the needs of the population?
- What services are needed?
- How much service is needed, over what time period?
- What service delivery arrangements are needed to provide those services to the population?

Questions about program conceptualization or design:

- What clientele should be served?
- What services should be provided?
- What are the best delivery systems for the services?
- How can the program identify, recruit, and sustain the intended clientele?
- How should the program be organized?
- What resources are necessary and appropriate for the program?

Questions about program operations and service delivery:

- Are administrative and service objectives being met?
- Are the intended services being delivered to the intended persons?
- Are there needy but unserved persons the program is not reaching?

- Once in service, do sufficient numbers of clients complete service?
- Are the clients satisfied with the services?
- Are administrative, organizational, and personnel functions handled well?

Questions about program outcomes:

- Are the outcome goals and objectives being achieved?
- Do the services have beneficial effects on the recipients?
- Do the services have adverse side effects on the recipients?
- Are some recipients affected more by the services than others?
- Is the problem or situation the services are intended to address made better?

Questions about program cost and efficiency:

- Are resources used efficiently?
- Is the cost reasonable in relation to the magnitude of the benefits?
- Would alternative approaches yield equivalent benefits at less cost?

These families of evaluation questions are not mutually exclusive, of course. Questions in more than one category, and maybe in all the categories, could be relevant to a program for which an evaluation was being planned. To develop an appropriate evaluation plan, the many possible questions that might be asked about the program must be narrowed down to those that are most relevant to the program context and the information needs of the key stakeholders (see Exhibit 3-C for an example of evaluation questions for an actual program). We turn now to a discussion of how the evaluator can identify the critical evaluation questions.

DETERMINING THE QUESTIONS ON WHICH THE EVALUATION SHOULD FOCUS

Occasionally, the evaluator is also the evaluation sponsor and primary stakeholder in a program. For instance, an academic researcher who heads a university counseling clinic may have an innovative program concept, implement it in the university clinic, and then conduct an evaluation. It is far more typical, however, for persons other than the evaluator to be the ones who have responsibility for the program, initiate the evaluation, and use the findings. In such circumstances, the evaluation is a project of the sponsor and other involved stakeholders; the evaluator is only the instrument for accomplishing that project. Correspondingly, it is only fitting that the evaluation give central attention to the issues and questions of the evaluation sponsor and the other principal stakeholders. In the discussion that follows, therefore, we first examine the matter of obtaining appropriate input from the evaluation sponsor and relevant stakeholders prior to and during the design stage of the evaluation.

However, it is rarely appropriate for the evaluator to rely only on input from the evaluation sponsor and stakeholders to determine the questions on which the evaluation should focus. Because of their close familiarity with the program, stakeholders may overlook critical, but relatively routine, aspects of program performance. Also, the experience and knowledge of the evaluator may yield distinctive insights into program issues and their inter-

EXHIBIT 3-C Evaluation Questions for a Neighborhood Afterschool Program

An afterschool program located in an economically depressed area uses the facilities of a local elementary school to provide free afterschool care from 3:30 to 6:00 for the children of the neighborhood. The program's goals are to provide a safe, supervised environment for latchkey children and to enhance their school performance through academic enrichment activities. The following are examples of the questions that an evaluation might be designed to answer for the stakeholders in this program:

Is there a need for the program?

Question: How many latchkey children reside within a radius of 1.5 miles of the school? Latchkey children are defined as those of elementary school age who are without adult supervision during some period after school at least once a week during the school year.

Standard: There should be at least 100 such children in the defined neighborhood. The planned enrollment for the program is 60, which should yield enough children in attendance on any given day for efficient staffing, and it is assumed that some eligible children will not enroll for various reasons.

Question: What proportion of the children enrolled in the program are actually latchkey children?

Standard: At least 75% of the enrolled children should meet the definition for latchkey children. This is an administrative target that reflects the program's intent that a large majority of the enrollees be latchkey children while recognizing that other children will be attracted to, and appropriate for, the program even though not meeting that definition.

Is the program well designed?

Question: Are the planned educational activities the best ones for this clientele and the purposes of enhancing their performance in school?

Standard: There should be indications in the educational research literature to show that these activities have the potential to be effective. In addition, experienced teachers for the relevant grade levels should endorse these activities.

Question: Is there a sufficient number of staff positions in the program?

Standard: The staff-student ratio should exceed the state standards for licensed child care facilities.

Is the program implemented effectively?

Question: What is the attendance rate for enrolled children?

Standard: All enrolled children should either be in attendance every afternoon for which they are scheduled or excused with parental permission.

Question: Is the program providing regular support for school homework and related tasks?

Standard: There should be an average of 45 minutes of supervised study time for completion of homework and reading each afternoon, and all the attending children should participate.

Does the program have the intended outcomes?

Question: Is there improvement in the attitudes of the enrolled children toward school?

(continued)

EXHIBIT 3-C Continued

Standard: At least 80% of the children should show measurable improvement in their attitudes toward school between the beginning and end of the school year. Norms for similar students show that their attitudes tend to get worse each year of elementary school; the program objective is to reverse this trend, even if the improvement is only slight.

Question: Is there an improvement in the academic performance of the enrolled children in their regular school work?

Standard: The average term grades on academic subjects should be at least a half letter grade better than they would have been had the children not participated in the program.

Is the program cost-effective?

Question: What is the cost per child for running this program beyond the fixed expenses associated with the regular operation of the school facility?

Standard: Costs per child should be near or below the average for similar programs run in other school districts in the state.

Question: Would the program be equally effective and less costly if staffed by community volunteers (except the director) rather than paid paraprofessionals?

Standard: The annual cost of a volunteer-based program, including recruiting, training, and supporting the volunteers, would have to be at least 20% less than the cost of the current program with no loss of effectiveness to justify the effort associated with making such a change.

relations that are important for identifying relevant evaluation questions. Generally, therefore, it is desirable for the evaluator to make a relatively independent analysis of the program for the purpose of identifying areas of program performance that may be pertinent for investigation.

The second topic addressed in the discussion that follows, therefore, is how the evaluator can analyze a program in a way that will uncover potentially important evaluation questions for consideration in designing the evaluation. An especially useful tool for this purpose is the concept of *program theory*, a depiction of the significant assumptions and expectations on which the program depends for its success. We therefore discuss the different components of program theory, how the evaluator can describe and represent it, and how it can be used diagnostically to identify those program functions that relate most directly to its effectiveness.

Representing the Concerns of the Evaluation Sponsor and Major Stakeholders

In planning and conducting an evaluation, evaluators usually find themselves confronted with multiple stakeholders who hold different

EXHIBIT 3-D Diverse Stakeholder Perspectives on an Evaluation of a Multiagency Program for the Homeless

The Joint Program was initiated to improve the accessibility of health and social services for the homeless population of Montreal through coordinated activities involving provincial, regional, and municipal authorities and more than 20 nonprofit and public agencies. The services developed through the program included walk-in and referral services, mobile drop-in centers, an outreach team in a community health center, medical and nursing care in shelters, and case management. To ensure stakeholder participation in the evaluation, an evaluation steering committee was set up with representatives of the different types of agencies involved in the program and which, in turn, coordinated with two other stakeholder committees charged with program responsibilities.

Even though all the stakeholders shared a common cause to which they were firmly committed —the welfare of the homeless—they had quite varied perspectives on the evaluation. Some of these were described by the evaluators as follows:

> The most glaring imbalance was in the various agencies' different organizational cultures, which led them to experience their participation in the evaluation very differently. Some of the service agencies involved in the Joint Program and its evaluation were front-line public organizations that were accustomed to viewing their actions in terms of a mandate with a target clientele. They were familiar with the evaluation process, both as an administrative procedure and a measurement of accountability. Among the nonprofit agencies, however, some relative newcomers who had been innovators in the area of community-based intervention were hoping the evaluation would recognize the strengths of their approach and make useful suggestions for improvement. Other nonprofit groups were offshoots of religious or charitable organizations that had been involved with the homeless for a very long time. For those groups the evaluation (and the logical, planning-based program itself) was a procedure completely outside of anything in their experience. They perceived the evaluators as outsiders meddling in a reality that they had managed to deal with up until now, under very difficult conditions. Their primary concern was the client. More than the public agencies, they probably saw the evaluation as a waste of time, money, and energy. Most of the day centers involved in the program fell into this category. They were the ones who were asked to take part in a process with which they were unfamiliar, alongside their counterparts in the public sector who were much better versed in research procedures. (p. 471)

SOURCE: Quoted, with permission, from Céline Mercier, "Participation in Stakeholder-Based Evaluation: A Case Study," *Evaluation and Program Planning*, 1997, 20(4):467-475.

and sometimes conflicting views on the program or its evaluation and whose interests will be affected by the outcome (see Exhibit 3-D for an illustration). At the planning stage of an

evaluation, the evaluator usually attempts to identify all the stakeholders with an important point of view on what questions should be addressed in the evaluation, set priorities among those viewpoints, and integrate as many of the relevant concerns as possible into the evaluation plan.

The starting point, of course, is with the evaluation sponsors. Those who have commissioned and funded the evaluation rightfully have priority in defining the issues it should address. Sometimes evaluation sponsors have stipulated the evaluation questions and methods completely and want the evaluator only to manage the practical details. In such circumstances, the evaluator should assess which, if any, stakeholder perspectives are excluded and whether they are sufficiently distinct and important that their omission compromises the evaluation. If so, the evaluator must then decide whether to conduct the evaluation under the specified constraints, reporting the limitations and biases along with the results, or attempt to negotiate an arrangement whereby the evaluation is broadened to include additional perspectives.

More often, however, the evaluation sponsors' initial specifications are not so constrained or nonnegotiable that the concerns of other stakeholders cannot be considered. In this situation, the evaluator typically makes the best attempt possible within the constraints of the situation to consult fully with all stakeholders, set reasonable priorities, and develop an evaluation plan that will enhance the information available about the respective concerns of all parties.

Given the usual multiplicity of program stakeholders and their perspectives, and despite an evaluator's efforts to be inclusive, there is considerable inherent potential for misunderstandings to develop between the evaluator and one or more of the stakeholders regarding what issues the evaluation should address. It is especially important, therefore, that there be full and frank communication between the evaluator and the pertinent stakeholder groups from the earliest possible point in the planning process. Along with obtaining critical input from the stakeholders about the program and the evaluation, this exchange should emphasize realistic, shared understanding of what the evaluation will and will not do, and why. Most essentially, the evaluator should strive to ensure that the key stakeholders understand, and find acceptable, the nature of the evaluation process, the type of information the evaluation will produce, what it might mean if the results come out one way or another, and what ambiguities or unanswered questions may remain.

Obtaining Input From Stakeholders

The major stakeholders, by definition, have a significant interest in the program and the evaluation. It is thus generally straightforward to identify them and obtain their views about the issues and questions to which the evaluation should attend. The evaluation sponsor, program administrators (who may also be the evaluation sponsor), and intended program beneficiaries are virtually always major stakeholders. Identification of other important stakeholders can usually be accomplished by analyzing the network of relationships surrounding a program. The most revealing relationships involve the flow of money to or from the program, political influence on and by the program, those whose actions affect or are affected by the program, and the set of direct interactions between the program and its various boards, patrons, collaborators, competitors, clients, and the like.

A *snowball sampling* approach is often helpful in identifying the various stakeholder groups and persons involved in relationships with the program. As each such representative is identified and contacted, the evaluator asks for nominations of other persons or groups who have a significant interest in the program or are likely to have useful information about it. Those representatives, in turn, are asked the same question. When this process no longer produces consequential new nominations, the evaluator can be reasonably assured that all major stakeholders have been identified.

If the evaluation is structured as an explicitly collaborative or participatory endeavor so that certain stakeholders are directly involved in designing and conducting the evaluation (as described in Chapter 2), they will, of course, have a firsthand role in shaping the evaluation questions. Similarly, an internal evaluator who is part of the organization that administers the program will likely receive forthright counsel from program personnel. Even when such stakeholder involvement is built into the way the evaluation is organized, however, this arrangement is usually not sufficient to represent the full range of pertinent stakeholder perspectives. There may be important stakeholder groups that are not involved in the participatory structure but have distinct and significant perspectives on the program and the evaluation. Moreover, there may be a range of viewpoints among the members of those groups that are represented in a participatory evaluation process so that a broader sampling of opinion is needed than that brought by the designated participant on the evaluation team.

Generally, therefore, formulating responsive evaluation questions requires some discussion with members of stakeholder groups who are not directly represented on the evaluation team. Fewer such contacts may be needed by evaluation teams that already represent many stakeholders and more by those on which few or no stakeholders are represented. In cases where the evaluation has not initially been organized as a collaborative endeavor with stakeholders, the evaluator may wish to consider configuring such an arrangement to ensure engagement by key stakeholders and full representation of their views in the evaluation design and implementation. Similarly, various participatory arrangements might be made through stakeholder advisory boards, steering committees, or simply involvement of key stakeholder representatives in regular consultation with the evaluator. More information about the procedures and benefits of such approaches can be found in Fetterman, Kaftarian, and Wandersman (1996), Greene (1988), Mark and Shotland (1985), and Patton (1997).

Outside of organized arrangements, evaluators generally obtain stakeholder views about the important evaluation issues through personal or telephone interviews. Because these early contacts with stakeholders are primarily for orientation and reconnaissance, such interviews are typically unstructured or, perhaps, semistructured around a small set of themes of interest to the evaluator. Input from some number of individuals representing one or more stakeholder groups might also be obtained through focus groups (Krueger, 1988). Focus groups have the particular advantages of efficiency in getting information from a number of people and the facilitative effect of group interaction in stimulating ideas and observations. They also may have some disadvantages for this purpose, notably the potential for conflict in politically volatile situations and the lack of confidentiality in group settings. In some cases, therefore, stakeholder informants may speak more frankly about the program and

the evaluation one-on-one with the evaluator than they will in a focus group.

The evaluator will rarely be able to obtain input from every member of every stakeholder group, nor will that ordinarily be necessary to identify the major issues and questions with which the evaluation should be concerned. A modest number of carefully selected stakeholder informants who are representative of significant groups or distinctly positioned in relation to the program is typically sufficient to identify the principal issues. When the evaluator no longer hears new themes in discussions with diverse stakeholders, the most significant prevailing issues have probably all been discovered.

Topics for Discussion With Stakeholders

As mentioned in the previous chapter, the issues identified by the evaluation sponsor when the evaluation is requested usually need further discussion with the sponsor and other stakeholders to clarify what they mean to the various parties and what sort of information would usefully bear on them. This endeavor may then lead to refinement and revision of the questions the evaluation will address. The topics that should be addressed in these discussions will depend in large part on the particulars of the evaluation situation. We will review some of the general topics that are often relevant.

Why is an evaluation needed? It is usually worthwhile for the evaluator to probe the reasons an evaluation is desired with the evaluation sponsor and other stakeholders. The evaluation may be motivated by an external requirement, in which case it is important to know the nature of that requirement and what use is likely to be made of the results. The evaluation may be desired by program managers to determine if the program is effective, to find ways to improve it, or to "prove" its value to potential funders, donors, critics, or the like. Sometimes the evaluation is politically motivated only, for example, as a stalling tactic for a controversial program. Whatever the reasons, they provide an important starting point for determining what questions will be most important for the evaluation to answer and for whom.

What are the program goals and objectives? Inevitably, whether a program achieves certain of the goals and objectives ascribed to it will be pivotal questions for the evaluation to answer. The distinction between goals and objectives is critical. *Goals* are typically stated by programs in broad and rather abstract terms. For evaluation purposes, such goal statements must be refined and restated in terms that can be measured. For example, a program for the homeless may have as its goal "the reduction of homelessness" in its urban catchment area. Although easily understood, such a goal is too vague to support agreement that it has or has not been met. Is a "reduction of homelessness" 5%, 10%, or 100%? Does it refer to only those who are homeless or also to those who are marginally housed and at imminent risk of homelessness? For evaluation purposes, these broad goals must be translated into concrete statements that specify the condition to be dealt with together with one or more measurable criteria of success. Evaluators generally refer to these more specific statements of measurable attainments as *objectives*. Exhibit 3-E presents helpful rules for specifying objectives.

An important task for the evaluator, therefore, is to collaborate with the evaluation sponsors, program managers, and other relevant

EXHIBIT 3-E Some Rules for Specifying Objectives

Four techniques are particularly helpful for writing useful objectives: (a) using strong verbs, (b) stating only one purpose or aim, (c) specifying a single end-product or result, and (d) specifying the expected time for achievement (Kirschner Associates, 1975).

A "strong" verb is an action-oriented verb that describes an observable or measurable behavior that will occur. For example, "to increase the use of health education materials" is an action-oriented statement involving behavior which can be observed. In contrast, "to promote greater use of health education materials" is a weaker and less specific statement. The term "promote" is subject to many interpretations. Examples of action-oriented, strong verbs include: "to write," "to meet," "to find," "to increase," and "to sign." Examples of weaker, nonspecific verbs include: "to understand," "to encourage," "to enhance," and "to promote."

A second useful suggestion for writing a clear objective is to state only a single aim or purpose. Most programs will, of course, have multiple objectives, but within each objective only a single purpose should be delineated. An objective that states two or more purposes or desired outcomes may well require different implementation and assessment strategies, making achievement of the objective difficult to determine. For example, the statement "to begin three prenatal classes for pregnant women and provide outreach transportation services to accommodate twenty-five women per class" creates difficulties. This objective contains two aims—to provide prenatal classes and to provide outreach services. If one aim is accomplished but not the other, to what extent has the objective been met?

Specifying a single end-product or result is a third technique contributing to a useful objective. For example, the statement "to begin three prenatal classes for pregnant women by subcontracting with City Memorial Hospital" contains two results, namely, the three classes and the subcontract. It is better to state these objectives separately, particularly since one is a higher-order objective (to begin three prenatal classes) which depends partly on fulfillment of a lower-order objective (to establish a subcontract).

A clearly written objective must have both a single aim and a single end-product or result. For example, the statement "to establish communication with the Health Systems Agency" indicates the aim but not the desired end-product or result. What constitutes evidence of communication —telephone calls, meetings, reports? Failure to specify a clear end-product makes it extremely difficult for assessment to take place.

Those involved in writing and evaluating objectives need to keep two questions in mind. First, would anyone reading the objective, with or without knowledge of the program, find the same purpose as the one intended? Second, what visible, measurable, or tangible results are present as evidence that the objective has been met? Purpose or aim describes what will be done; end-product or result describes evidence that will exist when it has been done. This is assurance that you "know one when you see one."

Finally, it is useful to specify the time of expected achievement of the objective. The statement "to establish a walk-in clinic as soon as possible" is not a useful objective because of the vagueness of "as soon as possible." It is far more useful to specify a target date, or in cases where some uncertainty exists about some specific date, a range of target dates—for example, "sometime between March 1 and March 30"—is also useful.

SOURCE: Adapted, with permission, from Stephen M. Shortell and William C. Richardson, *Health Program Evaluation* (St. Louis, MO: C. V. Mosby, 1978), pp. 26-27.

stakeholders to identify the program goals and transform overly broad, ambiguous, or idealized representations of them into clear, explicit, concrete statements of objectives. The more closely the objectives describe situations that can be directly and reliably observed, the more likely it is that a meaningful evaluation will result. Furthermore, it is essential that the evaluator, evaluation sponsors, and other pertinent stakeholders achieve a workable agreement on which program objectives are most central to the evaluation and the criteria to be used in assessing whether those objectives have been met. For instance, if one stated objective of a job training program is to maintain a low drop-out rate, the key stakeholders should agree to its importance before it is accepted as one of the focal issues around which the evaluation will be designed.

If consensus about an appropriate criterion is weak, or not attained at all, it may be wise for the evaluator to employ multiple criteria that reflect the interests of the various stakeholders concerned with a particular objective. If consensus is weak or nonexistent about which objectives are important, one solution is to include all those put forward by the various stakeholders and, perhaps, additional objectives drawn from current viewpoints and theories in the relevant substantive field (Chen, 1990). For example, the sponsors of a job training program may be interested solely in the frequency and duration of postprogram employment. But the evaluator may propose that stability of living arrangements, competence in handling finances, and efforts to obtain additional education be examined as program outcomes because these lifestyle features also may undergo positive change with increased employment and job-related skills.

What are the most important questions for the evaluation to answer? With an understanding of why an evaluation is desired and by whom, and a careful specification of the program objectives that key stakeholders agree are central to the evaluation, attention can be given to formulating the questions the evaluation will be designed to answer. We echo Patton's (1997) view that the delineation of priority evaluation questions should be organized around a concept that generally concerns evaluators very much: *utilization*. Evaluation results are rarely intended by evaluators or evaluation sponsors to be "knowledge for knowledge's sake." Rather, they are intended to be useful, and to be used, by those with responsibility for making decisions about the program, whether at the day-to-day management level or at broader funding or policy levels (see Exhibit 3-F for an evaluation manager's view of this process).

Unfortunately, the experience of evaluators is replete with instances of evaluation findings that were virtually ignored by those to whom they were reported. There are numerous reasons why this may happen, many of which are not within the control of the evaluator. Program circumstances may change, for instance, between the initiation of the evaluation and its completion in ways that make the evaluation results irrelevant when they are delivered. But lack of utilization may also occur because the evaluation does not actually provide information useful to the decisionmakers for the decisions they must make. Moreover, this can happen rather innocently as well as through ineptness. It may well be, for instance, that an evaluation plan looks like it will produce relevant information but, when that information is generated, it is not as useful as the recipients expected. It may also happen that those to whom the evaluation results are directed are not initially altogether clear in their own minds about what information they need for what purposes.

With these considerations in mind, we advocate that the development of evaluation questions involve *backward mapping*, which starts with a specification of the desired endpoint then works backward to determine what must be done to get there (Elmore, 1980). Taking this approach, the essential discussion with the evaluation sponsor and other key stakeholders must establish who will use the evaluation results and for what purposes. Note that the question is not who is *interested* in the evaluation findings. Although relevant, that question does not probe the matter of what actions or decisions are potentially affected. The evaluator wants to come as close as possible to understanding, in an explicit, detailed fashion, who specifically will use the evaluation and what specifically they will use it for. For instance, the administrator and board of directors of the program may intend to use the evaluation results to set administrative priorities for the next fiscal year. Or the legislative committee that oversees a program area may desire the evaluation as input to their deliberations about continued funding for the program. Or the program monitors in the government agency that has initiated the program may want to know if it represents a successful model that should be disseminated to other sites.

EXHIBIT 3-F Lessons Learned About the Utilization of Evaluation

An evaluation manager for a social services organization summarized his observations about the use of evaluation findings by program decisionmakers as follows:

1. The utilization of evaluation or research does not take care of itself. Evaluation reports are inanimate objects, and it takes human interest and personal action to use and implement evaluation findings and recommendations. The implications of evaluation must be transferred from the written page to the agenda of program managers.
2. Utilization of evaluation, through which program lessons are identified, usually demands changed behaviors or policies. This requires the shifting of priorities and the development of new action plans for the operational manager.
3. Utilization of evaluation research involves political activity. It is based on a recognition and focus on who in the organization has what authority to make *x*, *y*, or *z* happen. To change programs or organizations as a result of some evaluation requires support from the highest levels of management.
4. Ongoing systems to engender evaluation use are necessary to legitimate and formalize the organizational learning process. Otherwise, utilization can become a personalized issue and evaluation advocates just another self-serving group vying for power and control.

SOURCE: Quoted, with permission, from Anthony Dibella, "The Research Manager's Role in Encouraging Evaluation Use," *Evaluation Practice*, 1990, 11(2):119.

In each case, the evaluator should work with the respective evaluation users to describe the range of potential decisions or actions that they might consider taking and the form and nature of information that they would find pertinent in their deliberation. To press this exercise to the greatest level of specificity, the evaluator might even generate dummy information of the sort that the evaluation might produce, for example, "20% of the clients who complete the program relapse within 30 days," and discuss with the prospective users what this would mean to them and how they would use such information.

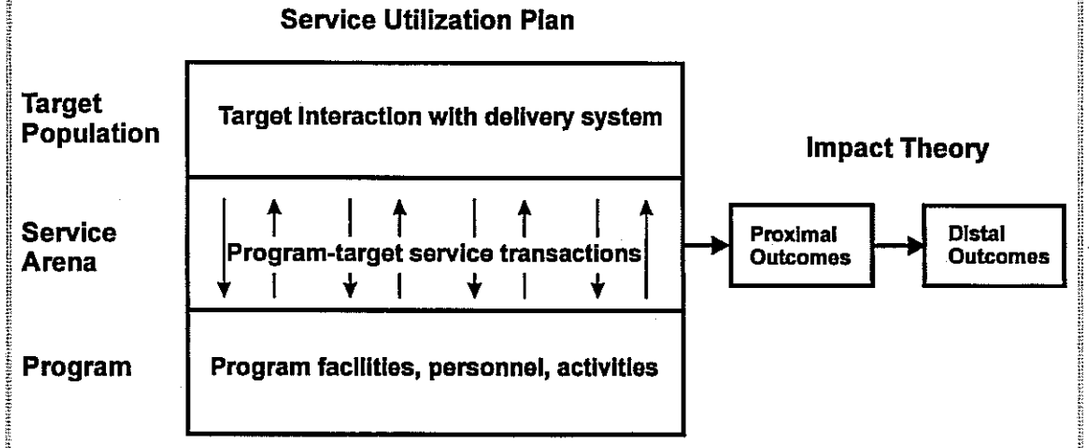
A careful specification of the intended use of the evaluation results and the nature of the information that is expected to be useful leads directly to the formulation of questions the evaluation must attempt to answer (e.g., "What proportion of the clients who complete the program relapse during the first month?") and provides a context within which to set priorities for which questions are most important. At this juncture, consideration must also be given to matters of timing. It may be that some questions must be answered before others can be asked, or users may need answers to some questions before others because of their own timetable for decision making. The important questions can then be organized into related groups, combined and integrated as appropriate, sequenced in appropriate time lines, and worked into final form in consultation with the designated users. With this in hand, developing the evaluation plan is largely a matter of working backward to determine what measures, observations, procedures, and the like must be undertaken to provide answers to the important questions in the form that the users require by the time they are needed.

Analysis of Program Assumptions and Theory

Evaluation is about assessing how the program is performing, whether at some global level or with regard to specific functions and aspects. Most evaluation questions, therefore, are variations on the theme of "Is what's supposed to be happening actually happening?" for example, "Are the intended target participants being reached?" "Are the services adequately delivered?" or "Are the goals being met?" A very useful analysis of a program for purposes of identifying relevant and important evaluation questions is to delineate in some detail just what it is that is supposed to be happening in a program. The evaluator can construct a representation, a conceptual model, of how the program is expected to work and the connections presumed between its various activities and functions and the social benefits it is intended to produce. This representation of the program assumptions and expectations can then be used to identify those aspects of the program most essential to effective performance. These, in turn, raise evaluation-related questions about whether the key assumptions and expectations are reasonable and appropriate and, if so, whether the program is enacting them in an effective manner.

What we are describing here is an explication of the program theory, the set of assumptions about the relationships between the strategy and tactics the program has adopted and the social benefits it is expected to produce. *Theory* has a rather grandiose sound to it and few program directors would claim that they were working from any distinct theory. Among the dictionary definitions of theory, however, we find "a particular conception or view of

EXHIBIT 3-G Overview of Program Theory



something to be done or of the method of doing it." It is generally this sense of the word that evaluators mean when they refer to program theory. It might alternatively be called the program conceptualization or, perhaps, the program plan, blueprint, or design.

Evaluators have long recognized the importance of program theory as a basis for formulating and prioritizing evaluation questions, designing evaluation research, and interpreting evaluation findings (Bickman, 1987; Chen and Rossi, 1980; Weiss, 1972; Wholey, 1979). It is, however, described and used under various different names, for example, logic model, program model, outcome line, cause map, action theory, and so forth. Moreover, there is no general consensus about how best to depict or represent program theory, and many different versions can be found in the evaluation literature, although all show common elements. Consequently, we will describe representations of several separate components of program theory that we have found useful in our own evaluation activities and that illustrate themes found in most variations of this type of analysis.

For this purpose, we depict the typical social program as centering on a set of program-target transactions, those points of direct contact between program operations and the target population the program serves that occur in some service arena (see Exhibit 3-G). These might involve counseling sessions for women with eating disorders in therapists' offices, recreational activities for high-risk youths at a community center, educational presentations to local citizens' groups, nutrition posters in a clinic, informational pamphlets about empowerment zones and tax law mailed to potential

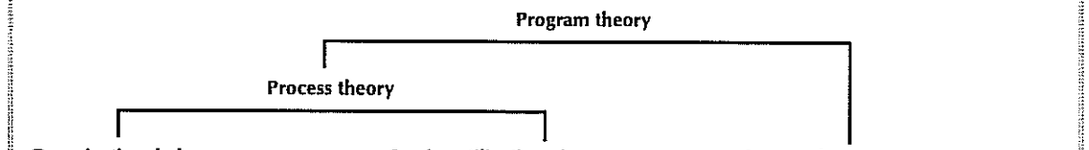
investors, delivery of meals to the front doors of elderly persons, or any such point of service contact. On one side of this program-target transaction, we have the program as an organizational entity, with its various facilities, personnel, resources, activities, and so forth. On the other side, we have the target participants in their lifespaces with their various situations and behaviors, including their circumstances and experiences in relation to the service delivery system that provides them with points of contact with the program.

For purposes of explicating and analyzing program theory, this simple scheme highlights three different, but interrelated, theory components, each of which focuses attention on an important facet of program performance. Most important are the program-target transactions, for they constitute the means by which the program expects to bring about its intended effects. These transactions are thus operationalizations of the program's *impact theory*, the assumptions about the change process actuated or facilitated by the program and the improved conditions expected to result from inducing that change. This impact theory may be as simple as presuming that exposure to information about the negative effects of drug abuse will motivate high school students to abstain or as complex as the ways in which an eighth-grade science curriculum will lead to deeper understanding of natural phenomena. It may be as informal as the commonsense presumption that providing hot meals to elderly persons improves their nutrition or as formal as classical conditioning theory adapted to treating phobias. Whatever its nature, however, an impact theory of some sort constitutes the essence of a social program. If the assumptions embodied in that theory about how desired changes are brought about by program action are faulty, or if they are valid but not well operationalized by the program, the intended social benefits will not be achieved.

To instigate the change process posited in the program's impact theory, the program must first provide the intended services to the target population. If we view the program from the perspective of the target population, attention focuses on the points of service delivery and their accessibility, whether the services are actually delivered to the intended targets, and the extent to which targets complete the full sequence of services. Every program works within a framework of assumptions and expectations about how to reach the target population, provide and sequence service contacts with those designated as clients, and conclude the relationship when services are no longer needed or appropriate. These assumptions and expectations constitute an important part of program theory that we will call the program's *service utilization plan*.

In simplest form, a service utilization plan proposes that if the intended targets experience particular encounters and opportunities provided by the program's service delivery system, they will receive the intended services. For a program to increase awareness of AIDS risk, for instance, the service utilization plan may be simply that appropriate persons will read informative posters if they are put up in subway cars. A multifaceted AIDS prevention program, on the other hand, may be organized on the assumptions that if high-risk drug abusers in specified neighborhoods encounter outreach workers and are referred to clinics, and if street-front clinics are available nearby, and if clients receive encouragement from case managers to maintain continuing program contact, and if they receive testing and information at the clinics, then high-risk drug abusers will have received the preventive service package the program intends to deliver.

EXHIBIT 3-H The Three Components of Program Theory



The program, of course, must be organized in such a way that it can, indeed, actually provide the intended services, which, in turn, are expected to produce the desired benefits. The third component of program theory, therefore, has to do with the nature of the program resources, personnel, administration, and general organization. It might be called the program's *organizational plan*. It can generally be represented as a set of propositions: If the program has such and such resources, facilities, personnel, and so on, is organized and administered in such and such a manner, and engages in such and such activities and functions, then a viable organization will result with the capability of developing and/or maintaining the intended service delivery system and corresponding service utilization. Elements of programs' organizational theories include such presumptions as that case managers should have master's degrees in social work and at least five years' experience, that at least 20 case managers should be employed, that the agency should have an advisory board that represents local business owners, that there should be an administrative coordinator assigned to each site, that working relations should be maintained with regard to referrals from the Department of Public Health, and so forth.

Adequate resources and effective organization, in this scheme, are the factors that make it possible to develop and maintain a service delivery system that enables utilization of the services so that the target population receives the intervention. Program organization and the service delivery system it supports are the parts of the program most directly under the control of program administrators and staff. These two aspects together are often referred to as *program process*, and correspondingly, the assumptions and expectations on which program process is based may be called the program's *process theory*.

The intervention the program implements as a result of its organizational and service delivery activities, in turn, is the means by which the program expects to bring about the desired changes in the target population or social conditions. Thus, all three theory components are closely interrelated and, collectively, can be viewed as constituting the overall program theory (Exhibit 3-H gives a summary

of the theory components). With this overview, we turn now to a more detailed discussion of each of these theory components with particular attention to how the evaluator can construct a workable representation of program theory and use it to analyze the program and generate potentially important evaluation questions.

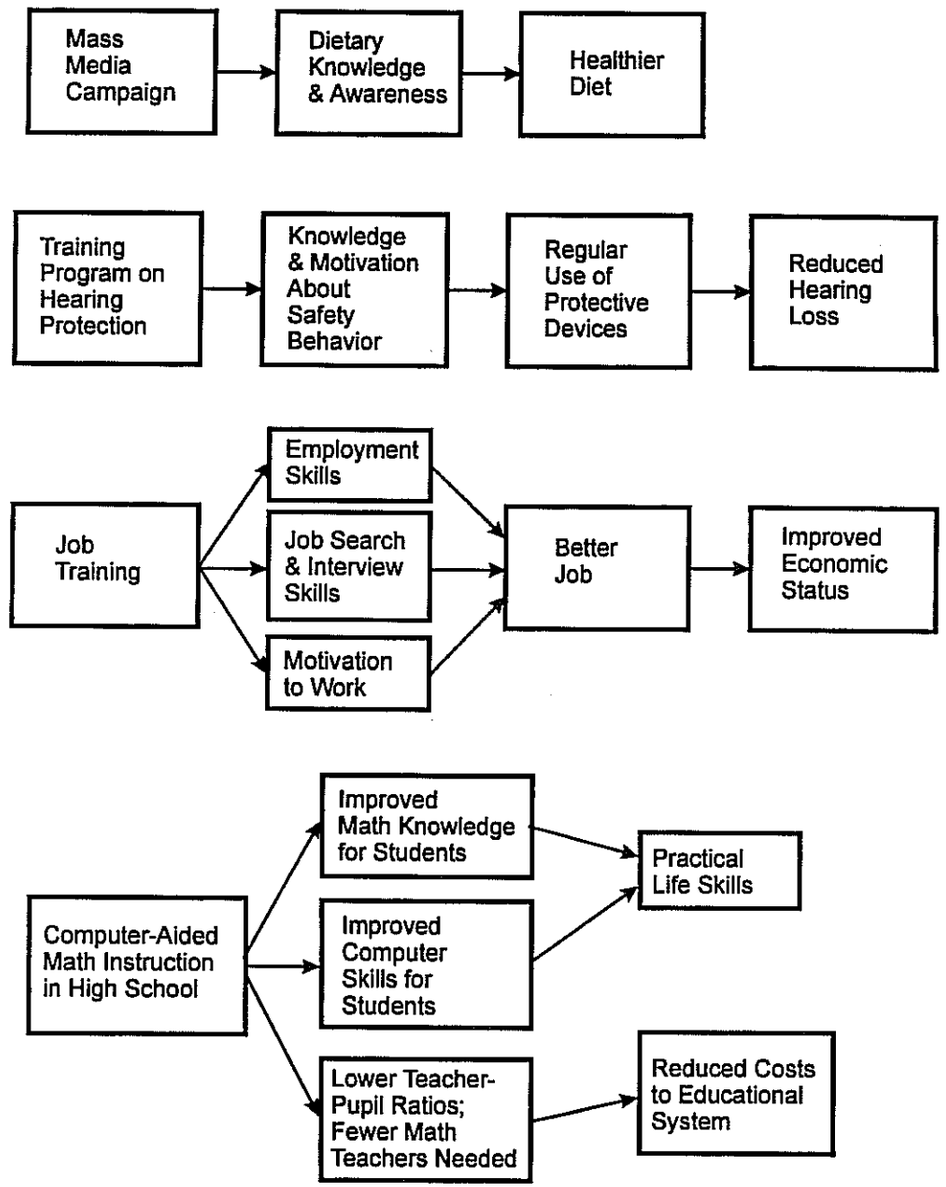
The Program Impact Theory

The central premise of any social program is that the services it delivers to the target population induce some change that improves social conditions. The program impact theory is the set of assumptions embodied in the program about how its services actuate or facilitate the intended change. Program impact theory, therefore, is causal theory: It describes a cause-and-effect sequence in which certain program activities are the instigating causes and certain social benefits are the effects they eventually produce. Evaluators, therefore, typically represent program impact theory in the form of a causal diagram showing the pattern of cause-and-effect linkages presumed to connect the program activities with the expected outcomes (Chen, 1990; Lipsey, 1993; Martin and Kettner, 1996). Because programs rarely exercise complete, direct control over the social conditions they are expected to improve, they must generally work indirectly by attempting to alter some critical but manageable aspect of the situation, which, in turn, is expected to lead to more far-reaching improvements. For instance, a program cannot make it impossible for people to abuse alcohol, but it can attempt to change their attitudes and motivation toward alcohol in ways that help them avoid abuse. Similarly, a program may not be able, at a stroke, to eliminate poverty in a target population, but it may be able to help unemployed persons prepare for and find jobs that pay a living wage.

The simplest program impact theory, therefore, is generally the basic "two step" in which services change some intermediate condition such as motivation or employability that, in turn, helps ameliorate the social conditions of concern, for example, by reducing alcohol abuse or unemployment (Lipsey and Pollard, 1989). More complex program theories may have more steps along the path between program and social benefit and, perhaps, involve more than one distinct path. Exhibit 3-I illustrates causal diagrams for several different program impact theories. The distinctive features of any representation of program impact theory are that each element is either a cause or an effect and that the causal linkages between those elements show a chain of events that begins with program actions and ends with change in the social conditions the program ultimately intends to improve.

Depiction of the program's impact theory has considerable power as a framework for analyzing a program and generating significant evaluation questions. First, the process of making that theory explicit brings a sharp focus to the nature, range, and sequence of program outcomes that are reasonable to expect and may be appropriate for the evaluator to investigate. Every event following the instigating program activity in the causal diagram representing a program's impact theory is an outcome. Those following directly from the instigating program activities are the most direct outcomes, often called *proximal* or immediate outcomes, whereas those further down the chain constitute the more *distal* or ultimate outcomes. Program impact theory highlights the dependence of the more distal, and generally more important, outcomes on successful attainment of the more proximal ones. For a

EXHIBIT 3-I Diagrams Illustrating Program Impact Theories



full understanding of program impact, therefore, it may be important for the evaluation to examine the proximal outcomes even when they are not themselves the accomplishments for which the program will be held accountable.

A second, and related, contribution of program impact theory to formulation of key evaluation questions is the distinction it reveals between two rather different sets of assumptions inherent in the program. The first set of assumptions represents the expectation that the program actions will have the intended effects on the proximal or immediate outcomes. For instance, a mass media campaign about AIDS must assume that the public service announcements, billboards, and other promotion it does (program actions) will result in heightened awareness and knowledge of the risk of unsafe sex practices (proximal outcomes). This set of assumptions thus links program actions to the immediate outcomes expected to follow from them and has been referred to as the program's "action theory" (Chen, 1990; see also Exhibit 3-J). Because it is only one link in the impact theory, however, we would prefer to call it the *action hypothesis*. Articulating the action hypothesis allows the evaluator to identify the important evaluation questions that relate to it, particularly with regard to whether the intended actions were implemented and the expected proximal effects achieved.

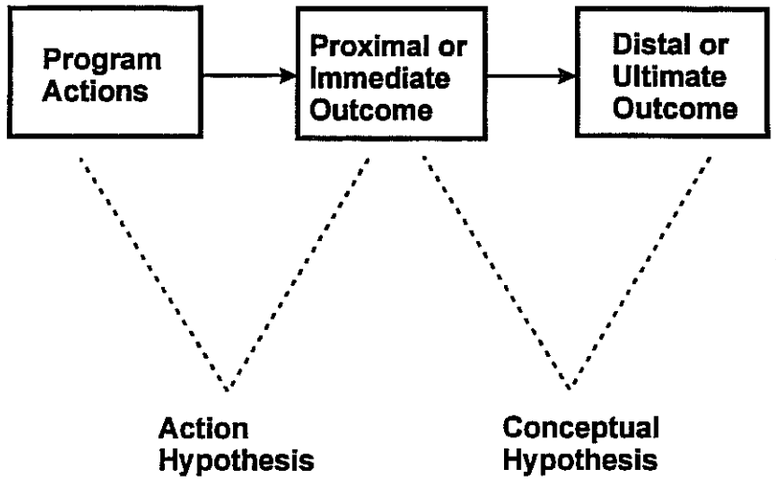
The second set of assumptions inherent in program impact theory connects the proximal outcomes with the distal ones. In the mass media campaign, for instance, it is expected that if knowledge and awareness are heightened (proximal outcomes), appropriate safe-sex behavior will follow with corresponding decreases in AIDS transmission (distal outcomes). This part of the process is completely out of the control of the program; it is only assumed that if the program does its part by implementing the campaign in such a way that knowledge and awareness are heightened, the intended social benefits will follow. These assumptions have been referred to as the program's "conceptual theory" (Chen, 1990; see also Exhibit 3-J), although, again, we would prefer *conceptual hypothesis*. This hypothesis is the part of the impact theory that assumes that success in changing the targeted aspect of the problem (proximal outcomes) will, in turn, result in the desired social benefits (distal outcomes).

This aspect of impact theory, of course, draws the evaluator's attention to another set of linkages that might bear investigation and helps formulate questions about whether, given proximal outcome *A*, distal outcome *B* actually follows. The evaluator, therefore, might find it important not only to ask if awareness and knowledge of AIDS risk increased, but if such increases were further associated with changed behavior and reduced incidence of AIDS. This line of analysis helps identify the full set of outcome variables potentially relevant to an impact evaluation.

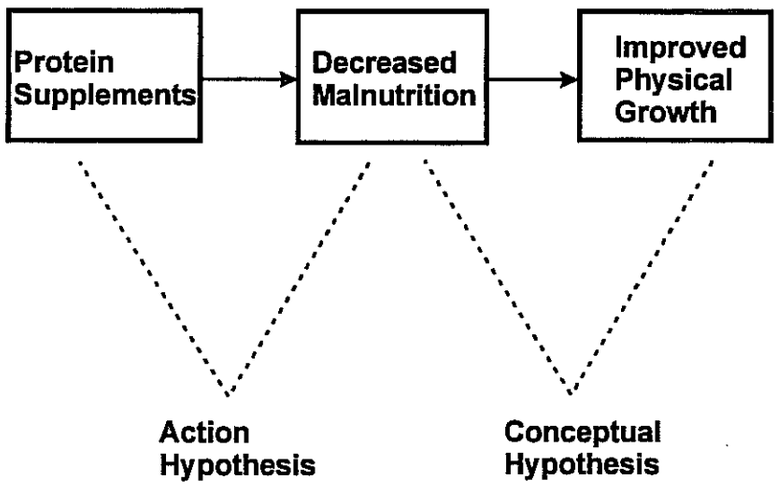
Of course, the intended outcomes may not be achieved by a program and the evaluator will generally want to be in a position to explain why unfavorable results occurred. One of the useful applications of program impact theory is for identification of the various points in the anticipated chain of events where things may not happen as expected. At a minimum, evaluators may wish to distinguish *implementation failure* from *theory failure*.

Implementation failure is the failure of the program to adequately deliver the services or perform the actions that are supposed to start off the change process expected to lead to improved social conditions (Exhibit 3-K provides an example). Obviously, if the program is not

EXHIBIT 3-J Program Impact Theory: The Action and Conceptual Hypotheses



Example:



SOURCE: Adapted from Huey-Tsyh Chen, *Theory-Driven Evaluations* (Newbury Park, CA: Sage, 1990).

EXHIBIT 3-K Implementation Failure: A Case Study of a School Not Ready to Support Change

A team of administrators and staff from a Southern school district worked with evaluators to plan a program to reduce problem behaviors—including drug and alcohol use, delinquent behavior, pregnancy, nonattendance, and misconduct—in a middle school with a large proportion of high-risk students. A program theory was developed on the basis of research showing that improvements in social bonding, social competency skills, and school success should result in a reduction in problem behaviors. Funding was obtained to support a four-year program with the following components:

- *Instructional improvement:* Cooperative learning techniques were to be used schoolwide. In addition, for high-risk students, one-on-one tutoring was to be provided by community volunteers.
- *Mentoring:* High-risk students were to be paired with teachers who volunteered as "academic godparents" to tutor them, monitor their progress, and share in recreational activities.
- *Social competency promotion:* The Botvin life skills training (LST) and Manning's cognitive self-instruction (CSI) curriculum were to be implemented schoolwide. These were augmented with a social problem solving (SPS) course for seventh graders, a violence prevention (VP) curriculum for eighth graders, and a career and educational decision skills (CED) class for sixth and eighth graders.

Assessment of the implementation of this multicomponent program revealed the following:

- The first school year was largely a start-up period. A group of teachers was trained to use cooperative learning techniques, but only 13 actually used them. The CED course was implemented but not all the lessons were covered. Several teachers were trained to teach the LST course and a portion of it was taught in health classes.
- During the second school year, cooperative learning was implemented by more than half the teachers but in fewer than half their lessons. The CED course was provided to most of the eighth graders with about half the intended number of sessions. Seventh-grade high-risk students received the SPS course but eighth graders got neither the LST nor VP modules. About one-third of the high-risk students received tutoring, but the average was only five sessions for the year.
- In the third year, all the program components except tutoring and mentoring were incorporated into a single life focus course adapted to each grade level. Not all the intended material was covered, however, so some students did not get some of the components and, in other cases, received fewer lessons than intended. The mentoring component improved from the previous year, but the high-risk tutoring component deteriorated. Cooperative learning was implemented more fully but still at only about two-thirds the intended level.
- During the final year of the program, implementation of a few of the components improved but, in general, the overall level of the program declined.

In summary, the program was never implemented according to the initial intentions of the team that developed it. The outcome evaluation examined change in measures of problem behavior and antisocial attitudes, positive school adjustment, and school attendance. Not surprisingly, the results showed no reductions on any of these variables.

SOURCE: Adapted from Denise C. Gottfredson, Carolyn M. Fink, Stacy Skroban, and Gary D. Gottfredson, "Making Prevention Work," in *Establishing Preventive Services,* eds. R. P. Weissberg, T. P. Gullotta, R. L. Hampton, B. A. Ryan, and G. R. Adams (Thousand Oaks, CA: Sage, 1997), pp. 219-252.

implemented, or implementation is incomplete or weak, we would not expect it to be very successful in producing the intended outcomes, either the most immediate proximal outcomes or the ultimate outcomes to which it aspires. Evaluators gather information on this aspect of program performance through assessments of program process, including attention to both service utilization and program organizational issues.

Programs can also fail when the intended program activities are implemented but those activities do not actually have the intended effects. In the example of the mass media campaign on AIDS risk, the campaign may be implemented just as planned but may not be widely noticed and, therefore, not result in any heightened awareness or greater knowledge of AIDS risk. This is one form of theory failure, in particular, a failure of the action hypothesis—the program services do not bring about the immediate outcomes that are expected.

Another form of theory failure involves the conceptual hypothesis. This form of theory failure occurs when the program implements the intended services and, indeed, achieves the expected proximal outcomes, but those, in turn, do not lead to the expected distal outcomes. Thus, a mass media campaign may be hugely successful in raising awareness and knowledge about AIDS risk and how to reduce it, but people may not translate that knowledge into changed sexual behavior (one of the distal outcomes expected), and consequently, there will be no reduction in the incidence of AIDS (the social benefit the program ultimately hopes to produce). Exhibit 3-L provides an example of theory failure.

Although it simplifies considerably, we might liken the causal sequence embodied in program impact theory to the assumption that flipping a switch turns on a light. If we analyze this situation closely, it first requires moving one's hand to properly manipulate the switch, then having the switch activate a flow of current that causes the light to illuminate. Our assumptions about the relationship between moving our hand and the position of the switch are the action hypothesis. We can fail to change the position of the switch because we do not move our hands at all, or not the way we intended (implementation failure), or because we move them exactly the way we intend but somehow those moves are not successful in flipping the switch (failure of the action hypothesis). If the switch is flipped, our assumption that the light will come on represents our conceptual hypothesis. We can quite successfully manipulate the switch and still get no light if that part of the impact theory is in error, for example, the circuit is broken or the switch is not hooked up to a circuit (failure of the conceptual hypothesis).

The concept of program impact theory, the distinctions between proximal and distal outcomes, and the related distinctions between the program's action and conceptual hypotheses, therefore, can alert the evaluator to different aspects of program performance that may be appropriate to assess. It is for this reason that articulation of program impact theory during the planning stage of an evaluation is an important form of analysis for the evaluator to undertake. That exercise almost always yields very relevant evaluation questions regarding whether key program actions were implemented as intended and, if so, whether they produced the expected effects. Of course, the evaluator must also identify those key program activities in some detail so that appropriate questions can be raised about program implementation. This is where program process theory, encompassing service utilization and program organization, can be helpful.

EXHIBIT 3-L Theory Failure: A Children's Mental Health Demonstration Project

Mental health services for children are often underfunded, fragmented, and limited in variety. The five-year, $80 million Fort Bragg Demonstration Project was designed to test an innovative alternative to traditional mental health systems for children. Developed around the concept of a "continuum of care," the Demonstration Project was organized to deliver needed services on an individualized basis at all levels of severity using case management and interdisciplinary treatment teams to integrate and coordinate care. This variant of managed care was expected to result in improved treatment outcomes and lower cost of care per client.

The Demonstration Project was set up for the 42,000 military dependents under age 18 in the vicinity of the Fort Bragg military base in North Carolina. For evaluation purposes, two comparison sites were selected—Fort Campbell, Kentucky, and Fort Stewart, Georgia. Dependent children in those areas received mental health care under a conventional health insurance plan in which parents used independent practitioners or agencies and were reimbursed, subject to deductibles, by CHAMPUS, the military insurance provider.

The evaluators identified critical implementation and outcome issues with the aid of a carefully developed program theory description. To assess implementation, the program-as-implemented was compared with the program-as-planned. The results showed that, as intended, the demonstration had implemented a single point of entry to services for the target population, provided a comprehensive range of services, and established case management and treatment teams to coordinate services. Moreover, relative to the comparison sites, the services the children received in the Demonstration Project began sooner, were more individualized, had more variety, lasted longer with fewer dropouts, showed greater continuity and more parent involvement, and represented better matches between treatment and needs as judged by parents. The conclusion of the evaluators was that "the Demonstration was executed with sufficient fidelity to provide an excellent test of the program theory—the continuum of care."

The impact evaluation examined parents' satisfaction, treatment costs, and mental health outcomes with the following results:

- Parents were more satisfied with the services from the Demonstration Project than in the comparison sites.
- The costs per treated child were substantially higher in the Demonstration Project, not lower as expected.
- Mental health data collected on 984 children and families within 30 days after entry into the system and in two follow-up waves six months apart showed essentially no differences in clinical outcomes between the Demonstration and comparison sites. Of 116 distinct comparisons representing general and individualized measures reported by children, parents, therapists, and trained raters, 101 show no significant difference, 7 favored the comparison, and 8 favored the Demonstration.

In short, the continuum of care concept was well implemented but did not produce the effects that were expected on the basis of the program theory. Or, as the evaluators put it, "Commonly accepted wisdom about what is a better quality system of care is called into question."

SOURCE: Adapted from Leonard Bickman, "Implications of a Children's Mental Health Managed Care Demonstration Evaluation," *Journal of Mental Health Administration*, 1996, 23(1):107-118.

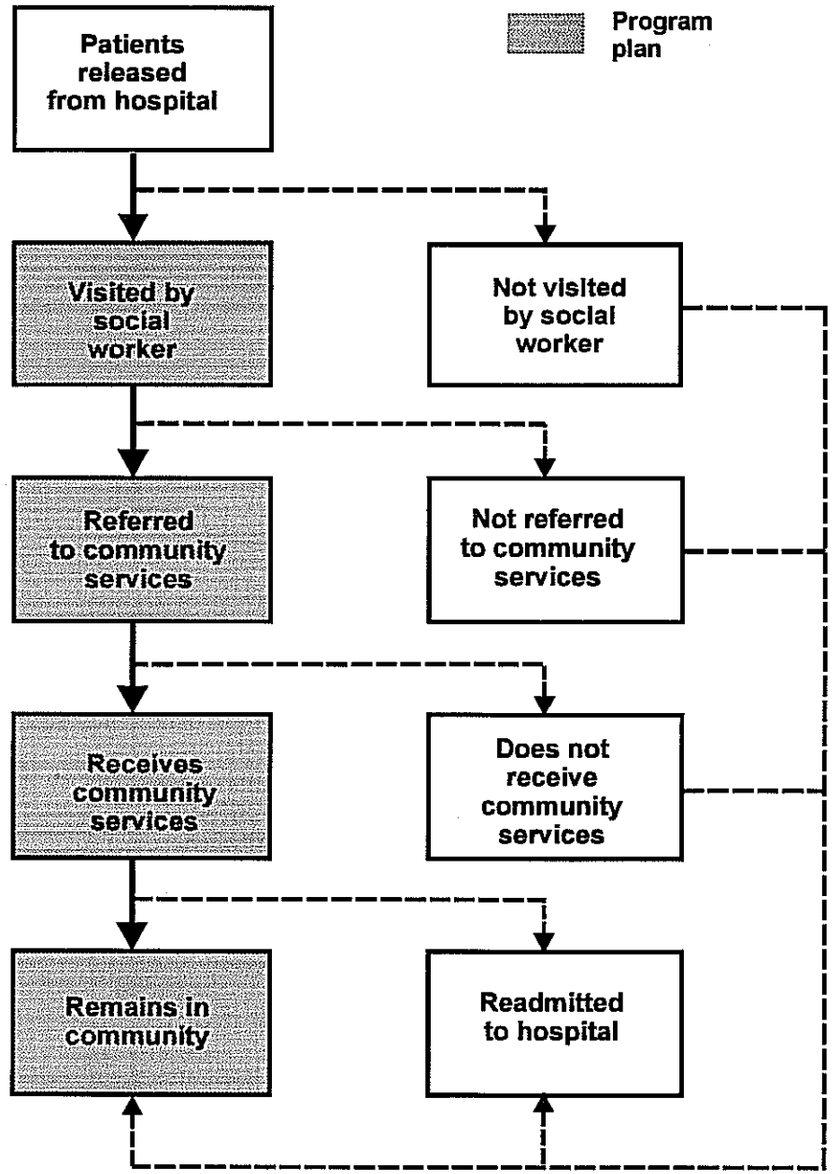
The Program Service Utilization Plan

Whereas program impact theory describes the chain of events leading from program-target transactions to the intended improvements in social conditions, the service utilization plan describes the sequence of events through which clients engage in those transactions. The service utilization plan is the set of assumptions and expectations about how the targets will make initial contact with the program and be engaged with it through the completion of the intended services. Its distinctive theme is that it describes the program-target transaction from the perspective of the targets and their experience and history of engagement with the program. An explicit, even if relatively informal, service utilization plan pulls into focus the critical assumptions about how and why the intended recipients of service will actually become engaged with the program and follow through to the point of receiving sufficient services to initiate the change process represented in the program impact theory. In the example of a mass media campaign to reduce AIDS risk, the service utilization plan would describe how persons at risk for AIDS will encounter the communications disseminated by the program and engage them sufficiently for their message to be received. Or, for another example, the service utilization plan for a neighborhood afterschool program for latchkey children would describe how parents are expected to learn of the program and enroll their children as well as how the children are expected to get to the program regularly and return home again afterward.

A program's service utilization plan can be usefully depicted in a flowchart that tracks the various paths program targets can follow from some appropriate point prior to first program contact through a point where there is no longer any contact. Exhibit 3-M shows an example of a simple service utilization flowchart for a hypothetical aftercare program for released psychiatric patients. One of the desirable features of such charts is the identification of possible situations in which the program targets are *not* engaged with the program as intended. For instance, for the community aftercare program in Exhibit 3-M, we see that formerly hospitalized psychiatric patients in the target population may not receive the planned visit from a social worker or referrals to community agencies and, as a consequence, may receive no service at all. The size of this group will be a function of how vigorously the program contacts potentially eligible cases and establishes case management for them. Similarly, a service utilization flowchart can highlight such issues as insufficient referrals from gateway agencies, program dropouts, early terminations prior to receiving the full-service package intended, and other such issues of incomplete service or service not offered or not delivered. Of course, at the same time, it portrays the pattern of outreach, intake, receipt of service, and exit from service that represents the program's scheme for making services available to the targets.

As a tool for program analysis and formulation of evaluation questions, articulating the service utilization plan contributes an important perspective on how the program is designed and what assumptions are made about the ways in which the target population is expected to engage the program services. That perspective facilitates the identification of important questions of program performance related to whether the appropriate target population is being served and in sufficient numbers,

EXHIBIT 3-M Service Utilization Flowchart for an Aftercare Program for Formerly Hospitalized Psychiatric Patients



what barriers there may be to entry into the program, the extent to which full and appropriate service is completed by an acceptable proportion of those beginning service, and whether desirable follow-up contact is made following service completion. An evaluator who has made the effort to explicate the program's service utilization plan and analyze its implications for program performance will be able to raise many important issues for consideration in developing the questions around which the evaluation will be designed.

The Program's Organizational Plan

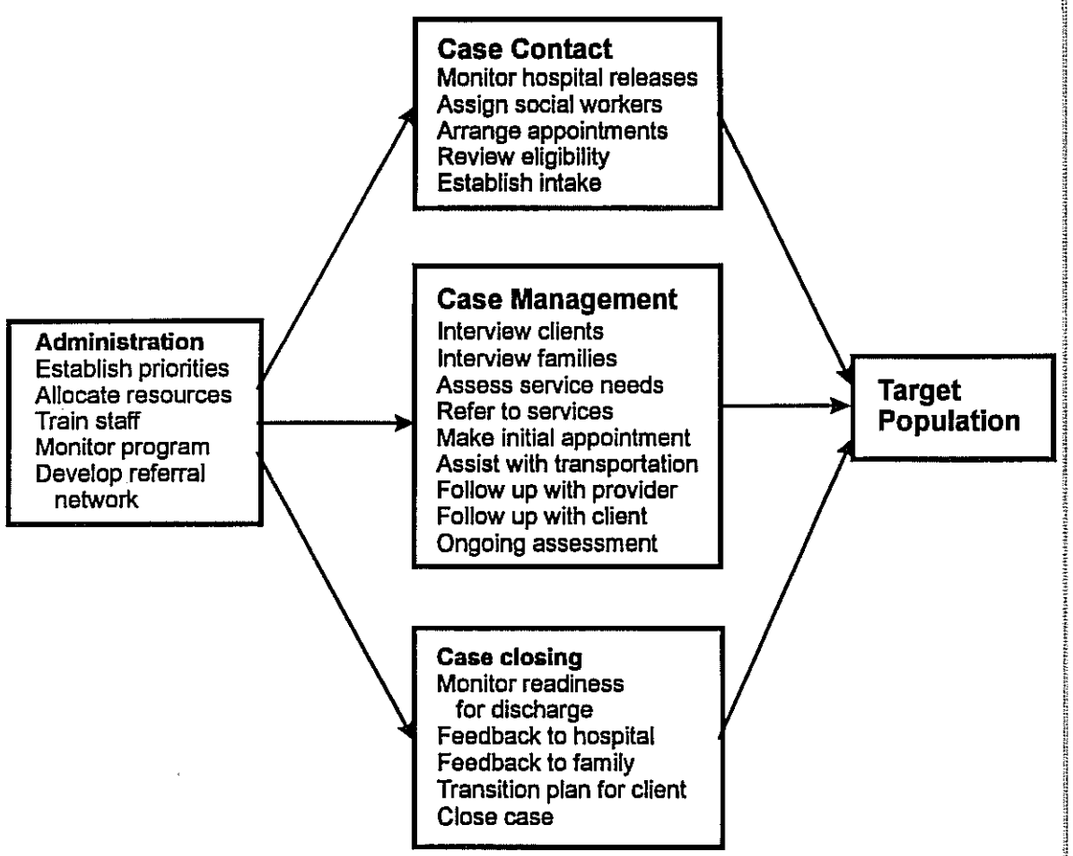
The organizational arrangements and program operations that constitute the routine functions of the program are based on a set of beliefs, assumptions, and expectations about what the program must do to bring about the intended target-program service transactions. These beliefs, assumptions, and expectations are what we call the program's organizational plan. The organizational plan is articulated from the perspective of program management and encompasses both the functions and activities the program is expected to perform and the human, financial, and physical resources required for that performance. Central to this scheme are the program services, those specific activities that constitute the program's role in the target-program transactions expected to lead to social benefits. However, it also must include those functions that provide essential preconditions and ongoing support to the organization's ability to provide its primary services, for instance, fund-raising, personnel management, facilities acquisition and maintenance, political liaison, and the like.

There are many ways the evaluator may depict the program's organizational plan. To be consistent with the schemes we have presented for impact theory and the service utilization plan, it is desirable to adopt a form that centers on those target-program service transactions that constitute the major points of contact between the program and the target population. The first element of the organizational plan, then, will be a description of the program's objectives with regard to the particular services it will provide: what those services are, how much is to be provided, to whom, and on what schedule. The next element of the organizational plan might then describe the resources and prior functions necessary to engage in those critical service activities. For instance, sufficient personnel with appropriate credentials and skills will be required as will logistical support, proper facilities and equipment, funding, supervision, clerical support, and so forth.

As with the other portions of program theory, it is often useful to describe a program's organizational plan with a chart or diagram. Exhibit 3-N presents an example in that form that describes the major organizational components of the hypothetical aftercare program for psychiatric patients whose service utilization scheme is depicted in Exhibit 3-M. A rather common way of depicting the organizational plan of a program is in terms of *inputs*, representing the resources and constraints applicable to the program, and *activities*, indicating the services the program is expected to provide. When included in a full *logic model*, these schemes typically represent receipt of services (service utilization) as program *outputs*, which, in turn, are related to the desired outcomes. Exhibit 3-O shows such a scheme, drawn from a widely used workbook prepared by the United Way of America.

Naturally, a description of program organization and articulation of the underlying as-

EXHIBIT 3-N Organizational Schematic for an Aftercare Program for Formerly Hospitalized Psychiatric Patients



sumptions, expectations, and intentions reveal many aspects of the program related to how effectively it performs its various functions. Thus, the evaluator can use the organizational plan to generate evaluation questions that may be relevant to the design and planning of the evaluation. One major category of issues, for instance, relates to whether the program is actually implementing the functions and activities that are intended and in the intended way. Questions on this topic can be rather general or, more often, quite specific and detailed, for instance, whether the case managers are contacting the teachers about the child's

EXHIBIT 3-O A Logic Model for a Teen Mother Parenting Education Program

Inputs	*Activities*	*Outputs*	*Outcomes*		
			Initial	*Intermediate*	*Longer Term*
Agency provides MSW program manager, part-time RN instructor, nationally certified education manuals, videos, and other teaching tools.	Program provides parenting classes on prenatal-through-infant nutrition, development, safety, and caretaking delivered in high schools twice a week for one hour to teen mothers from three months prior, to one year after, delivery of a child.	Pregnant teens attend program.	Teens are knowledgeable about prenatal nutrition and health guidelines.	Teens follow proper nutrition and health guidelines. Teens deliver healthy babies.	Babies achieve appropriate 12-month milestones for physical, motor, verbal, and social development.
Agency and high school identify pregnant teens to participate in program.			Teens are knowledgeable about proper care, feeding, and social interaction with infants.	Teens provide proper care, feeding, and social interaction to their babies.	

SOURCE: Adaped from United Way of America, *Measuring Program Outcomes: A Practical Approach.* Alexandria, VA: Author, 1996, p. 42. Used by permission, United Way of America.

schoolwork for every child in every family assigned to their caseload. A related question is whether those functions and activities are optimally configured for their purposes, represent appropriate standards of professional practice, are adequately supervised, and so forth.

Another set of important organizational questions relates to program resources. An evaluation may encompass questions about whether the personnel are sufficient in numbers or credentials for their assigned tasks, the adequacy of program funding, availability of the requisite facilities and equipment, and many other such matters. Still another category of organizationally important issues has to do with the administration, management, and governance of the program. Questions might be raised about specific managerial functions or the effectiveness of overall program administration. Similarly, the nature and relationship of governing and advisory boards may be of interest as well as relations with other agencies, superordinate organizational structures, political entities, and so on.

How to Articulate Program Theory

Program theory in the detail indicated above is rarely written down in the documentation the program has on hand for the evaluator to examine, although the broad outlines will likely appear in statements of the program's mission and services or in funding pro-

posals and annual reports. Typically, then, the evaluator must articulate the program theory on the basis of an initial investigation of the program being evaluated. Once articulated in a form endorsed by key program stakeholders, the program theory can itself be an object of evaluation. That is, an important evaluation question may relate to how well conceptualized the program is, and the evaluator may conduct an explicit assessment of the program theory. Chapter 5 of this volume describes in some detail how program theory can be evaluated. Because program theory must first be articulated before it can be evaluated, Chapter 5 also describes the ways the evaluator can reveal and express program theory. When program theory is formulated for the purpose of analyzing a program to identify pertinent evaluation questions, as discussed in this chapter, the same procedures are applicable. Because a full discussion is provided in Chapter 5, we will mention only a few general points here.

It is, for instance, important to recognize that articulation of program theory should be mainly a process of discovery and not one of invention. The evaluator is rarely the authoritative voice with regard to how the program is expected to work. The understandings of those persons who originate, plan, administer, and staff a social program are the primary source of information on this matter. The evaluator, of course, may play a large and creative role in interpreting and organizing that information. Moreover, few programs are so unique that they bear no resemblance to at least some other programs whose funders, administrators, staff, and so forth can be consulted by the evaluator for additional perspectives on how such programs should work. There may also be pertinent information available from professional and research literature about that type of program and sources of expertise and experience among the members of the professions involved, for example, social work, nursing, psychiatry, or teaching.

The greatest difficulty the evaluator will encounter is that the various components of program theory often are implicit rather than explicit and may be in the form of tacit knowledge that is so routinized in the program context that it is rarely thought about or discussed. The evaluator attempting to describe the program theory, therefore, must generally draw it out piecemeal from program informants, available documents, and the professional and research literature, and then attempt to synthesize the pieces into a coherent whole. This exercise must involve considerable interaction with program stakeholders, especially administrators, who should provide both critical input and feedback on each iteration the evaluator produces. A useful way to approach this task is to draw figures and charts such as those shown in Exhibits 3-I, 3-M, 3-N, and 3-O for the various components of program theory, then go over them in detail with program informants to obtain feedback for refinement.

It is wise to avoid evaluation jargon in this task. Most program administrators will have little notion of what is meant by "program theory" if asked outright and are likely to assume it means something more formal and abstract than it does in this context. On the other hand, inquiries about how the program works, what various personnel do and why, and other such questions at a practical level generally lead to fruitful and often lengthy discussions that can be very informative. It does sometimes happen, nonetheless, that the effort to explicate program theory will reveal that there are important areas of the program conceptualization that are vague, undetermined, or inconsistent. In such cases, it may be appropriate for the evaluation itself to encompass a systematic assessment of the program theory aimed at identifying weaknesses and assisting program personnel in clarifying and refining their understanding of what the program should be doing and why (Chapter 5, on assessing program theory, describes how this might be done).

COLLATING EVALUATION QUESTIONS AND SETTING PRIORITIES

The evaluator who thoroughly explores stakeholder concerns and conducts an analysis of program issues guided by carefully developed descriptions of program theory will turn up many questions that the evaluation might address. The task at this point becomes one of organizing those questions according to distinct themes and setting priorities among them.

Organization is generally rather straightforward. Evaluation questions tend to cluster around different program functions (e.g., recruitment, services, outcomes) and, as noted earlier, around different evaluation issues (need, design, implementation, impact, efficiency). In addition, evaluation questions tend to show a natural hierarchical structure with many very specific questions (e.g., "Are elderly homebound persons in the public housing project aware of the program?") nested under broader questions ("Are we reaching our target population?").

Setting priorities to determine which questions the evaluation should be designed to answer can be much more challenging. Once articulated, most of the questions about the program that arise during the planning process are likely to seem interesting to some stakeholder or another, or to the evaluators themselves. Rarely will resources be available to address them all, however. At this juncture, it is especially important for the evaluator to focus on the purpose of the evaluation and the expected uses to be made of its findings. There is little point to investing time and effort in developing information that is of little use to any stakeholder.

That said, we must caution against an overly narrow interpretation of what information is useful. Evaluation utilization studies have shown that practical, instrumental use, for example, for program decision making, is only one of the contributions evaluation information makes (Leviton and Hughes, 1981; Rich, 1977; Weiss, 1988). Equally important in many cases are conceptual and persuasive uses—the contribution of evaluation findings to the way in which a program and the social problems to which it responds are understood and debated. Evaluations often identify issues, frame analysis, and sharpen the focus of discussion in ways that are influential to the decision-making process even when there is no direct connection evident between any evaluation finding and any specific program decision. A fuller discussion of this issue is presented in Chapter 12; our purpose here is only to point out the possibility that some evaluation questions may be important to answer even though no immediate use or user is evident.

With the priority evaluation questions for a program decided on through some reasonable process, the evaluator is ready to design that substantial part of the evaluation that will be

devoted to trying to answer them. Most of the remainder of this book discusses the approaches, methods, and considerations related to that task. That discussion is organized to follow the natural logical progression of evaluation questions and thus addresses, in turn, how to assess the need for a program, the program theory or plan for addressing that need, the implementation of the program plan and the associated program process, the impact or outcome of the program implementation on the social need, and the efficiency with which the program attains its outcomes.

SUMMARY

- A critical phase in evaluation planning is the identification and formulation of the questions the evaluation will address. Those questions focus the evaluation on the areas of program performance most at issue for key stakeholders and guide the design so that it that will provide meaningful information about program performance. Good evaluation questions, therefore, must identify clear, observable dimensions of program performance that are relevant to the program's goals and represent domains in which the program can realistically be expected to have accomplishments.
- What most distinguishes evaluation questions, however, is that they involve criteria by which the identified dimensions of program performance can be judged. If the formulation of the evaluation questions can include performance standards on which key stakeholders agree, evaluation planning will be easier and the potential for disagreement over the interpretation of the results will be reduced.
- To ensure that the matters of greatest significance are covered in the evaluation design, the evaluation questions are best formulated through interaction and negotiation with the evaluation sponsors and other stakeholders representative of significant groups or distinctly positioned in relation to program decision making.
- Although stakeholder input is critical, the evaluator must also be prepared to identify program issues that might warrant inquiry. This requires that the evaluator conduct a somewhat independent analysis of the assumptions and expectations on which the program is based.
- One useful way to reveal aspects of program performance that may be important to assess in an evaluation is to make the program theory explicit. Program theory describes the assumptions inherent in a program about the activities it undertakes and how those relate to the social benefits it is expected to produce. It encompasses impact theory, which links program actions to the intended outcomes, and process theory, which describes a program's organizational plan and scheme for ensuring utilization of its services by the target population.

- When these various procedures have generated a full set of candidate evaluation questions, the evaluator must organize them into related clusters and draw on stakeholder input and professional judgment to set priorities among them. With the priority evaluation questions for a program determined, the evaluator is then ready to design the part of the evaluation that will be devoted to answering them.

KEY CONCEPTS FOR CHAPTER 4

Needs assessment	An evaluative study that answers questions about the social conditions a program is intended to address and the need for the program. Needs assessment may also be used to determine whether there is a need for a new program and to compare or prioritize needs within and across program areas.
Key informants	Persons whose personal or professional position gives them a perspective on the nature and scope of a social problem or a target population and whose views are obtained during a needs assessment.
Survey	Systematic collection of information from a defined population, usually by means of interviews or questionnaires administered to a sample of units in the population.
Focus group	A small panel of persons selected for their knowledge or perspective on a topic of interest that is convened to discuss the topic with the assistance of a facilitator. The discussion is usually recorded and used to identify important themes or to construct descriptive summaries of views and experiences on the focal topic.
Social indicator	Periodic measurements designed to track the course of a social condition over time.
Incidence	The number of new cases of a particular problem or condition that arise in a specified area during a specified period of time.
Prevalence	The number of existing cases with a particular condition in a specified area at a specified time.
Population at risk	The individuals or units in a specified area with characteristics judged to indicate that they have a significant probability of having or developing a particular condition.
Population in need	The individuals or units in a specified area that currently manifest a particular problematic condition.
Sensitivity	The extent to which the criteria used to identify a target population result in the inclusion of individuals or units that actually have or will develop the condition to which the program is directed.
Specificity	The extent to which the criteria used to identify the target population result in the exclusion of individuals or units who do not have or will not develop the condition to which the program is directed.
Rate	The occurrence or existence of a particular condition expressed as a proportion of units in the relevant population (e.g., deaths per 1,000 adults).

CHAPTER 4

ASSESSING THE NEED FOR A PROGRAM

Previous chapters provided an overview of evaluation and an orientation to the critical themes in tailoring an evaluation to program circumstances and formulating the specific questions an evaluation will be designed to answer. Beginning with this chapter, we turn to fuller discussion of the various methods and approaches evaluators use to address different categories of evaluation questions.

The category of evaluation questions that is logically most fundamental to program evaluation has to do with the nature of the social problem the program is expected to ameliorate and the needs of the population experiencing that problem. These questions follow from the assumption that the purpose of social programs is to bring about improvement in problematic social conditions and that they are accountable to those who fund and support them for making a good faith effort to do so.

Needs assessment, in general, is a systematic approach to identifying social problems, determining their extent, and accurately defining the target population to be served and the nature of their service needs. From a program evaluation perspective, needs assessment is the means by which an evaluator determines if, indeed, there is a need for a program and, if so, what program services are most appropriate to that need. Such an assessment is critical to the effective design of new programs. However, it is equally relevant to established programs because there are many circumstances in which it cannot merely be assumed that the program is needed or that the services it provides are well suited to the nature of the need.

What makes the assessment of the need for a program so fundamental, of course, is that a program cannot be effective at ameliorating a social problem if there is no problem to begin with or if the program services do not actually relate to the problem. The concepts and procedures an evaluator can use to conduct this critical investigation of the nature and extent of the need for a program are discussed in this chapter.

As we described in Chapter 1, a fundamental premise of program evaluation within the human service domain is that effective programs are instruments for improving social conditions. Indeed, bringing about such improvement is the primary mission and reason

for being of social programs (which is not to say that they are not also influenced by other political and organizational imperatives). Whether a program addresses a significant social need in a plausible way and does so in a manner that is responsive to the circumstances of those in need, therefore, are essential questions for evaluating the effectiveness of a social program.

Answering these questions for a given program first requires a description of the social problem the program intends to ameliorate. With that description in hand, the evaluator can ask if the program theory embodies a valid conceptualization of the problem and an appropriate means of remedying it. If that question is answered in the affirmative, attention can turn to whether the program is actually implemented in line with the program theory and, if so, whether the intended improvements in the social conditions actually result and at what cost. Thus, the logic of program evaluation builds upward from careful description of the social problem the program is expected to ameliorate.

Thorny issues in this domain revolve around deciding just what is meant by a need in contrast, say, to a want or desire, and what ideals or expectations should provide the benchmarks for distinguishing a need (cf. McKillip, 1998; Scriven, 1991). We will not attempt to resolve these issues here, if indeed they can be resolved, but will be content with the notion that a need is a social construction negotiated between a set of social agents with responsibility for social programs and policy and a set of claimants and their advocates who assert that a problem exists that warrants intervention.

The family of procedures used by evaluators and other social researchers to systematically describe and diagnose social needs is generally referred to as *needs assessment*. Its purpose is to determine if there is a need or problem and, if so, what its nature, depth, and scope are. In addition, needs assessment often encompasses the process of comparing and prioritizing needs according to how serious, neglected, or salient they are.

Within the context of program evaluation, however, the primary focus of needs assessment is not on human needs broadly defined but, rather, on social conditions deemed unsatisfactory through some process of social judgment and presumed remediable by social programs. The essential tasks for the program evaluator as needs assessor are to identify the decisionmakers and claimants who constitute the primary stakeholders in the program domain of interest, describe the "problem" that concerns them in a manner that is as careful, objective, and meaningful to both groups as possible, and help draw out the implications of that diagnosis for structuring effective intervention, whether through new or ongoing programs.

THE ROLE OF EVALUATORS IN DIAGNOSING SOCIAL CONDITIONS AND SERVICE NEEDS

In the grand scheme of things, evaluators' contributions to the identification and alleviation of social problems are modest compared with the weightier actions of political bodies, advocacy groups, investigative reporters, and sundry charismatic figures. The impetus for attending to social problems most often comes from political and moral leaders and community advocates who have a stake, either personally or professionally, in dealing with a particular condition. Thus, the post-World War II attention to mental illness was heavily influenced by the efforts of a single congressman; federal programs for mental retardation received a major boost during John F. Kennedy's presidency because he had a sibling with mental retardation; improved automobile safety can be credited to a considerable degree to Ralph Nader's advocacy; and efforts to control illegal and improper delivery of health and welfare services have most often come about because of exposés in the mass media and the activities of interest and pressure groups, including the organized efforts of those in need themselves.

Nevertheless, evaluators do contribute significantly to efforts to improve the human and social condition, though not by mobilizing the disaffected, storming the barricades, or shooting from the lip. Rather, they contribute in mundane but essential ways by applying their repertoire of research techniques to systematically describe the nature of social problems, gauge the appropriateness of proposed and established intervention programs, and assess the effectiveness of those programs for improving social conditions.

This chapter focuses on the role of evaluators in diagnosing social problems through systematic and reproducible procedures in ways that can be related to the design and evaluation of intervention programs. The importance of the resulting diagnostic information cannot be overstated. Speculation, impressionistic observations, political pressure, and even deliberately biased information may spur policymakers, planners, and funding organizations to initiate action, support ongoing programs, or withdraw support from programs. But if sound judgment is to be reached about such matters, it is essential to have an adequate understanding of the nature and scope of the problem the program is meant to address as well as precise information about the corresponding program targets and the context in which the intervention operates or will operate. Here are a few examples of what can happen when adequate diagnostic procedures are ignored:

- The problem of high unemployment rates in inner-city neighborhoods frequently has been defined as reflecting the paucity of employment opportunities in those neighborhoods. Programs have therefore been established that provided substantial incentives to businesses for locating in inner-city neighborhoods. Subsequent experiences often found that most of the workers these businesses hired came from outside the neighborhood that was supposed to be helped.

- After a social intervention designed to prevent criminal behavior by adolescents was put in place in a Midwestern suburb, it was discovered that there was a very low rate of juvenile crime in the community. The program planners had assumed that because juvenile delinquency was a serious problem nationally, it was a problem in their community as well.

- Planners of many of the urban renewal projects undertaken during the 1960s assumed that persons living in what the planners regarded as dilapidated buildings also viewed their housing as defective and would therefore support the demolition of their homes and accept relocation to replacement housing. In city after city, however, residents of urban renewal areas vigorously opposed these projects.

- Media programs designed to encourage people to seek physical examinations to detect early signs of cancer had the effect of swamping

health centers with more clients than they could handle. The media effort stimulated many hypochondriacal persons without cancer symptoms to believe they were experiencing warning signs.

• In an effort to improve the clinical identification of AIDS, community physicians were provided with literature about the details of diagnosing the syndrome among high-risk patients using blood tests. Only after the materials had been disseminated was it recognized that few physicians take sex histories as a routine practice and thus they were unlikely to know which of their patients were high risk. Consequently, the only way they could make use of their new knowledge was by testing all their patients. The result was an excessive amount of testing, at high cost and some risk to the patients.

• A birth control project was expanded to reduce the reportedly high rate of abortion in a large urban center, but the program failed to attract many additional participants. Subsequently, it was found that most of the intended urban clients were already being adequately served and a high proportion practiced contraception. The high abortion rate was caused mainly by young women who came to the city from rural areas to have abortions.

• The problem of criminal use of handguns has led to legislative proposals to forbid the sale of such guns to persons convicted of felony offenses. However, most criminals do not purchase their guns from legitimate gun dealers, nor do dealers have reliable ways of ascertaining whether purchasers have criminal records.

In all of these examples, a good needs assessment would have provided information leading to a valid description of the problem that would have prevented programs from implementing inappropriate or unneeded services. In some cases, unnecessary programs were designed because the problem did not exist. In others, the intervention was not effective because the target population did not desire the services provided, was incorrectly identified, or was unlikely or unable to act in the way the program expected.

All social programs rest on a set of assumptions and representations of the nature of the problem they address and the characteristics, needs, and responses of the target population they intend to serve. Any evaluation of a plan for a new program, a change in an existing program, or the effectiveness of an ongoing program must necessarily engage those assumptions and representations. Of course, the problem diagnosis and target population description may already be well and convincingly established, in which case the evaluator can move forward with that as a given. Or the nature of the evaluation task may be stipulated in such a way that the need for the program and the nature of that need is not a matter for independent investigation. Indeed, program personnel and sponsors often believe they know the social problems and target population needs so well that further inquiry is a waste of time. Such situations must be approached cautiously. As the examples above show, it is remarkably easy for a program to be based on faulty assumptions, either through insufficient initial problem diagnosis, changes in the problem or target population since the program was initiated, or selective exposure or stereotypes that lead to distorted views.

In all instances, therefore, the evaluator should scrutinize the assumptions about the target problem and population that shape the nature of a program. Where there is any ambiguity, it may be advisable for the evaluator to work with key stakeholders to formulate those assumptions explicitly so that they may serve as touchstones for assessing the adequacy of the program design and theory. Often it will also be useful for the evaluator to conduct at least some minimal independent investigation of the nature of the program's target problem and population. For new program initiatives, or established programs whose utility has been called into question, it may be essential to conduct a thorough assessment of the social need and target population to be served by the program at issue. In other cases, a needs assessment may be virtually mandated. For example, the 1974 community mental health legislation called for periodic community mental health needs assessments, and the 1987 McKinney Act indicated that states and local communities should use needs assessments as the basis for planning programs for the homeless. The role of the federal government in fostering needs assessment for educational programs is described in Exhibit 4-A.

It should be noted that needs assessment is not always done with reference to a specific social program or program proposal. The techniques of needs assessment are also used as planning tools and decision aids for policymakers who must prioritize among competing

EXHIBIT 4-A The Rise and Fall of the Government's Role in Educational Needs Assessment

The widespread use of needs assessment (NA) as a systematic, rational means of determining goals and priorities for program planning and evaluation in the United States dates from 1965, with the passage of the Elementary and Secondary Education Act (ESEA, PL 89-10). In the next 15 years over 35 titles in the 54 largest grants-in-aid programs in health, education, and social services required applicants for categorical and competitive grants to document their needs.

Although satisfying granting agencies was not the only reason for needs assessment, the social legislation was a powerful stimulus to the development of models and the conduct of exemplary studies. The period of 1966-1981 was characterized by the conduct of large-scale needs assessments of whole systems, the dissemination of kits of materials and survey instruments (Witkin, 1977), the spread of NA to city planning and the private sector, and some development of theoretical perspectives.

With the passage of the Omnibus Budget Reconciliation Act of October 1981, about 90 percent of the legislation that included mandates for NA was eliminated. In the following year, there was an abrupt drop in NAs, especially in local education agencies. Although applications for categorical grants such as ESEA Chapter 1 (compensatory education) still required evidence of need, those NAs often consisted merely of reporting demographic data and test scores of the students to be served.

SOURCE: Quoted, with permission, from Belle Ruth Witkin, "Needs Assessment Since 1981: The State of the Practice," *Evaluation Practice*, 1994, 15(1):17.

EXHIBIT 4-B Steps in Analyzing Need

1. *Identification of users and uses.* The users of the analysis are those who will act on the basis of the results and the audiences who may be affected by it. The involvement of both groups will usually facilitate the analysis and implementation of its recommendations. Knowing the uses of the need analysis helps the researcher focus on the problems and solutions that can be entertained, but also may limit the problems and solutions identified in Step 3, below.
2. *Description of the target population and service environment.* Geographic dispersion, transportation, demographic characteristics (including strengths) of the target population, eligibility restrictions, and service capacity are important. Social indicators are often used to describe the target population either directly or by projection. Resource inventories detailing services available can identify gaps in services and complementary and competing programs. Comparison of those who use services with the target population can reveal unmet needs or barriers to solution implementation.
3. *Need identification.* Here problems of the target population(s) and possible solutions are described. Usually, more than one source of information is used. Identification should include information on expectations for outcomes; on current outcomes; and on the efficacy, feasibility, and utilization of solutions. Social indicators, surveys, community forums, and direct observation are frequently used.
4. *Need assessment.* Once problems and solutions have been identified, this information is integrated to produce recommendations for action. Both quantitative and qualitative integration algorithms can be used. The more explicit and open the process, the greater the likelihood that results will be accepted and implemented.
5. *Communication.* Finally, the results of the need analysis must be communicated to decisionmakers, users, and other relevant audiences. The effort that goes into this communication should equal that given the other steps of the need analysis.

SOURCE: Adapted from Jack McKillip, "Need Analysis: Process and Techniques," in *Handbook of Applied Social Research Methods*, eds. L. Bickman and D. J. Rog (Thousand Oaks, CA: Sage, 1998), pp. 261-284.

needs and claims. For instance, a regional United Way or a metropolitan city council might commission a needs assessment to help them determine how funds should be allocated across various service areas. Or a state department of mental health might assess community needs for different mental health services to distribute resources optimally among its service units. Although different in scope and purpose from the assessment of the need for a particular program, whether existing or proposed, the methods for these broader needs assessments are much the same, and they also are generally conducted by evaluation researchers. Exhibit 4-B provides an overview of the general steps in a needs assessment. Useful book-length discussion of needs assessment applications and techniques can be found in McKillip (1987), Reviere et al. (1996), Soriano (1995), and Witkin and Altschuld (1995).

As the examples and commentary above indicate, needs assessment has a number of facets and applications relevant to program evaluation. The next sections discuss the evaluator's role in identifying social problems, analyzing their location and scope, defining the targets of proposed interventions, and describing the nature of the associated service needs.

DEFINING SOCIAL PROBLEMS

Proposals for policy changes, new or modified programs, or evaluation of existing programs generally arise out of the dissatisfaction of one or more groups of stakeholders with the effectiveness of existing policies and programs or realization that a new social problem is emerging. Either case assumes that a social problem has been identified, a matter that is not as straightforward as it may seem. Indeed, the question of what defines a social problem has occupied spiritual leaders, philosophers, and social scientists for centuries. For our purposes, the key point is that social problems are not themselves objective phenomena. Rather, they are social constructions that emerge from the interests of the parties involved as they relate to observed conditions. In this sense, community members, together with the stakeholders involved in a particular issue, literally create the social reality that constitutes a recognized social problem (Miller and Holstein, 1993; Spector and Kitsuse, 1977).

It is generally agreed, for example, that poverty is a social problem. The observable facts are the statistics on the distribution of income and assets. However, those statistics do not define poverty; they merely permit one to determine how many are poor when a definition is given. Nor do they establish poverty as a social problem; they only characterize a situation that individuals and social agents may view as problematic. Moreover, both the definition of poverty and the goals of programs to improve the lot of the poor vary over time, between communities, and among stakeholders. Initiatives to reduce poverty, therefore, may range from increasing employment opportunities and reducing barriers to economic mobility to simply lowering the expectations of those persons with low income.

Defining a social problem and specifying the goals of intervention are thus ultimately political processes that do not follow simply from the inherent characteristics of the situation. This circumstance is illustrated nicely, for instance, in an analysis of legislation designed to reduce adolescent pregnancy that was conduced by the U.S. General Accounting Office (GAO, 1986). The GAO found that none of the pending legislative proposals defined the problem as involving the fathers of the children in question; every one addressed adolescent pregnancy as an issue of young mothers. Although this view of the problem of adolescent pregnancy may lead to effective programs, clearly there are alternative definitions that include the adolescent fathers.

Indeed, the social definition of a problem is so central to the political response that the preamble to proposed legislation usually shows some effort to specify the conditions for which the proposal is designed as a remedy. For example, two contending legislative proposals may both be addressed to the issue of homeless persons, but one may identify the homeless as needy persons who have no kin on whom to be dependent, whereas the other defines homelessness as the lack of access to conventional shelter. The first definition centers attention

primarily on the social isolation of potential clients; the second focuses on housing arrangements. The ameliorative actions that are justified in terms of these definitions will likely be different as well. The first definition, for instance, would support programs that attempt to reconcile homeless persons with alienated relatives; the second, subsidized housing programs.

It is usually informative, therefore, for an evaluator to determine what the problem a program addresses is thought to be in its particular political context. To investigate this, the evaluator might, for instance, study the implicit or explicit definitions that appear in policy and program proposals. Revealing information may also be found in legislative proceedings, including committee hearings and floor debates, journals of opinion, newspaper and magazine editorials, and other sources in which discussions of the problem appear. The operative definition of the problem a particular program addresses can usually be found in program documents, newspaper accounts of its launch, proposals for funding it, and the like. Such materials may explicitly describe the nature of the problem and the program's plan of attack, as in funding proposals, or implicitly define the problem through the assumptions that underlie statements about program activities, successes, and plans.

This inquiry will almost certainly turn up information that will be useful for a preliminary description of the social need to which the program is presumably designed to respond. As such, it can guide a more probing needs assessment, both with regard to how the problem is defined and what alternative perspectives might be applicable. If the evaluation circumstances do not permit further systematic investigation, this information can nonetheless be the basis for a thoroughgoing discussion with stakeholders about their perceptions and assumptions about the nature of the social conditions the program addresses. This, then, provides the evaluator with some basis for analyzing the structure and goals of the program and assessing its design.

Also, an important role evaluators may play at this stage is to provide policymakers and program managers with a critique of the problem definition inherent in their policies and programs and propose alternative definitions that may be more serviceable. For example, evaluators could point out that a definition of the problem of teenage pregnancies as primarily one of illegitimate births ignores the large number of births that occur to married teenagers and suggest program implications that follow from that definition.

SPECIFYING THE EXTENT OF THE PROBLEM: WHEN, WHERE, AND HOW BIG?

The design and funding of a social program should be geared to the size, distribution, and density of the problem it addresses. In assessing, say, emergency shelters for homeless persons in a community, it makes a very significant difference whether the total homeless population is 350 or 3,500. It also makes a big difference whether the problem is located primarily in poor neighborhoods or affluent ones and how many of the homeless suffer from mental illness, chronic alcoholism, and physical disabilities.

It is much easier to establish that a problem exists than to develop valid estimates of its density and distribution. Identifying a handful of battered children may be enough to convince

a skeptic that child abuse exists. But specifying the size of the problem and where it is located geographically and socially requires detailed knowledge about the population of abused children, the characteristics of the perpetrators, and the distribution of the problem throughout the political jurisdiction in question. For a problem like child abuse, which is not generally public behavior, this can be difficult. Such social problems are mostly "invisible," so that only imprecise estimates of their rates of occurrence are possible. In such cases, it is often necessary to use data from several sources and use different approaches to estimating rates of occurrence (e.g., Ards, 1989).

It is also generally important to have at least reasonably representative samples to estimate rates of occurrence. It can be especially misleading to draw estimates from at-risk populations, such as those found in service programs, when general population estimates are needed to determine the extent of a problem. Estimation of the rate of spousal abuse during pregnancy based on women in shelters, for instance, results in considerable overestimation of the frequency of occurrence in the general population of pregnant women. An estimate from a more representative sample still indicates that battering of pregnant women is a serious problem, but places the extent of the problem in a realistic perspective (see Exhibit 4-C).

Using Existing Data Sources to Develop Estimates

Through their knowledge of existing research and data sources and their understanding of which designs and methods lead to conclusive results, evaluation researchers are in a good position to collate and assess whatever information already exists on a given social problem. Here we stress both *collate* and *assess*—unevaluated information can be as bad as no information at all.

For some social issues, existing data sources may be of sufficient quality to be used with confidence. For example, accurate and trustworthy information can usually be obtained about issues on which measures are routinely collected either by the Current Population Survey or the decennial U.S. Census. Moreover, through the census tract coding, that information can be disaggregated to state and local levels. As an illustration, Exhibit 4-D describes the use of vital statistics records and census data to assess the nature and magnitude of the problem of poor birth outcomes in a Florida county. This needs assessment was aimed at estimating child and maternal health needs so that appropriate services could be planned. Even when such direct information about the problem of interest is not available from existing records, indirect estimates may be possible if the empirical relationships between available information and problem indicators are known (e.g., Ciarlo et al., 1992).

In addition to the decennial census, data available in many of the statistical series routinely collected by federal agencies are often trustworthy. There are, unfortunately, exceptions. For example, it is widely acknowledged that the U.S. Census undercounts the numbers of African Americans and Hispanics and, to a considerable extent, the number of homeless. For the nation as a whole, these undercounts are relatively small and for many purposes can be ignored. For jurisdictions with large populations of African Americans and Hispanics, however, these undercounts may result in significant misestimation of the size of certain target populations relevant to assessing the need for social programs.

EXHIBIT 4-C Estimating the Frequency of Domestic Violence Against Pregnant Women

All women are at risk of battering; however, pregnancy places a woman at increased risk for severe injury and adverse health consequences, both for herself and her unborn infant. Local and exploratory studies have found as many as 40%-60% of battered women to have been abused during pregnancy. Among 542 women in a Dallas shelter, for example, 42% had been battered when pregnant. Most of the women reported that the violence became more acute during the pregnancy and the child's infancy. In another study, interviews of 270 battered women across the United States found that 44% had been abused during pregnancy.

But most reports on battering during pregnancy have been secured from samples of battered women, usually women in shelters. To establish the prevalence of battering during pregnancy in a representative obstetric population, McFarlane and associates randomly sampled and interviewed 290 healthy pregnant women from public and private clinics in a large metropolitan area with a population exceeding three million. The 290 black, white, and Latina women ranged in age from 18 to 43 years; most were married, and 80% were at least five months pregnant. Nine questions relating to abuse were asked of the women, for example, whether they were in a relationship with a male partner who had hit, slapped, kicked or otherwise physically hurt them during the current pregnancy and, if yes, had the abuse increased. Of the 290 women, 8% reported battering during the current pregnancy (one out of every twelve women interviewed). An additional 15% reported battering before the current pregnancy. The frequency of battering did not vary as a function of demographic variables.

SOURCE: Adapted from J. McFarlane, "Battering During Pregnancy: Tip of an Iceberg Revealed," *Women and Health*, 1989, 15(3):69-84.

Because many federal programs are tied to the size of particular populations, such undercounting can also translate into substantial losses of federal funds for those jurisdictions. This circumstance has led to a stream of lawsuits by cities and states, and to advocacy of statistical adjustments of the undercount. Although many statisticians regard such adjustments as a sound approach, adjustments have been rejected in the 1980 and 1990 censuses. In planning for the 2000 census, the Bureau of the Census has formulated a plan for statistical adjustments based on sampling nonrespondents. These plans are quite controversial, however, and Congress is considering legislation that would instruct the Bureau of the Census to conduct a "complete census" and abandon any plans to adjust census returns on the basis of sampling.

As this example illustrates, it is often difficult to separate technical decisions from political interests in diagnosing social problems. Adjusting for the undercount of African American and, particularly, Hispanic persons would increase the amount of resources allocated to them under many federal programs. At the

EXHIBIT 4-D Using Vital Statistics and Census Data to Assess Child and Maternal Health Needs

The Healthy Start Initiative in Florida, a series of legislative measures intended to improve pregnancy and birth outcomes within the state, provides for the establishment of community-based prenatal and infant health care coalitions composed of health care providers, representatives of state and local government, community alliances, maternal and child health organizations, and consumers of family planning, prenatal care, and primary care services. Each coalition is required to conduct a needs assessment within its service delivery area and develop a service delivery plan. The needs assessment of the Gadsden Citizens for Healthy Babies, Inc., representing a small, primarily rural, majority African American county in north Florida, used existing data from the State of Florida vital statistics records and the U.S. Census of Population and Housing to estimate the magnitude and distribution of child and maternal health problems in the county.

First, pregnancy outcomes and related maternal characteristics within the county were investigated using data from the *Florida Vital Statistics* volumes, which report birth and death information collected annually within the state. In particular, the following indicators were examined:

- *Infant mortality.* The county's rate was far higher than national or state rates.
- *Fetal mortality.* The overall rate for the county was higher than the state goal and the rate for African American mothers was higher than for white mothers.
- *Neonatal mortality.* The rates were higher than the state goal for white mothers but below for African American mothers.
- *Postneonatal mortality.* The rates were below state goals.
- *Low birth rate babies.* There was a higher incidence for adolescents and women over age 35.
- *Very low birth weight births.* The overall rate was twice that for the whole state and exceeded state goals for both African American and white mothers.
- *Adolescent pregnancy.* The proportion of births to teens was over twice the state average; the rate for African American teens was more than twice that for white teens of the same age.
- *Age of mother.* The infant mortality and low birth rates were highest among children born to mothers 16-18 years of age.
- *Education of mother.* Mothers with less than high school education were slightly more likely to have low birth weight newborns but almost eight times more likely to have newborns identified as high risk on infant screening measures.

Based on these findings, three groups were identified with high risk for poor birth outcomes:

- Mothers less than 19 years of age
- Mothers with less than a high school education
- African American mothers

U.S. Census data on CD-ROM discs available from the Bureau of Census were then used to identify the number of women of childbearing age in each of these risk categories, the proportions who were in various low-income strata, and their geographical concentrations within the county according to census tract and zip code. This information was used by the coalition to identify the major problem areas in the county, set goals, and plan services.

SOURCE: Adapted from E. Walter Terrie, "Assessing Child and Maternal Health: The First Step in the Design of Community-Based Interventions," in *Needs Assessment: A Creative and Practical Guide for Social Scientists*, eds. R. Reviere, S. Berkowitz, C. C. Carter, and C. G. Ferguson (Washington, DC: Taylor & Francis, 1996), pp. 121-146.

same time, however, it would modify the distribution of congressional seats from state to state and, within states, result in the need to redraw electoral boundaries. Thus, although social researchers advise assessing the quality of data in terms of their measurement properties, the political implications of estimates of social problems and the distribution of the affected population may play a large role in the procedures used to collect the data.

When sources are used whose validity is not as widely recognized as that of the census, it is always necessary to examine carefully how the data were collected. A good rule of thumb is to anticipate that, on any issue, different data sources will provide disparate or even contradictory estimates. For needs assessment purposes, sometimes data on the same topic collected by opposing stakeholders can be especially useful. For example, both the Coalition Against Handguns and the National Rifle Association (NRA) have sponsored sample surveys of the U.S. population concerning approval or disapproval of gun control legislation. Although their reports differed widely in their conclusions—the one finding popular support for gun control measures and the other the opposite—close inspection of the data showed that many of the specific findings were nearly identical in the two surveys (Wright, Rossi, and Daly, 1983). Both surveys found that guns were owned by about half of U.S. households, for instance. Findings on which different surveys substantially agree can be regarded as having greater credibility.

Using Social Indicators to Identify Trends

On some topics, existing data sources will provide periodic measures that can be used to chart historical trends in the society. For example, the Current Population Survey of the Bureau of the Census collects data annually on the characteristics of the U.S. population using a large household sample. The data include measures of the composition of households, individual and household income, and household members' age, sex, and race. The regular Survey of Income and Program Participation provides data on the extent to which the U.S. population participates in various social programs: unemployment benefits, Aid to Families With Dependent Children, food stamps, job training programs, and so on. The National Crime Survey compiles annual data on crime victimization from a national survey of households (see Exhibit 4-E).

These regularly occurring measures, called *social indicators*, can provide important information for assessing social problems and needs in several ways. First, when properly analyzed, the data can often be used to estimate the size and distribution of the social problem whose course is being tracked over time. Second, the trends shown can be used to alert decisionmakers to whether certain social conditions are improving, remaining the same, or deteriorating. Finally, the social indicator trends can be used to provide a first, if crude, estimate of the effects of social programs that have been in place. For example, the Survey of Income and Program Participation can be used to estimate the coverage of such national programs as food stamps or job training.

Similarly, the proportions of U.S. households at or below the poverty level can be followed from year to year over the post-World War II years through data obtained by the Current Population Survey. The question whether there was more or less poverty in the 1980s than in the preceding decades can thus be answered by referring to this social indicator. This is not to say that the trend data provided by the Current Population Survey are without controversy; many believe that they underestimate the current level of poverty, whereas others believe the opposite (Ruggles, 1990).

Considerable effort is currently going into the collection of social indicator data on poor households in an effort to judge whether their circumstances have worsened or improved after the radical reforms in welfare enacted in the Personal Responsibility and Work Opportunity Reconciliation Act of 1996. Special surveys, concentrating on the well-being of children, are being conducted by the Urban Institute and the Manpower Development Research Corporation. In addition, the Bureau of the Census has extended the Survey of Income and Program Participation to constitute a panel of households repeatedly interviewed before and after the welfare reforms were instituted.

Unfortunately, the social indicators currently available are limited in their coverage of social problems, focusing mainly on issues of poverty and employment, criminal victimization, national program participation, and household composition. For many social problems, no social indicators exist or those that do support analysis of national trends but cannot be broken down to provide useful indicators of local trends.

Estimating Problem Parameters Through Social Research

In many instances, no existing data source will provide estimates of the extent and distribution of a problem of interest. For example, there are no ready sources of information about household pesticide misuse that would indi-

EXHIBIT 4-E Tracking Crime Victimization Trends Using Social Indicators

Since 1973 the Bureau of Justice Statistics in the Department of Justice has conducted an annual survey of a national sample of households that asks if each person in the household has been the victim of a crime during the year prior to the interview, whether reported to the police or not. The table below charts the ten-year trends in crimes with persons or households as their victims.

Victimization Rates per 1,000 Persons Age 12 and Older or per 1,000 Households

	1987	*1988*	*1989*	*1990*	*1991*	*1992*	*1993*	*1994*	*1995*	*1996*
Crimes of violence[a]	43.7	44.1	44.4	44.1	48.0	49.3	49.9	51.8	46.6	42.0
Property crimes[b]	298.4	295.3	295.3	276.5	282.3	325.3	318.9	310.2	290.5	266.3

NOTE: The 1987-1992 figures incorporate a statistical adjustment for a change in survey methodology.
a. Rape, robbery, assault.
b. Burglary, theft, auto theft.

SOURCE: U.S. Department of Justice, Bureau of Justice Statistics, *Criminal Victimization in the United States: 1973-92 Trends* (Washington, DC: U.S. Department of Justice, July 1994). U.S. Department of Justice, Bureau of Justice Statistics, *Criminal Victimization 1996* (Washington, DC: U.S. Department of Justice, November 1997).

cate whether it is a problem, say, in households with children. In other instances, good information about a problem may be available for a national or regional sample that cannot be disaggregated to a relevant local level. The National Survey of Household Drug Use, for instance, uses a nationally representative sample to track the nature and extent of substance abuse. However, the number of respondents from most states is not large enough to provide good state-level estimates of drug abuse, and no valid city-level estimates can be derived at all.

When pertinent data are nonexistent or insufficient, the evaluator must consider collecting new data. There are several ways of making estimates of the extent and distribution of social problems, ranging in increasing degrees of effort from relying on "expert" testimony to conducting large-scale sample surveys. Decisions about the kind of research effort to undertake must be based in part on the funds available and in part on how important it is to have precise estimates. If, for legislative or program design purposes, it is critical to know the number of malnourished infants in a political jurisdiction, a carefully planned health interview survey may be necessary. In contrast, if the need is simply to determine whether there is any malnutrition among infants, input from knowledgeable informants may be all that is required. This section describes the various procedures that can be used to determine the size of a social problem and its geographical and social distribution.

Agency Records

Records of organizations that provide services to the population in question are information sources that may be useful for estimating the extent of a social problem (Hatry, 1994). Some agencies keep excellent records on their clients, but others do not keep records of high quality or do not keep records at all. When an agency's clients include all the persons manifesting the problem in question and records are faithfully kept, then the evaluator need not search any further. Unfortunately, these conditions do not occur often.

It would be tempting to try to estimate, say, the extent of drug abuse by extrapolating from the records of persons treated in drug abuse clinics. To the extent that the drug-using community is fully covered by existing clinics, such estimates may be quite accurate. However, if drug abuse clinics did cover all or most of the drug-abusing population, drug abuse treatment programs might not be problematic. Hence, to the extent that a problem is being adequately handled by existing programs, data from such programs may be useful and accurate, but that is not the situation in which data are usually needed. In the case of drug abuse clinics, of course, it is doubtful that all drug abusers are in fact served by the clinics. (The different prevalence estimates obtained from a served population and a sample survey of the general population are illustrated in the example of battered pregnant women in Exhibit 4-C.)

Surveys and Censuses

When it is necessary to get very accurate information on the extent and distribution of a problem and there are no existing credible data, the evaluator may need to undertake original research using sample surveys or complete enumerations. Because they come in a variety of sizes and degrees of technical complexity, either of these techniques can involve considerable effort and skill, not to mention a substantial commitment of resources.

To illustrate one extreme, Exhibit 4-F describes a needs assessment survey undertaken

EXHIBIT 4-F Using Sample Surveys to Study the Chicago Homeless

Most sample surveys are based on the assumption that all persons can be enumerated and surveyed in their dwellings, an assumption that fails by definition in any study of the homeless. The strategy devised for the Chicago study therefore departed from the traditional survey in that persons were sampled from non-dwelling units and interviews were conducted at times when the separation between the homed and homeless was at a maximum. Two complementary samples were taken: (1) a probability sample of persons spending the night in shelters provided for homeless persons, and (2) a complete enumeration of persons encountered between the hours of midnight and 6 a.m. in a thorough search of non-dwelling-unit places in a probability sample of Chicago census blocks. Taken together, the shelter and street surveys constitute an unbiased sample of the homeless of Chicago.

A person was classified as homeless at the time of the survey if that person was a resident of a shelter for homeless persons or was encountered in the block searches and found not to rent, own, or be a member of a household renting or owning a conventional dwelling unit. Conventional dwelling units included apartments, houses, rooms in hotels or other structures, and mobile homes.

In the street surveys, teams of interviewers, accompanied by off-duty Chicago policemen, searched all places on each sampled block to which they could obtain access, including all-night businesses, alleys, hallways, roofs and basements, abandoned buildings, and parked cars and trucks. All persons encountered in the street searches were awakened if necessary and interviewed to determine whether or not they were homeless. In the shelter samples, all persons spending the night in such places were assumed to be homeless. Once identified, homeless persons were interviewed to obtain data on their employment and residence histories as well as their sociodemographic characteristics. All cooperating respondents were paid $5.00.

SOURCE: Adapted from P. H. Rossi, *Down and Out in America: The Origins of Homelessness* (Chicago: University of Chicago Press, 1989).

to estimate the size and composition of the homeless population of Chicago. The survey covered both persons in emergency shelters and homeless persons who did not use shelters. Surveying the latter involved searching Chicago streets in the middle of the night. The survey was undertaken because the Robert Wood Johnson Foundation and the Pew Memorial Trust were planning a program for increasing the access of homeless persons to medical care. Although there was ample evidence that serious medical conditions existed among the homeless populations in urban centers, there was virtually no precise, reliable information on either the size of the homeless population or the extent of medical problems in that population. Hence, the foundations funded a research project to collect the missing information. Although many regard the effort as less than satisfactory, this research stimulated further efforts to count the homeless and improve data collection procedures. For example, the

EXHIBIT 4-G Assessing the Extent of Knowledge About HIV Prevention

To gauge the extent of knowledge about how to avoid HIV infection, a sample of Los Angeles County residents was interviewed on the telephone. They were asked to rate the effectiveness of four methods that "some people use to avoid getting AIDS through sexual activity" (see table). Their highest rating was for monogamous sex between HIV-negative people, although 12% felt that even in these circumstances there were no assurances of safety. Condom use, despite reported problems with breakage, leakage, and misuse, was rated as very effective by 42% of the respondents and as somewhat effective by another 50%. Respondents were much less certain about the effectiveness of spermicidal agents, regardless of whether they were used in conjunction with an alternative method.

Percentage Distribution of Ratings of the Effectiveness of Different Prevention Methods

Prevention Method	*Very Effective*	*Somewhat Effective*	*Not at All Effective*	*Don't Know*
Monogamous sex between HIV-negative individuals	73	14	12	1
Using a condom alone	42	50	7	1
Using a diaphragm with spermicide	9	35	50	6
Using spermicide alone	7	32	53	8

SOURCE: Adapted from D. E. Kanouse et al., *AIDS-Related Knowledge, Attitudes, Beliefs, and Behaviors in Los Angeles County* R-4054-LACH (Santa Monica, CA: RAND, 1991).

1990 census gave special attention to counting the homeless. One learning experience from subsequent research is the need to take into account the differences in the ways the homeless spend their time from community to community and the extent to which they are found in shelters, indoor settings that provide meals, on the streets, and in other outside areas.

Although time-consuming and costly, such extensive efforts are sometimes required for diagnostic purposes. To illustrate, Burnam and Koegel (1988) made a strenuous effort to obtain a representative sample and found that 44% of Los Angeles' homeless spent the night before being interviewed in a shelter bed and 26% slept on the street. An earlier study, with a poorer quality sample, had resulted in estimates of 66% in shelter beds and 14% sleeping on the streets. Given the costs of shelter care and the hostility of residents to persons sleeping on the streets in their neighborhoods, accurate estimates were worth obtaining for both program planning and political reasons.

Usually, however, needs assessment research is not as elaborate as that described in Exhibit 4-F. In many cases, conventional sample surveys can provide adequate information. If, for example, reliable information is required about the number and distribution of children needing child care so that new facilities can be planned, it will usually be feasible to obtain it from sample surveys conducted on the telephone. To illustrate, Exhibit 4-G describes a telephone survey conducted with more than 1,100 residents of Los Angeles County to ascertain the extent of public knowledge concerning the effectiveness of different AIDS prevention behaviors. For mass media educational programs aimed at increasing awareness of ways to prevent AIDS, a survey such as this identifies both the extent and the nature of the gaps in public knowledge.

Many survey organizations have the capability to plan, carry out, and analyze sample surveys for needs assessment. In addition, it is often possible to add questions to regularly conducted studies in which a number of organizations buy "time," thereby reducing costs. Whatever the approach, it must be recognized that designing and implementing sample surveys can be a complicated endeavor requiring quite specific skills. For discussion of the various aspects of sample survey methodology, see Fowler (1993), Henry (1990), Rossi, Wright, and Anderson (1983), and Sudman and Bradburn (1982).

Key Informant Surveys

Perhaps the easiest approach to obtaining estimates of the extent of a social problem is to ask *key informants*, those persons whose position or experience should give them some perspective on the magnitude and distribution of the problem. Unfortunately, such reports are generally not especially accurate. Although key informants can often provide very useful information about the characteristics of certain target populations and the nature of service needs, as we will discuss later in this chapter, few are likely to have a vantage point or information sources that permit very good estimation of the number of persons affected by a social condition or the demographic and geographical distribution of those persons.

Consider, for example, the task of estimating the number of homeless persons in a community. Although well-placed key informants may have experience with some subset of that population, it will be difficult for them to extrapolate from that experience to an estimate of the size of the total population. Indeed, it can be shown that selected informants' guesses about the numbers of homeless in their localities vary widely and tend to be overestimates, sometimes quite large overestimates (see Exhibit 4-H).

On the grounds that key informants' reports of the extent of a problem are better than no information at all, evaluators may wish to conduct a key informant survey when no other research is possible or when available funds are insufficient to support a better approach. Given those circumstances, evaluators must take care to ensure that the key informant survey is of the highest quality. The researchers should choose the persons to be surveyed very carefully, attempting to ensure that they have the necessary expertise, that they are questioned in a careful manner, and that any qualifications they may have about their reports are obtained (Averch, 1994).

Forecasting Needs

Both in formulating policies and programs and evaluating them, it is often important to be able to estimate what the magnitude of a social problem is likely to be in the future. A problem that is serious now may become more or less serious in later years, and program planning must attempt to take such trends into account. Yet the forecasting of future trends can be quite risky, all the more so as the time horizon lengthens.

EXHIBIT 4-H Using Key Informant Estimates of the Homeless Populations

To ascertain how close "expert" estimates of the number of homeless persons in downtown Los Angeles came to actual counts of homeless persons "on the streets," in shelters, or in single-room occupancy (SRO) hotels, a team of researchers asked eight service providers in the Skid Row area—shelter operators, social agency officials, and the like—to estimate the total homeless population in that 50-block area. The estimates obtained were as follows:

- Provider 1: 6,000 to 10,000
- Provider 2: 200,000
- Provider 3: 30,000
- Provider 4: 10,000
- Provider 5: 10,000
- Provider 6: 2,000 to 15,000
- Provider 7: 8,000 to 10,000
- Provider 8: 25,000

Clearly, the estimates were all over the map. Two providers (4 and 5) came fairly close to what the researchers estimated as the most likely number, based on shelter, SRO, and street counts.

SOURCE: Adapted from Hamilton, Rabinowitz, and Alschuler, Inc., *The Changing Face of Misery: Los Angeles' Skid Row Area in Transition—Housing and Social Services Needs of Central City East* (Los Angeles: Community Redevelopment Agency, July 1987).

There are a number of technical and practical difficulties in forecasting that derive in part from the necessary assumption that the future will be much like the past. For example, at first blush a projection of the number of persons in the population aged 18 to 30 a decade from now seems easy to construct from current data—it is almost completely determined by the present age structure of the population. However, had demographers made forecasts ten years ago for central Africa, they would have been substantially off the mark because of the unanticipated and tragic impact of the AIDS epidemic, which is most prevalent among young adults. Projections with longer time horizons would be even more problematic because they would have to take into account trends in fertility as well as mortality.

We are not arguing against the use of forecasts in a needs assessment. Rather, we wish to warn against accepting forecasts uncritically without a thorough examination of how they were produced. Moreover, such critical examination may itself involve some difficulty. For simple extrapolations of existing trends, the assumptions on which a forecast is based may be relatively few and easily ascertained. Even if the assumptions are known, however, it may not be easy to determine whether they are reasonable. For sophisticated projections such as those developed from multiple-equation, computer-based models, examining the assumptions may require the skills of an advanced programmer and the expertise of an experienced statistician. In any event, it must be recognized that all but the simplest forecasts are technical activities that require specialized knowledge and procedures.

DEFINING AND IDENTIFYING THE TARGETS OF INTERVENTIONS

Correctly defining and identifying the targets for intervention is crucial to the success of social programs from the very early stage when stakeholders begin to converge in their definition of a social problem to the extended period over which the program is operated. Specifying those targets is complicated by the fact that the definition and corresponding estimates of the size of the population may shift over this period. As a new social problem emerges or becomes increasingly visible, one definition of the targets of an intervention may be adopted; as stakeholders plan and eventually implement a program initiative, however, that definition may well be modified or abandoned.

As an illustration, during the early 1980s the problem of homelessness became extremely salient. Initially, the homeless were identified as those individuals who lived in streets and alleyways or in shacks they constructed for themselves. As advocates of the homeless became increasingly active, however, the targets of interventions began to also include persons who spent periods of time sleeping in shelters (sensibly so, because many persons who sleep in shelters also sleep out on the streets, and vice versa). Then, as programs began to emerge, some of them took the view that persons who had no regular place to live but moved in for brief periods with various relatives, friends, and sometimes strangers should be included. For some stakeholders and programs, the homeless population also encompassed the large number of individuals who lived in single rooms, usually paying for them daily or weekly and without the protection of leases or other contractual arrangements. (For further discussion of the technical and policy issues in homeless research, see Carr, 1991.)

As we will discuss shortly, the targets of an intervention need not be persons or groups of persons: They can also be organizations or "conditions." Here the same point about shifts in target definition applies. The target of a social intervention might be defective housing, which, at first, might be defined in terms of serious violations of the building code and later come to also include, for instance, buildings that are not earthquake proof.

Although the definition and estimated number of targets for a program are always in flux to some degree, they are critically important at two points in time. First, in the predesign phase of laying out policy and program alternatives, it is important to clearly identify the intended targets. Obviously, this is necessary so that the size and character of the social problem as a basis for realistic planning can be estimated. In addition, how the targets are defined will affect the potential effectiveness of a policy or program because that definition will shape the program emphasis and approach. For these reasons, new legislation and program proposals ordinarily contain statements about who or what the targets are to be.

Second, the issue of target definition is critical during the course of designing the specific content and structure of a program. Although targets may be initially defined in legislation or program proposals, in practice, programs often have to limit the targets to which they actually direct an intervention to some portion of those defined targets. This often happens simply because the funds provided are insufficient to cover all the originally

intended targets or because the nature of the program is such that it cannot be appropriately provided to all those in the initially defined target population.

Sometimes the reverse is true and, in operation, the program is directed at a broader or larger group than originally planned. For example, health clinics set up in shelters for the homeless were initially intended to provide services for persons sleeping there and in the streets. However, when poor persons housed in the same neighborhoods learned of the clinics, some of them began to use them too and were accepted as patients.

What Is a Target?

The targets of social programs are usually individuals. But they also may be groups (families, work teams, organizations), geographically and politically related areas (such as communities), or physical units (houses, road systems, factories). Whatever the target, it is imperative at the outset of a needs assessment to define the units in question clearly.

In the case of individuals, targets are usually identified in terms of social and demographic characteristics, location, or their problems, difficulties, and conditions. Thus, targets of an educational program may be designated as children aged 10 to 14 who are between one and three years below their normal grade in school. Or targets of a maternal and infant care program may be pregnant women and mothers of infants with annual incomes less than 150% of the poverty line.

When aggregates (groups or organizations) are targets, they are often defined in terms of the characteristics of the individuals that constitute them: their informal and formal collective properties and their shared problems. An organizational-level target for an educational intervention might be elementary schools (kindergarten to eighth grade) with at least 300 pupils in which at least 30% of the pupils qualify for the federal free lunch program.

Direct and Indirect Targets

Targets may also be regarded as direct or indirect depending on whether services are delivered to them immediately (directly) or eventually (indirectly). Most programs specify direct targets. This is clearly the case in medical interventions, for example, where persons with a given affliction directly receive medical treatments. However, in some cases, either for economic or feasibility reasons, planners may design programs to affect a target population indirectly by acting on an intermediary population or condition that will, in turn, have an impact on the intended target population. In a rural development project, for example, influential farmers were selected from small communities for intensive training programs. The intention was that afterward they would return to their communities and communicate their new knowledge to other farmers, the indirect targets of the program. Similarly, a project that identifies substandard dwelling units as its direct targets may be intended to benefit (indirectly) the current and future occupants of those dwellings.

When targets are defined as indirect, the program's effectiveness depends to a large extent on whether the pathways leading from immediate to ultimate targets are correctly identified in the program theory. The effectiveness of the project that used influential farmers, for instance, depended heavily on the ability and motivation of those farmers to communicate their knowledge persuasively to other farmers in their communities. Similarly, if there is a strong relationship between housing quality and health, investing in the physical improvement of housing to indirectly promote householders' well-being may be justified, but if the correlation is low or zero, the investment is likely to be ineffective for that purpose.

Specifying Targets

At first glance, specification of the size and distribution of target populations may seem simple. Although target definitions may be easy to write, however, it is often difficult to employ such definitions in the more precise work of needs assessment and program design. There are few human and social problems that can be easily and convincingly described in terms of simple, unambiguous characteristics of the individuals experiencing that problem.

Take a single illustration: What is the population of persons with cancer in a given community? The answer depends, first, on whether one counts only permanent residents or includes temporary ones as well (a decision that would be especially important in any community with a large number of vacationers such as Orlando, Florida). Second, are "recovered" cases to be counted, or are those without a relapse for, say, five years to be eliminated from the estimate? Third, is having cancer to be defined only as diagnosed cases or does it also include those persons whose cancer had not yet been detected? Finally, the estimate must take into account the purpose for which it is being used. If it is to be used in designing a special nursing-home program, for instance, persons with skin cancer should not be included because their condition rarely requires inpatient services.

An illustration of the considerations that go into specifying targets is provided in Exhibit 4-I, which is extracted from a landmark article that greatly influenced the development of the "poverty line" concept, a definition of poverty that is still employed today, adjusted to the current value of the dollar. Note that Orshansky's (1969) article dealt both with technical issues, such as the availability of appropriate data, and with substantive issues, such as trying to arrive at a definition that would satisfy stakeholders and social scientists. It is a tribute to her skill in balancing those objectives that the poverty-level concept she developed is still in use more than two decades later. The Orshansky measures have been heavily criticized, however, and several alternatives have been proposed (Ruggles, 1990).

Benchmarks for the size of major problem populations, such as the poor, and the definition underlying their identification have important consequences for the way governmental and private resources are allocated. For example, federal officials use the poverty-level approach just discussed in determining how much to appropriate for food stamp, medical care, and housing assistance programs.

Target Boundaries

Adequate target specification establishes boundaries, that is, rules determining who or what is included and excluded when the specification is applied. One risk in specifying target populations is to make a definition too broad or overinclusive. For example, specifying that a criminal is anyone who has violated any law or administrative regulation is useless; only saints have not at one time or another violated some law or regulation, wittingly or otherwise. This definition of criminal is too inclusive, lumping together in one category trivial and serious offenses and infrequent violators with habitual felons.

EXHIBIT 4-I Defining a Target Population: How Poverty Is Measured

Counting the poor is an exercise in the art of the possible. For deciding who is poor, prayers are more relevant than calculation because poverty, like beauty, lies in the eye of the beholder. . . . To say who is poor is to use all sorts of value judgments. The concept has to be limited by the purpose which is to be served by the definition. There is no particular reason to count the poor unless you are going to do something about them. When it comes to defining poverty, you can only be more subjective or less so. You cannot be nonsubjective.

Defining the Issue

We wanted to be sure that every family or consumer unit had its fair chance to be numbered among those who would be considered as needing attention. Indeed, it was precisely to ensure consideration of the needs of large families as well as small, and of young people as well as old, that we refined the initial standard developed by the Council of Economic Advisers. Their standard said that any family of two or more with less than $3,000 annual income, and any single person living alone with less than $1,500, would be considered poor for purposes of antipoverty program planning—but not for program eligibility. This original standard led to the odd result that an elderly couple with $2,900 income for the year would be considered poor, but a family with a husband, wife, and four little children with $3,100 income would not be.

Moreover, when we looked at the poor distributed demographically, by comparison with the total population, we made some unusual discoveries. For example, the percentage of the families classified as poor who had no children was higher than that for the population as a whole, and to make it even more unrealistic, the percentage of the poor families with four children or more was actually less than the representation of such families in the population. We did not think this was correct, so we tried to vary the poverty line—the necessary minimum of resources—with the size and composition of the family.

Setting the Benchmark

A concept which can help influence public thinking must be socially and politically credible. We need benchmarks to distinguish the population group that we want to worry about. A benchmark should neither select a group so small, in relation to all the population, that it hardly seems to deserve a general program, nor so large that a solution to the problem appears impossible. For example, in the 1930s, President Roosevelt said, "I see before me one-third of a nation ill-clothed, ill-housed, and ill-fed." This fraction is now part of our history. No matter how we get our numbers today, if more than a third of the population is called poor, it will lose value as a public reference point.

Definitions may also prove too restrictive, or underinclusive, sometimes to the point that almost no one falls into the target population. Suppose that the designers of a program to rehabilitate released felons decided to include only those who have never been drug or alcohol abusers. The prevalence of substance abuse is so great among released prisoners that only a small proportion would be eligible given this exclusion. In addition, because persons with long arrest and conviction histories are more likely to be substance abusers, this definition may eliminate those most in need of rehabilitation as targets of the proposed intervention.

In addition to specifying appropriate boundaries, useful target definitions must be

EXHIBIT 4-I Continued

At the Social Security Administration, we decided that we would develop two measures of need, and state, on the basis of the income sample of the Current Population Survey, how many and what kinds of families these measures delineated. It was not the Social Security Administration that labeled the poverty line. It remained for the Office of Economic Opportunity and the Council of Economic Advisers to select the lower of the two measures and decide they would use it as the working tool. The best you can say for the measure is that at a time when it seemed useful, it was there. It is interesting that few outside the Social Security Administration ever wanted to talk about the higher measure. Everybody wanted only to talk about the lower one, labeled the "poverty line," which yielded roughly the same number of people in poverty as the original $3,000 measure did, except that fewer families with more children were substituted for a larger number of older families without children.

Thresholds of Poverty

We have developed two poverty thresholds, corresponding to what we call the "poor" and "near-poor." These thresholds are set separately for 124 different kinds of families, based on the sex of the head, the number of children under 18, the number of adults, and whether or not the household lives on a farm. The threshold is defined as an attempt to "specify the minimum money income that could support an average family of given composition at the lowest level consistent with the standards of living prevailing in this country. It is based on the amount needed by families of different size and type to purchase a nutritionally adequate diet on the assumption that no more than a third of the family income is used for food." The two thresholds were developed from food consumption surveys conducted by the Department of Agriculture. . . . These revealed that the average expenditure for food by all families was about one-third of income.

An assumption was made that the poor would have the same flexibility in allocating income as the rest of the population but that, obviously, their margin for choice would be less. The amount allocated to food from the average expenditure was cut to the minimum that the Agriculture Department said could still provide American families with an adequate diet. We used the low-cost plan to characterize the near-poor and for the poor an even lower one, the economy food plan.

SOURCE: Quoted, with permission, from Mollie Orshansky, "Perspectives on Poverty: How Poverty Is Measured," *Monthly Labor Review*, 1969, 92(2):37-38.

feasible to apply. A specification that hinges on a characteristic that is difficult to observe or for which existing data contain no measures—for example, a definition of the targets of a job training program as persons who hold favorable attitudes toward accepting job training—may be virtually impossible to put into practice. Overly complex definitions requiring much detailed information are similarly difficult to apply. The data required to select targets defined as "farmer members of producers' cooperatives who have planted barley for at least two seasons and have an adolescent son" would be difficult, if not impossible, to gather. Moreover, in general the more criteria a definition has, the smaller the number of units that can qualify for inclusion in the target population. (The farmers satisfying the criteria just given would be a small group indeed.) Complex specifications are therefore kin to narrow ones and carry the same risks.

Varying Perspectives on Target Specification

Another issue in the definition of target populations arises from the differing perspectives of professionals, politicians, and the other stakeholders involved—including, of course, the potential recipients of services. During all phases of intervention, beginning with the emergence of a social problem, there can be differences in opinion as to the exact parameters of the target population.

Discrepancies may exist, for instance, between the views of legislators at different levels of government. At the federal level, Congress may plan to alleviate the financial burden on the government for natural disasters by encouraging states to invest in such disaster-mitigating measures as improved land-use management of flood plains and building codes that lower risks of damage and injury. From the federal perspective, the target population would be viewed as all those areas in which 100-year floods may occur. Because the federal government must be concerned with all the flood plains in the United States, their perspective recognizes that such a flood may occur somewhere as often as once every few days. True to their name, however, 100-year floods occur in any one place only once in every century (on average). From the local perspective, therefore, a given flood plain may not be viewed as a reasonable target at all and local governments may object strongly to the burdens of a program targeted on their flood plains.

Differences in perspective can arise in program design as well. The planners of programs concerned with improving the quality of housing available to poor persons may have a conception of housing quality much different from those of the people who will live in those dwellings. Their definition of what constitutes the target population of substandard housing for renewal, therefore, may result in a great outcry from residents of those dwellings who find them adequate.

Although needs assessment cannot establish which perspective on program targets is "correct," it can help eliminate conflicts that might arise from groups talking past each other. This is accomplished by investigating the perspectives of all the significant stakeholders on target definition and helping ensure that none is left out of the decision process through which the program focus is determined. Information collected about needs from varying perspectives may lead to a reconceptualization of the target population or of the prospective intervention, or even indicate the advisability of abandoning the program (especially if the different perspectives turn out to be contradictory and intensely held by the various stakeholders).

Useful Concepts in Target Definition

Understanding the nature of a social problem and estimating the size and characteristics of a target population are prerequisite to documenting the need for a program. Delivering service to a target population, however, requires that the definition of the target population permit targets to be distinguished from nontarget units in a relatively unambiguous and efficient manner as part of the program's normal operating procedures. To be effective, a program must not only know what its target population is but also be able to readily direct its services to that population and screen out individuals who are not part of that population. This section discusses a number of concepts that underlie appropriate target definition and selection.

Incidence and Prevalence

A useful distinction is the difference between *incidence* and *prevalence*. Incidence refers to the number of *new* cases of a particular problem that are identified or arise in a specified geographical or otherwise defined area during a specified period of time. Prevalence refers to the number of *existing* cases in a particular area at a specified time. These concepts are derived from the field of public health where generally they are sharply distinguished. For example, the incidence of influenza during a particular month is defined as the number of new cases reported during the month; its prevalence during that month is the number of afflicted people at any given time, regardless of when they were first stricken. In the health sector, project planners generally are interested in incidence when dealing with disorders of short duration, such as upper-respiratory infections and minor accidents. They are more interested in prevalence when dealing with problems that cannot be eradicated quickly but require long-term management and treatment efforts, including chronic diseases such as cancer and clinically observable long-term illnesses such as severe malnutrition.

The concepts of incidence and prevalence have been adapted to the study of social problems. In studying the impact of crime on victims, for instance, the critical measure is the incidence of victimization: the numbers of new cases (or persons victimized) per interval of time in a given area. Similarly, in programs aimed at lowering drunken-driver accidents, the incidence of accidents involving a drunken driver in a specified area and period of time may be the best measure of the need for intervention. But for chronic conditions such as low educational attainment, criminality, or poverty, prevalence is generally the appropriate measure. In the case of poverty, for instance, prevalence may be defined as the number of poor individuals or families in a community at a given time, regardless of when they became poor.

For other social problems, it is often unclear whether one should define target populations in terms of prevalence or incidence. In dealing with the problem of unemployment, it is important to know its prevalence, the numbers or proportions of the total population unemployed at a particular time. If the concern is with providing financial support for the unemployed, however, it is not clear whether the definition should refer to persons who are unemployed at a particular time or those who become unemployed in a given period. The principle involved centers on the issue of whether one is concerned with detecting and treating new cases as they appear or with existing cases whatever their time of origin.

Population at Risk

Another public health concept, *population at risk*, is helpful in specifying targets, particularly in projects that are preventive in character. Population at risk refers to that group of persons or units that has a significant probability of developing a given condition. Thus, the population at risk in fertility control programs is usually defined as women of childbearing age. Similarly, projects designed to mitigate the effects of typhoons and hurricanes may define targets as communities located in the typical paths of such storms and, hence, at risk of experiencing a disaster.

A population at risk can be defined only in probabilistic terms. Women of childbearing age may be the population at risk in a fertility control project, but a given woman may or may not conceive a child within a given period of time. In this instance, specifying the population at risk simply in terms of age results unavoidably in overinclusion; that is, the definition includes many women as targets who may not be in need of family planning efforts because they are not sexually active or are otherwise incapable of getting pregnant.

Sensitivity and Specificity

The *sensitivity* of a criterion for target identification refers to the likelihood of correctly selecting those targets who should be in a program in contrast to those who might also be selected by the criterion but not be appropriate for the program. If the program is designed to serve those with a specified condition, sensitivity is the ability of a screening or selection procedure to identify *true positives*, that is, those who actually have the condition. *Specificity*, on the other hand, refers to correctly excluding those persons or units who do not have the relevant condition, that is, *false positives*. A program that selects its clientele with poor sensitivity overlooks many who need and qualify for service. A program that selects with poor specificity uses its resources to serve many who do not need or qualify for service. Ideally, both high sensitivity and high specificity are desired in defining a target population and selecting individuals for a program.

Need and Demand

Whereas a population at risk includes all those with a high probability of having or acquiring a given condition, a *population in need* is a group of potential targets who currently manifest the condition. A population in need can usually be defined rather exactly; that is, one can identify a precise criterion for including a unit among targets (e.g., a screening technique). For instance, there are reliable and valid tests for determining an individual's degree of literacy. These tests can be used to specify a target population of functionally illiterate persons. For projects directed at alleviating poverty, one may define the population in need as families whose income, adjusted for family size, is below a certain minimum.

Just because individuals constitute a population in need by some criteria representing the social construction of need used in program or policy context, however, does not mean that they necessarily want the program or service at issue. Desiring service or being willing to participate in a program defines *demand* for service, a concept that usually only partially overlaps the applicable criteria of need. Community leaders and service providers, for instance, may quite reasonably define a "need" for overnight shelter among homeless persons sleeping on the streets but may find that some of these persons will not use such facilities. Thus, there may be a need but not a demand.

Some needs assessments undertaken to estimate the extent of a problem and serve as the basis for designing programs are actually *at-risk assessments* or *demand assessments* according to the definitions just offered. Such assessments may do duty for true needs assessments either because it is technically infeasible to measure need or because it is impractical to implement a program that deals only with the at-need population. For example, although only sexually active females may require family planning information, the target population for most such programs is those women assumed to be at risk, generally defined by an age span such as 15 to 50, because it would be difficult to identify and designate only those who are sexually active. Similarly, whereas the in-need group for an evening educational program may be all nonliterate adults, only those who are willing or who can be persuaded to participate can be considered the target population (an "at demand" definition). Clearly, the distinctions between populations at risk, in need, and at demand are important for estimating the scope of a problem, anticipating the size of the target population, and subsequently designing, implementing, and evaluating the program.

Rates

In addition to estimating the size of a problem group, it is generally also important to know the proportion of a population with a particular problem. Many times it is critical to be able to express incidence or prevalence as a *rate* to compare areas or problem groups. Thus, the number of new cases of unemployment or underemployment during a given period in an area (incidence) might be described per 100 or per 1,000 of a population (e.g., 133 new unemployed persons per 1,000 population).

Rates or percentages are especially critical in identifying the characteristics of the target population. For example, in describing the population of crime victims, it is important to have estimates by sex and age group. Although almost every age group is subject to some kind of crime victimization, young people are much more likely to be the victims of robbery and assault, whereas older persons are more likely to be the victims of burglary and larceny; men are considerably less likely than women to be the victims of sexual abuse; and so on. The ability to estimate targets by various characteristics allows a program to be planned and developed in ways that maximize opportunities to include the most appropriate participants and to tailor the program to the particular characteristics of sizable groups.

Estimates of target populations and their characteristics may be made at several levels of disaggregation. For example, illiteracy rates, calculated by dividing the number of functionally illiterate persons in various age groups by the total number of persons in each group, allow one to estimate the target populations that can be reached by tailoring a project to specific age cohorts. More powerful statistical techniques may usefully be employed to take into account additional sociodemographic variables simultaneously.

In most cases, it is not only traditional but also useful to specify rates by age and sex. In communities in which there are marked subcultural differences, racial, ethnic, and religious groups also become important denominators for the disaggregation of characteristics. Other variables useful in identifying characteristics of the target population include socioeconomic status, geographic location, and residential mobility. (See Exhibit 4-J for an example of crime

EXHIBIT 4-J Rates of Violent Crime Victimization, by Sex, Age, and Race

Victimization per 1,000 Persons Age 12 or Older: 1996

Characteristic of Victim	*All Violent Crimes*	*Rape; Sexual Assault*	*Robbery*	*Aggravated Assault*	*Simple Assault*
Sex					
Male	49.9	0.4	7.2	11.6	30.8
Female	34.6	2.3	3.4	6.2	22.7
Age					
12-15	95.0	2.6	10.0	15.6	66.8
16-19	102.7	4.9	12.0	25.3	60.4
20-24	74.3	2.1	10.0	15.9	46.4
25-34	51.1	1.8	7.1	9.8	32.4
35-49	32.8	1.3	3.8	7.4	20.3
50-64	15.7	0.1	1.8	3.8	10.0
65+	4.9	0.0	1.1	0.8	3.0
Race					
White	40.9	1.3	4.2	8.2	27.2
Black	52.3	1.8	11.4	13.4	25.6
Hispanic	44.0	1.2	8.4	10.6	23.9
Other	33.2	2.1	7.4	7.2	16.6

SOURCE: U.S. Department of Justice, Bureau of Justice Statistics, *Criminal Victimization 1996* (Washington, DC: U.S. Department of Justice, November 1997).

victimization rates disaggregated by sex, age, and race.)

DESCRIBING THE NATURE OF SERVICE NEEDS

As described above, a central function of needs assessment research is to develop estimates of the extent and distribution of a given problem and the associated target population. However, it is also often important for such research to yield useful descriptive information about the specific character of the need within that population. This is important because it is often not sufficient for a social program to merely deliver some standard services in some standard way presumed to be responsive to a given problem or need. To be effective, a program may need to adapt its services to the local nature of the problem and the distinctive circumstances of the persons in need. This, in turn, requires information about the way in which the problem is experienced by those in need, their perceptions and attributions about relevant services and programs, and the barriers and difficulties they encounter in attempting to access services.

A needs assessment might, for instance, probe into the matter of why the problem exists and what other problems are linked with it. For example, a search for information on how many high school students study a non-English language may reveal that many schools do not offer such courses; thus, part of the problem turns out to be that opportunities to learn foreign languages are insufficient. Similarly, the fact that many primary school children of low socioeconomic backgrounds appear to be tired and listless in class may be explained with a finding that many regularly do not eat breakfast, which, in turn, reflects their families' economic problems. Of course, different stakeholders are likely to have different views about the nature and source of the problem so it is important that the full range of perspectives be represented (see Exhibit 4-K for an example of diverse stakeholder perspectives).

Cultural factors or perceptions and attributions that characterize a target population may be especially relevant to the effectiveness of a program's outreach to members of the target population and the way in which it delivers its service. A thorough needs assessment on poverty in Appalachian mountain communities, for instance, should reflect the sensitivities of the target population about their self-sufficiency and independence. Programs that are construed as charity or that give what are perceived as handouts are likely to be shunned by needy but proud families.

Another important dimension of service needs may involve difficulties some members of the target population have in using services. This may result from transportation problems, limited service hours, lack of child care, or a host of similar such obstacles. The difference between a program with an effective service delivery to needy persons and an ineffective one is often chiefly a matter of how much attention is paid to overcoming these barriers. Job training programs that provide child care to the participants, nutrition programs that deliver meals to the homes of elderly persons, and community health clinics that are open during evening hours all illustrate approaches that have based service delivery on a recognition of the complexity of their clients' needs.

Qualitative Methods for Describing Needs

Qualitative research can be especially useful for obtaining detailed, textured knowledge of the specific needs in question. Such research can range in complexity from interviews of a few persons or group discussions to elaborate ethnographic research such as that employed by anthropologists. As an example of the utility of such research, qualitative data on the structure of popular beliefs can contribute substantially to the effective design of educational campaigns. What, for instance, are the trade-offs people believe exist between the pleasures of cigarette smoking and the resulting health risks? A good educational program must be adapted to those perceptions.

Carefully and sensitively conducted qualitative studies are particularly important for uncovering process information of this sort. Thus, ethnographic studies of disciplinary problems within high schools may not only provide some indication of how widespread disciplinary problems are but also suggest why some schools have fewer disciplinary problems than others. The findings on how schools differ might have implications for the ways programs are designed. Or consider the qualitative research on household energy consumption that revealed the fact that few householders had any information about the energy consumption characteristics of their appliances. Not knowing how they consumed energy, these householders could not very well develop effective strategies for reducing their consumption.

A popular and useful technique for obtaining rich information about a social problem is the *focus group* approach made popular in the

EXHIBIT 4-K Stakeholders Have Different Perceptions of the Problems With Local Health Services

Telephone interviews were conducted in three rural Colorado communities to identify health service problems related to cancer. In each community the study participants included (a) health care providers (physicians, nurses, public health personnel), (b) community influentials (teachers, librarians, directors of community agencies, business leaders), and (c) patients or family members of patients who had a cancer experience. While there was general agreement about problems with availability and access to services, each stakeholder group had somewhat different perceptions of the nature of the problems:

Physicians and health care providers:

- Regional facilities only accept paying patients or close down.
- The remoteness of the community creates a lack of services.
- Physician shortage exists because of low salaries, large workloads, and difficult patients.
- We don't have training or equipment to do high-tech care.

Community influentials:

- People are on waiting lists for services for several months.
- There are not enough professionals or volunteers here.
- There is inadequate provider knowledge about specialized services.

Patients and family members:

- A time or two we have had no doctor here.
- We have a doctor here now but his patients have no money and I hear he's going to leave.
- We need treatment locally.
- I was on a waiting list for three weeks before the mammography van got here.

SOURCE: Adapted from Holly W. Halvorson, Donna K. Pike, Frank M. Reed, Maureen W. McClatchey, and Carol A. Gosselink, "Using Qualitative Methods to Evaluate Health Service Delivery in Three Rural Colorado Communities," *Evaluation & the Health Professions*, 1993, 16(4):434-447.

past several presidential elections when small panels of voters were assembled to provide perceptions of how the campaign rhetoric was being received (Morgan, 1988). Focus groups bring together selected knowledgeable persons for a discussion of a particular topic or theme under the supervision of a facilitator (Dean, 1994; Krueger, 1988). With a careful selection and grouping of individuals, a modest number of focus groups can provide a wealth of descriptive information about the nature and nuances of a social problem and the service needs of those who experience it (Exhibit 4-L provides a helpful protocol for a needs assessment focus

EXHIBIT 4-L Sample Protocol for a Needs Assessment Focus Group

A focus group protocol is a list of topics or open-ended questions to be covered in a focus group session that is used to guide the group discussion. The protocol should (a) cover topics in a logical, developmental order so that they build on one another; (b) raise open-ended issues that are engaging and relevant to the participants and that invite the group to make a collective response; and (c) carve out manageable "chunks" of topics to be examined one at a time in a delimited period. For example, the following is a protocol for use in a focus group with low-income women to explore the barriers to receiving family support services:

- Introduction—greetings; explain purpose of the session; fill out name cards; introduce observers, ground rules and how the focus group works (10 minutes).
- Participant introductions—give first names only; where participants live, age of children; which family support services are received and for how long, and other services received (10 minutes).
- Introduce idea of barriers to services—ask participants for their views on what have been the most important barriers to receipt of family support services (probe regarding transportation, treatment by agency personnel, regulations, waiting lists); have they discontinued any services or been unable to get ones they want? (30 minutes).
- Probe for reasons behind their choices of most important barriers (20 minutes).
- Ask for ideas on what could be done to overcome barriers in the future—what would make it or would have made it easier to enter and remain in the service loop? (30 minutes).
- Debrief and wrap up—moderator summary, clarifications, and additional comments or questions (10 minutes).

SOURCE: Adapted from Susan Berkowitz, "Using Qualitative and Mixed-Method Approaches," in *Needs Assessment: A Creative and Practical Guide for Social Scientists*, eds. R. Reviere, S. Berkowitz, C. C. Carter, and C. G. Ferguson (Washington, DC: Taylor & Francis, 1996), pp. 121-146.

group). A range of other group techniques for eliciting information for needs assessment can be found in Witkin and Altschuld (1995).

Appropriate participants in focus groups would generally include various knowledgeable community leaders, directors of service agencies, the line personnel in those agencies who deal firsthand with clients, representatives of advocacy groups, persons experiencing the social problem or service needs directly, and other such stakeholders. Of course, for these interactions to be productive and comfortable for the participants, care must be taken in mixing some of these stakeholders in the same focus group.

Any use of key informants in needs assessment must, therefore, involve a careful selection of the persons or groups whose perceptions are going to be taken into account. A useful approach to identifying key informants for a needs assessment is *snowball sampling*. This technique requires that an initial set of appropriate informants be located through some reasonable means and surveyed. They are then

EXHIBIT 4-M Homeless Men and Women Report Their Needs for Help

As efforts to help the homeless move beyond the provision of temporary shelter, it is important to understand homeless individuals' perspectives on their needs for assistance. Responses from a representative sample of 1,260 homeless men and women interviewed in New York City shelters revealed that they had multiple needs not easily met by a single service. The percentage reporting a need for help on each of 20 items was as follows:

Finding a place to live	87.1	Problems with drugs	18.7
Having a steady income	71.0	Learning to get along better with other people	18.5
Finding a job	63.3	Nerves and emotional problems	17.9
Improving my job skills	57.0	Learning how to protect myself	17.6
Learning how to get what I have coming from agencies	45.4	Learning how to read and fill out forms	17.3
Getting on public assistance	42.1	Legal problems	15.0
Health and medical problems	41.7	Drinking problems	13.0
Learning how to manage money	40.2	Getting around town	12.4
Getting along with my family	22.8	Getting veteran's benefits	9.6
Getting on SSI/SSD	20.8	Problems with the police	5.1

SOURCE: Adapted from Daniel B. Herman, Elmer L. Struening, and Susan M. Barrow, "Self-Reported Needs for Help Among Homeless Men and Women," *Evaluation and Program Planning*, 1994, 17(3):249-256.

asked to identify other informants whom they believe are knowledgeable about the matter at issue. These other informants are then contacted and asked, in turn, to identify still others. When this process no longer produces relevant new names, it is likely that most of those who would qualify as key informants have been identified. Because those persons active and involved in any matter of public interest in a community tend to know of each other, snowball sampling works especially well for key informant surveys about social problems.

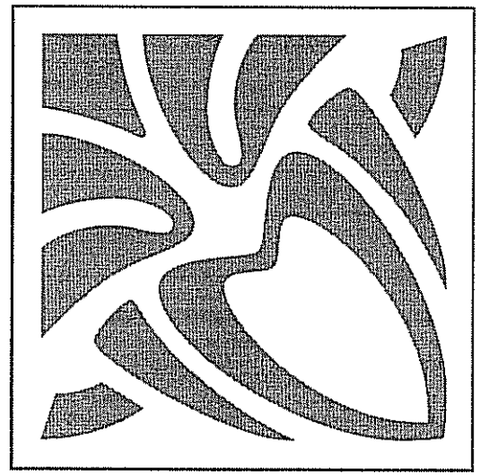
An especially useful group of informants that should not be overlooked in a needs assessment consists of a program's current clientele or, in the case of a new program, representatives of its potential clientele. This group, of course, is especially knowledgeable about the characteristics of the problem and the associated needs as they are experienced by those whose lives are most affected by the problem. Although they are not necessarily in the best position to report on how widespread the problem is, they are the key witnesses with regard to how seriously the problem affects individuals and what dimensions of it are most pressing. Exhibit 4-M illustrates the unique perspective of potential service beneficiaries.

Because of the distinctive advantages of qualitative and quantitative approaches, a useful and frequently used strategy is to conduct needs assessment in two stages. The initial, exploratory stage uses qualitative research approaches to obtain rich information on the nature of the problem (e.g., Mitra, 1994). The second stage, estimation, builds on this information to design a more quantitative assessment that provides reliable estimates of the extent and distribution of the problem.

SUMMARY

- Within evaluation research, needs assessment attempts to answers questions about the social conditions a program is intended to address and the need for the program, or to determine whether a new program is needed. More generally, it may be used to identify, compare, and prioritize needs within and across program areas.
- Adequate diagnosis of social problems and identification of the target population for interventions are prerequisites to the design and operation of effective programs. Nonetheless, it must be recognized that social problems cannot be objectively defined but, rather, are social constructions that emerge from the interests and political actions of the parties involved.
- To specify the size and distribution of a problem, evaluators may gather and analyze data from existing sources, such as the U.S. Census, or use ongoing social indicators to identify trends. Because the needed information often cannot be obtained from such sources, however, evaluators frequently conduct their own research on a social problem. Useful approaches include studies of agency records, surveys, censuses, and key informant surveys. Each of these has its uses and limitations; for example, key informant surveys may be relatively easy to conduct but of doubtful reliability; agency records generally represent persons in need of services but may be incomplete; surveys and censuses can provide valid, represen- tative information but can also be expensive and technically demanding.
- Forecasts for future needs are often very relevant to needs assessment but are complex and technical activities ordinarily performed by specialists. In using forecasts, evaluators must take care to examine the assumptions on which the forecasts are based.
- Appropriate definitions and accurate information about the numbers and characteristics of the targets of interventions are crucial throughout the intervention process, from initial planning through all the stages of program implementation. Targets may be individuals, groups, geographical areas, or physical units, and they may be defined as direct or indirect objects of an intervention.

- Good target specifications establish appropriate boundaries, so that an intervention correctly addresses the target population, and are feasible to apply. In defining targets, care must be taken to allow for the varying perspectives of different stakeholders. Useful concepts in target definition include incidence and prevalence, population at risk, sensitivity and specificity, need and demand, and rates.
- For purposes of program planning or evaluation, it is important to have detailed information about the local nature of a social problem and the distinctive circumstances of those in need of program services. Such information is usually best obtained through qualitative methods such as ethnographic studies or focus groups with selected representatives of various stakeholders and observers.



KEY CONCEPTS FOR CHAPTER 5

Program theory	The set of assumptions about the manner in which the program relates to the social benefits it is expected to produce and the strategy and tactics the program has adopted to achieve its goals and objectives. Within program theory we can distinguish *impact theory,* relating to the nature of the change in social conditions brought about by program action, and *process theory,* which depicts the program's organizational plan and service utilization plan (see Chapter 3 for fuller descriptions).
Articulated program theory	An explicitly stated version of program theory that is spelled out in some detail as part of a program's documentation and identity or as a result of efforts by the evaluator and stakeholders to formulate the theory.
Implicit program theory	Assumptions and expectations inherent in a program's services and practices that have not been fully articulated and recorded.
Evaluability assessment	Negotiation and investigation undertaken jointly by the evaluator, the evaluation sponsor, and possibly other stakeholders to determine if a program meets the preconditions for evaluation and, if so, how the evaluation should be designed to ensure maximum utility.
Black box evaluation	Evaluation of program outcomes without the benefit of an articulated program theory to provide insight into what is presumed to be causing those outcomes and why.

CHAPTER 5

EXPRESSING AND ASSESSING PROGRAM THEORY

Mario Cuomo, former governor of New York, once described his mother's rules for success as (a) figure out what you want to do and (b) do it. These are pretty much the same rules that social programs must follow if they are to be effective. In the last chapter, we discussed how the evaluator could assess the need for a program. Given an identified need, program decisionmakers must (a) conceptualize a program capable of alleviating that need and (b) implement it. In this chapter, we review the concepts and procedures an evaluator can apply to the task of assessing the quality of the program conceptualization. In the next chapter, we describe the ways in which the quality of the program implementation can be assessed.

The social problems that programs address are often so complex and difficult that bringing about even marginal improvement may pose formidable challenges. The foundation on which every program rests is some conception of what must be done to bring about the intended social benefits, whether that conception is expressed in a detailed program plan and rationale or is only implicit in the program's structure and activities. That conception is what we have referred to as the program theory.

A program's theory can be a good one, in which case it represents the "know-how" necessary for the program to attain the desired results, or it can be a poor one that would not produce the intended effects even if implemented well. One aspect of evaluating a program, therefore, is to assess how good the program theory is—in particular, how well it is formulated and whether it presents a plausible and feasible plan for improving the target social conditions. For program theory to be assessed, however, it must first be expressed clearly and completely enough to stand for review. This chapter describes how evaluators can tease out the theory implicit in a program and then, after it has been made explicit, assess how good it is.

In Chapter 3, we advocated that evaluators analyze a program's critical assumptions and expectations about the way in which it is intended to improve social conditions as an aid to identification of potentially important evaluation questions. This advice was presented in the context of planning an evaluation and setting priorities for the issues it would address. In this chapter, we return to the topic of program theory, not as a framework for identifying

important evaluation questions, but as a constituent part of the program itself.

Every program embodies a conception of the structure, functions, and procedures appropriate to attain its goals. This conception constitutes the "logic" or plan of the program, which we have called *program theory*. The program theory explains why the program does what it does and provides the rationale for expecting that doing things that way will achieve the desired results.

Evaluators and other informed observers recognize that there is little basis for presuming that program theory is universally sound and thus warrants little concern. There are many poorly designed social programs in operation with faults that reflect deficiencies in their underlying conception of how the desired social benefits can be attained. This circumstance stems in large part from the fact that careful, explicit conceptualization of program objectives and how they are supposed to be achieved is often not given sufficient attention during planning for new programs. Sometimes the political context within which programs originate does not permit extensive planning, but even when that is not the case, conventional practices for designing programs are not very probing with regard to the nature and plausibility of the underlying theory. The human service professions operate with repertoires of established modes and types of intervention associated with their respective specialty areas. As a result, program design is often principally a matter of configuring a variation of familiar "off the shelf" services into a package that seems appropriate for a social problem without a close analysis of the match between those services and the nature of the problem.

For example, many social problems that involve deviant behavior among the target population, such as alcohol and drug abuse, criminal behavior, early sexual activity, or teen pregnancy, are addressed by programs that provide some mix of counseling and educational services. Although rarely made explicit during planning, these programs are based on the assumption that people will change their problem behavior if given information and interpersonal support for doing so. Such theories may seem reasonable in the general case, but experience and research provide ample evidence that many such behaviors are very resistant to change despite knowledge by the participants about how to change and strong encouragement from loved ones to do so. Thus, the theory that education and supportive counseling will reduce deviant behavior may not be a sound basis for program design.

The rationale and conceptualization on which a program is based, therefore, should be subject to critical scrutiny within an evaluation just as any other important aspect of the program. If the program's goals and objectives do not relate in a reasonable way to the social conditions the program is intended to improve, or the assumptions and expectations embodied in a program's functioning do not represent a credible approach to bringing about that improvement, there is little prospect that the program will be effective. Evaluations of program process, impact, and efficiency thus ride on the presumption that the program theory is sound. Accordingly, evaluators must often make some assessment of the quality of a program's theory. Evaluating program theory, however, is not an easy task and certainly does not lend itself to structured and formalistic procedures. It is an important task, nonetheless, and one the evaluator must be prepared to undertake in many situations.

The first step in assessing program theory is to articulate it, that is, produce an explicit description of the conceptions, assumptions,

and expectations that constitute the rationale for the way the program is structured and operated. It is rare for a program to be able to immediately provide the evaluator with a statement of its program theory in a sufficiently explicit and detailed form to allow meaningful assessment. Although always implicit in program structure and operations, a full description of the program theory is seldom written down and available in program documents. Moreover, when some write-up of program theory is available, it is often in material that has been prepared for funding proposals or public relations purposes and may not correspond well with actual program practice.

Assessment of program theory, therefore, almost always requires that the evaluator first draw on program sources to synthesize and articulate the theory in a form amenable to analysis. Accordingly, the discussion in this chapter is organized around two themes: (a) how the evaluator can explicate and express program theory in a form that will be representative of key stakeholders' actual understanding of the program and workable for purposes of evaluation, and (b) how the quality of the articulated program theory can then be evaluated.

THE EVALUABILITY ASSESSMENT PERSPECTIVE

One of the earliest systematic attempts to describe and assess program theory arose from the experiences of an evaluation research group at the Urban Institute in the 1970s (Wholey, 1979). They found it often difficult, sometimes impossible, to undertake evaluations of public programs and began to analyze the obstacles. This led to the view that a qualitative assessment of whether minimal preconditions for evaluation were met should precede most evaluation efforts. Wholey and his colleagues termed the process *evaluability assessment* (see Exhibit 5-A).

Evaluability assessment generally involves three primary activities: (a) description of the "program model" with particular attention to defining the program goals and objectives, (b) assessment of how well defined and evaluable that program model is, and (c) identification of stakeholder interest in evaluation and the likely use of the findings. Evaluators conducting evaluability assessments operate much like program ethnographers. They seek to describe and understand the program through interviews and observations that will reveal its "social reality" as viewed by program personnel and other significant stakeholders. The evaluator begins with the conception of the program presented in documents and official information, but then tries to see the program through the eyes of those closest to it. The intent is to end up with a description of the program as it exists and an understanding of the program issues that really matter to the various parties involved. Although this process clearly involves considerable judgment and discretion on the part of the evaluator, various practitioners have attempted to codify its procedures so that evaluability assessments will be reproducible by other evaluators (see Rutman, 1980; Smith, 1989; Wholey, 1994).

A common outcome of evaluability assessments is that program managers and sponsors recognize the need to modify their programs. The evaluability assessment may reveal faults in a program's delivery system, that the program's target population is not well defined, or that the intervention itself needs to be reconceptualized. Or there may be few program

EXHIBIT 5-A The Rationale for Evaluability Assessment

If evaluators and intended users fail to agree on program goals, objectives, information priorities, and intended uses of program performance information, those designing evaluations may focus on answering questions that are not relevant to policy and management decisions. If program goals and objectives are unrealistic because insufficient resources have been applied to critical program activities, the program has been poorly implemented, or administrators lack knowledge of how to achieve program goals and objectives, the more fruitful course may be for those in charge of the program to change program resources, activities, or objectives before formal evaluation efforts are undertaken. If relevant data are unavailable and cannot be obtained at reasonable cost, subsequent evaluation work is likely to be inconclusive. If policymakers or managers are unable or unwilling to use the evaluation information to change the program, even the most conclusive evaluations are likely to produce "information in search of a user." Unless these problems can be overcome, the evaluation will probably not contribute to improved program performance.

These four problems, which characterize many public and private programs, can be reduced and often overcome by a qualitative evaluation process, *evaluability assessment*, that documents the breadth of the four problems and helps programs—and subsequent program evaluation work—to meet the following criteria:

- Program goals, objectives, important side effects, and priority information needs are well defined.
- Program goals and objectives are plausible.
- Relevant performance data can be obtained.
- The intended users of the evaluation results have agreed on how they will use the information.

Evaluability assessment is a process for clarifying program designs, exploring program reality, and—if necessary—helping redesign programs to ensure that they meet these four criteria. Evaluability assessment not only shows whether a program can be meaningfully evaluated (any program can be evaluated) but also whether evaluation is likely to contribute to improved program performance.

SOURCE: Quoted, with permission, from Joseph S. Wholey, "Assessing the Feasibility and Likely Usefulness of Evaluation," in *Handbook of Practical Program Evaluation*, eds. J. S. Wholey, H. P. Hatry, and K. E. Newcomer (San Francisco: Jossey-Bass, 1994), p. 16.

objectives that stakeholders agree on or no feasible performance indicators for the objectives. In such cases, the evaluability assessment has uncovered problems with the program design, which program managers must correct before any meaningful performance evaluation can be undertaken.

The aim of evaluability assessment is thus to create a climate favorable to evaluation work and an agreed-on understanding of the nature and objectives of the program that will facilitate evaluation design. As such, it can be integral to the approach the evaluator employs to tailor an evaluation and formulate evaluation questions

EXHIBIT 5-B Evaluability Assessment for the Appalachian Regional Commission

Evaluators from the Urban Institute worked with managers and policymakers in the Appalachian Regional Commission (ARC) on the design of their health and child development program. In this evaluability assessment, the evaluators

- Reviewed existing data on each of the 13 state ARC-funded health and child development programs;
- Made visits to five states and then selected two states to participate in evaluation design and implementation;
- Reviewed documentation related to congressional, commission, state, and project objectives and activities (including the authorizing legislation, congressional hearings and committee reports, state planning documents, project grant applications, ARC contract reports, local planning documents, project materials, and research projects);
- Interviewed approximately 75 people on congressional staffs and in commission headquarters, state ARC and health and child development staffs, local planning units, and local projects;
- Participated in workshops with approximately 60 additional health and child development practitioners, ARC state personnel, and outside analysts.

Analysis and synthesis of the resulting data yielded a *logic model* that presented program activities, program objectives, and the assumed causal links between them. The measurability and plausibility of program objectives were than analyzed and new program designs more likely to lead to demonstrably effective performance were presented. These included both an overall ARC program model and a series of individual models, each concerned with an identified objective of the program.

In reviewing the report, ARC staff were asked to choose explicitly among alternative courses of action. The review process consisted of a series of intensive discussions in which ARC and Urban Institute staff focused on one objective and program model at a time. In each session, the evaluators and staff attempted to reach agreement on the validity of the flow models presented, the importance of the respective objective, and the extent to which any of the information options ought to be pursued.

ARC ended up adopting revised project designs and deciding to systematically monitor the performance of all their health and child development projects and to evaluate the effectiveness of the "innovative" ones. Twelve of the 13 ARC states have since adopted the performance monitoring system. Representatives of those states report that project designs are now much more clearly articulated and that they believe the projects themselves have improved.

SOURCE: Adapted from Joseph S. Wholey, "Using Evaluation to Improve Program Performance," in *Evaluation Research and Practice: Comparative and International Perspectives*, eds. R. A. Levine, M. A. Solomon, G.-M. Hellstern, and H. Wollmann (Beverly Hills, CA: Sage, 1981), pp. 92-106.

(see Chapters 2 and 3). To illustrate the typical procedure, Exhibit 5-B presents an example of an evaluability assessment.

Evaluability assessment requires program stakeholders to articulate the program design and logic (the program model); however, it can

also be carried out for the purposes of describing and assessing program theory (Wholey, 1987). Indeed, the evaluability assessment approach represents the most fully developed set of concepts and procedures available in the evaluation literature for describing and assessing a program's conceptualization of what it is supposed to be doing and why. We turn now to a more detailed discussion of procedures for identifying and evaluating program theory, drawing heavily on the writings associated with the practice of evaluability assessment.

ELICITING AND EXPRESSING PROGRAM THEORY

Sometimes, though not often, a program's theory is spelled out in some detail in program documents and well understood by staff and stakeholders. In this case, we might say the program is based on an *articulated theory* (Weiss, 1997). This is most likely to occur when the original planning and design of the program are theory based. For instance, the design and delivery of a school-based drug use prevention program that features role-playing of refusal behavior in peer groups may be derived directly from social learning theory and its implications for peer influences on adolescent behavior.

In many cases, however, programs involve services and practices that are viewed as reasonable for the purposes of the program but the underlying assumptions and explanations of just how they are presumed to accomplish those purposes has not been fully articulated and recorded. In these cases, we might say that the program has an *implicit theory* or, as Weiss (1997) put it, a *tacit theory*. This might be the case for a counseling program to assist couples with marital difficulties. Although it may be reasonable to assume that discussing marital problems with a trained professional would be helpful, the way in which such interaction translates into improvements in the marital relationship is not described by an explicit theory nor would different counselors necessarily agree about that process.

When a program's theory is implicit rather than articulated, the evaluator must extract and describe it through some appropriate means before it can be analyzed and assessed. The first topics we must discuss, therefore, are how program theory can be elicited if it is not already fully articulated, how it might most usefully be expressed, and how it can be validated to ensure that it is an accurate representation of a program's actual working assumptions.

The evaluation literature presents diverse ways of defining and depicting program theory and thus muddies the waters for anyone wanting to see clearly to the bottom of this issue (Weiss, 1997). As indicated in Chapter 3, we view program theory as a relatively detailed description of the relationships between program resources, program activities, and program outcomes that shows how the program is supposed to work and, in particular, how it is supposed to bring about the intended outcomes. The objective for an articulation of program theory is to depict the "program as intended," that is, the actual expectations held by program decisionmakers about what the program is supposed to do and what results are expected to follow.

For expository purposes, we highlighted two components of a complete program theory: program impact theory (consisting of an action hypothesis and a conceptual hypothesis) and program process theory (consisting of the service utilization plan and the organizational plan). A brief review of these basic theory components may be helpful at this point.

Program impact theory delineates the cause-and-effect sequence through which the program is expected to bring about change in the social conditions it addresses. An agricultural extension program to increase the use of disease-resistant seeds among corn growers, for instance, may consist of distribution of educational materials, promotional talks with farmers' groups, and supply of seeds at discount prices. The presumptions that information, persuasion, and financial incentives will influence farmers' motivation (the action hypothesis) and that increased motivation will lead them to use the new seeds (conceptual hypothesis) constitute the program's impact theory.

Program process theory provides an account of how the program intends to bring about the desired interactions with the target population and provide the planned services. The service utilization plan describes how the target population will be engaged with the program from initial contact to completion of the intended services. For the agricultural extension program outlined above, the service utilization plan would lay out the sequence of interactions the target farmers are expected to have with the educational materials, the agricultural extension agents, and the suppliers of the disease-resistant seeds. The organizational plan would describe the program activities and resources and how they are to be organized and managed. To effectively promote the new seeds, for instance, the agricultural extension program would have to prepare and distribute educational materials, train agents and arrange promotional opportunities for them, acquire sufficient quantities of the new seeds, and organize distribution and the cost subsidy through, say, local commercial suppliers.

With this review as background, we turn to consideration of the concepts and procedures an evaluator can use to extract and articulate program theory as a prerequisite for assessing it.

What Is a Program for Purposes of Program Theory?

A crucial early step in articulating program theory is to define the boundaries of the program at issue (Smith, 1989). A human service agency may have many programs and provide multiple services; a regional program may have many agencies and sites. Depicting program theory thus almost always requires a clear definition of which components, activities, objectives, and target populations are encompassed in the program at issue. There is usually no one correct definition of a program for this purpose and the boundaries that the evaluator applies will depend, in large part, on the scope of the evaluation sponsor's concerns and the program domains to which they apply.

One way to circumscribe the program entity at issue is to work from the perspective of the decisionmakers expected to act on the evaluation findings and the nature of the decisions they are expected to make. What constitutes the program for which theory is to be articulated, then, should at minimum represent the relevant jurisdiction of those decisionmakers and the organizational structures and activities about which decisions are likely to be made. If the evaluation sponsor is the director of a single local community mental health agency, for instance, the decisions at issue and, hence, the boundaries of the program to be assessed may be defined primarily around one of the distinct combinations of service packages and target patients administered by that agency, such as outpatient counseling for eating

disorders. When the evaluation sponsor is the state director of mental health, however, the relevant program boundaries may be defined around effectiveness questions that relate to outpatient counseling services statewide, that is, the outpatient counseling component of all the local mental health agencies in the state.

Because program theory deals mainly with means-ends relations, the most critical aspect of defining program boundaries is to ensure that they encompass all the important activities, events, and resources linked to one or more outcomes recognized as central to the endeavor. This involves a form of backward mapping in which the point of departure is a set of well-defined program objectives relating to the social benefits the program intends to produce. From there, all the activities and resources under the relevant organizational auspices that are presumed to contribute to attaining those objectives are identified as part of the program. From this perspective, the eating disorders program at either the local or state level would be defined as the set of activities organized by the respective mental health agency that have an identifiable role in attempting to alleviate eating disorders for the eligible population in the respective jurisdiction.

Note, however, that although these approaches to defining a program for purposes of articulating program theory are straightforward in concept, they can be problematic in practice. Not only can programs be complex, with cross-cutting resources, activities, and goals, but the characteristics described above as linchpins for program definition can themselves be difficult to establish. Thus, in this matter, as with so many other aspects of evaluation, the evaluator must be prepared to negotiate a program definition agreeable to the evaluation sponsor and key stakeholders and be flexible about modifications as the evaluation progresses.

Procedures for Explicating Program Theory

For a program in the planning stage, theory might be derived from prior practice and research. For an existing program, however, the appropriate task is to describe the theory that is actually embodied in the program structure and operation. To accomplish this, the evaluator must interact with the program stakeholders to draw out their implicit program theory, that is, the theory represented in their actions and assumptions.

The Implicit Theory of Program Personnel and Other Stakeholders

The most straightforward approach to developing a description of a program's theory is to obtain it from program personnel and other pertinent stakeholders. The general procedure for this involves successive iteration. Draft descriptions of the program theory are generated, usually by the evaluator, and discussed with knowledgeable stakeholder informants to get feedback. The draft is then refined on the basis of their input and shown again to appropriate stakeholders. This process continues until the stakeholders find little to criticize in the description. The theory description developed in this fashion may involve impact theory or process theory or any components or combination deemed relevant to the evaluation purposes. Exhibit 5-C presents one evaluator's account of how a program process theory was formulated.

The primary sources of information for developing and differentiating descriptions of program theory are (a) review of program documents; (b) interviews with program personnel, stakeholders, and other selected informants; and (c) site visits and observation of various

EXHIBIT 5-C Formulating Program Process Theory for Adapted Work Services

Adapted Work Services (AWS) was initiated at the Rochelle Center in Nashville, Tennessee, to provide low-stress, paid work and social interaction to patients in the early stages of Alzheimer's disease. It was based on the belief that the patients would benefit emotionally and cognitively from working in a sheltered environment and their family members would benefit from being occasionally relieved of the burden of caring for them. The evaluator described the procedures for formulating a program process theory for this program as follows:

> The creation of the operational model of the AWS program involved using Post-it notes and butcher paper to provide a wall-size depiction of the program. The first session involved only the researcher and the program director. The first question asked was, "What happens when a prospective participant calls the center for information?" The response was recorded on a Post-it note and placed on the butcher paper. The next step was then identified, and this too was recorded and placed on the butcher paper. The process repeated itself until all (known) activities were identified and placed on the paper. Once the program director could not identify any more activities, the Post-it notes were combined into clusters. The clusters were discussed until potential component labels began to emerge. Since this exercise was the product of only two people, the work was left in an unused room for two weeks so that the executive director and all other members of the management team could react to the work. They were to identify missing, incorrect, or misplaced activities as well as comment on the proposed components. After several feedback sessions from the staff members and discussions with the executive director, the work was typed and prepared for presentation to the Advisory Board. The board members were able to reflect on the content, provide further discussion, and suggest additional changes. Several times during monthly board meetings, the executive director asked that the model be revisited for planning purposes. This helped further clarify the activities as well as sharpen the group's thinking about the program.

SOURCE: Quoted, with permission, from Doris C. Quinn, "Formative Evaluation of Adapted Work Services for Alzheimer's Disease Victims: A Framework for Practical Evaluation in Health Care" (doctoral diss., Vanderbilt University, 1996), pp. 46-47.

program functions and circumstances. Each of these warrants some discussion.

Documents. Some written description of the program or crucial aspects of it will almost always be available. However, the form, variety, and amount of such documentation will vary considerably depending on the nature of the program. For programs with legislative origins, there will likely be pertinent information in the authorizing legislation or legislative history and there may be accompanying regulations and guidelines. Descriptive information for programs is often found in conjunction with fund-raising and fiscal accountability. For example, grants, grant applications, contract documents, budget justifications, audit reports, and annual financial reports may provide

information about program goals and characteristics. Similarly, documents delineating formal commitments with other agencies or groups, for instance, interagency agreements to provide certain services, often include descriptive information about clients, services, and program objectives. Internal documents involving formal commitments, such as mission or vision statements, manuals of operating procedures, contracts with clients, job descriptions, and other such material, may also be informative.

In addition, many programs prepare and distribute promotional material. This may include flyers, brochures, newsletters, reports of program accomplishments, and listings in service and agency directories. Finally, any available management reports, organization charts, flowcharts, program monitoring documents, or evaluation studies relating to the program are good prospects for providing important descriptive information. Even though all these various sources of written information are not necessarily available or useful to the evaluator, thorough canvassing will generally turn up enough documentation to permit creation of a first approximation to a program theory description.

Although program documents can be informative for the evaluator attempting to describe the program and articulate the components of program theory, their limitations must be kept in mind. All program documents are prepared for some purpose, and that purpose will rarely be to present program theory in a valid and straightforward manner. The most descriptive documents are usually written to persuade some outside party to support the program and, naturally, have a self-serving bias. Others may describe an official or historical view of the program that does not coincide well with the program reality as it exists at the time of an evaluation. Thus, although program documents can be very illuminating to an evaluator attempting to understand and describe a program, their original context and purposes must be taken into account when they are interpreted.

Interviews. The most important sources of information describing a program and contributing to the articulation of program theory are those persons with firsthand knowledge and experience of the program. Generally, the best way for the evaluator to interact with these informants is through face-to-face discussion, individually or in small groups. Whereas written surveys and questions might be useful for some limited purposes, that approach lacks the flexibility to tailor the line of discussion to the expertise of the individual, probe and explore issues in depth, and engage the informant in careful reflection.

Central among the informants whose input is needed to properly depict program theory are the various members of the program staff in positions to know about key aspects of the program. Program managers and administrators, of course, are especially relevant because of their positions of oversight and responsibility. Line personnel should not be neglected when selecting informants, however. They often have the most detailed firsthand knowledge of how things actually work and generally are the personnel most closely in contact with the target population the program serves. Unique vantage points are also held by the program sponsors, funders, and policymakers, who often have a broader view of the objectives and goals of a program and its significance to the community than program personnel.

Critical sources of information related to the various components of program theory are members of the target population the program

serves. Surprisingly often this group is overlooked by evaluators formulating program descriptions and depicting program theory, perhaps because they are generally less accessible than program personnel and other such stakeholders. Nonetheless, representatives of the target population will generally have a unique perspective on the program, sometimes one at variance from that of other informants. They can be especially helpful to the process of formulating the service utilization plan by reporting on the nature of their contact and access to the program. As the recipients of a program's attempts to bring about change, their stories are also frequently illuminating for purposes of articulating the program's impact theory.

Finally, of course, the evaluator should obtain input from representatives of major stakeholders outside of the circle of persons directly involved with the program. This might include informants from other agencies, advocacy groups, community leaders, professional groups, and the like who have some interest in the program and some awareness of its purposes and activities. Many informants from these groups will not possess detailed knowledge of the program but may, nonetheless, provide useful insights about the perception of the program's purposes in the informed community, its relationships with other agencies and programs, and how it relates to social conditions and social needs recognized in that community.

Because theory description is worked out chiefly with stakeholders, evaluators experienced with evaluability assessment recommend that one or more stakeholder groups be organized to facilitate interaction for this purpose (e.g., Smith, 1989; Wholey, 1994). For example, Wholey (1994) reports that in many evaluability assessments it has proven useful to organize two groups, a policy group and a work group. A work group consists of program managers, staff, and representatives of stakeholders who are knowledgeable about program details and interact extensively with the evaluator to fashion a valid and useful rendition of program theory. The policy group, on the other hand, is composed of upper-level administrators, policymakers, and significant stakeholder representatives in decision-making roles whose feedback and endorsement are important to the acceptability and credibility of the theory description. This group is convened periodically for briefing and discussion as the work progresses.

Observation. Although program documents and stakeholder interviews will usually prove very informative, an evaluator is wise not to rely exclusively on them for describing the program and the theory it embodies. Documents and informants both have inherent limitations resulting from their partisan role in relationship to the program and the particular purposes and vantage points of their accounts. Based on experience with evaluability assessment, Wholey (1994) recommends that evaluators "explore program reality" firsthand through site visits and direct observation. In particular, evaluators should observe what they can of the program resources and routine operations so that they may make independent input to the formulation of program theory and so that they can be assured that the input from other sources is realistic with relation to the program capability.

The articulation of program theory necessarily and appropriately represents the program as intended more than the program as it actually is. Program managers and policymakers will generally think of the idealized program as the "real" one with various shortfalls from that ideal as glitches that do not represent what the

program is really about. Those further away from the day-to-day operations, on the other hand, may be unaware of such shortfalls and will naturally describe what they presume the program to be even if in actuality it does not quite live up to that image.

Some discrepancy between program theory and program reality is therefore natural. Indeed, examination of the nature and magnitude of that discrepancy is the task of process or implementation evaluation, as discussed in the next chapter. However, if the discrepancy is so great that the program theory describes activities and accomplishments that the program clearly cannot attain given its actual nature and resources, then the theory is overblown and cannot be realistically held up as a depiction of what is supposed to happen in the program context. For instance, suppose that a job training program's service utilization plan calls for monthly contacts between each client and a case manager. If the program resources are insufficient to support case managers, and none are employed by the program, this part of the theory is fanciful and should be revised to more realistically depict what the program might actually be able to accomplish.

The purpose of supplementing the accounts from program documents and stakeholders with direct observation, therefore, is not for the evaluator to verify that the program actually lives up to the intentions represented in its various theory components, but to ascertain that those intentions are generally realistic. When the program reality falls well short of the design envisioned by the key stakeholders, and that shortfall is readily apparent, there is little point in pursuing assessment of the theory or the details of how well the program implements the theory. Program redesign or reconceptualization is more in order, and the evaluator should provide that feedback to the pertinent stakeholders.

Collating information from different sources. Evaluators typically handle the information gleaned from program documents, interviews, and observations with some form of informal content analysis. Summaries or transcripts are made from the source material and then reviewed so that ambiguous or incomplete portions can be clarified with appropriate informants. The evaluator next extracts the pertinent information from each document in the form of thematic notes or excerpts and sorts them according to the aspect of the program to which they relate, such as goals, services, clients, personnel, program components, and resources. The information in each category is then used, along with other available information, to depict program theory in whatever representational form is preferred, for instance, a chart or graphic. Discussions of the general nature of this process can be found in Boyatzis (1998), Miles and Huberman (1994), Patton (1990), and Strauss and Corbin (1990). Exhibit 5-D reveals something of what the evaluator must bring to this process.

Topics for Attention During Document Review, Interviews, and Observations

Above, we reviewed the common sources of information useful to the task of articulating program theory but gave little attention to what information the evaluator might attempt to obtain from those sources. In this section, we turn attention to that matter.

Program goals and objectives. Perhaps the most important matter to be determined from program sources relates to the goals and objectives

EXHIBIT 5-D Theoretical Sensitivity

Theoretical sensitivity is the ability to recognize what is important in data and to give it meaning. It helps to formulate theory that is faithful to the reality of the phenomena under study. Theoretical sensitivity has two sources. First, it comes from being well grounded in the technical literature as well as from professional and personal experience. You bring this complex knowledge into the research situation. However, theoretical sensitivity is also acquired during the research process through continual interactions with the data—through your collection and analyses of the data. While many of the analytic techniques that one uses to develop theoretical sensitivity are creative and imaginative in character, it is important to keep a balance between that which is created by the researcher and the real. You can do so by (a) asking, what is really going on here? (b) maintaining an attitude of skepticism toward any categories or hypotheses brought to or arising early in the research, and validating them repeatedly with the data themselves; and (c) by following the data collection and analytic procedures as discussed in this book. Good science (good theory) is produced through this interplay of creativeness and the skills acquired through training.

SOURCE: Quoted, with permission, from Anselm Strauss and Juliet Corbin, *Basics of Qualitative Research: Grounded Theory Procedures and Techniques* (Newbury Park, CA: Sage, 1990), pp. 46-47.

of the program; these are necessarily an integral part of program theory, especially impact theory. The goals and objectives that must be represented in program theory, however, are not necessarily the same as those identified in program mission statements or in response to questions about goals to stakeholders. To be meaningful within an evaluation context, program goals must identify a state of affairs that could realistically be attained as a result of program actions; that is, there must be some reasonable connection between what the program does and what it intends to accomplish. Smith (1989) suggests that the evaluator use a line of questioning that does not ask about goals directly but, rather, about consequences. For instance, in a review of major program activities, the evaluator might ask about each, "Why do it? What are the expected results? How could you tell if those results actually occurred?" This approach attempts to keep the discussion concrete and specific rather than abstract and general as is typically the case if program goals are asked about directly (see Exhibit 5-E).

Given a set of relatively concrete and realistic goal statements, they must be integrated into the descriptions of program theory in a meaningful way. Within the context of the division of program theory into impact theory and process theory used here, the first distinction to be made is between goals appropriate to these different theory components. The ultimate goal of any social program should always be a specifiable improvement in the social conditions the program addresses. Thus, the goals and objectives that describe the outcome of the change process the program aims to bring about

EXHIBIT 5-E Asking About Program Goals and Objectives

The Illinois Cooperative Extension Service's teleconferencing network (TeleNet) was initiated to provide information and educational assistance to public officials, civic organizations, planning groups, and the general public on locally identified issues such as county jails, collective bargaining, and financial and personnel management. In an evaluability assessment, Midge Smith developed the following interview questions for program staff to help formulate program goals and objectives and the program activities and resources associated with them:

Goals and objectives:

- What do you think the TeleNet programs are trying to accomplish?
- What changes or differences, if any, is this program making with regard to participants, county advisors, the community, and the county?
- What negative effects, if any, might the program have or be having? (If some are mentioned, ask: What do you think could/should be done to avoid these negative effects?)

Tasks and activities:

- What tasks do you perform with the program?
- How does each of these tasks contribute to accomplishing the objectives?
- What problems, if any, do you face in performing these tasks?
- To what extent do you feel you reach the target audience?

Resources:

- What resources are used or are available at the local level to carry out the different program activities?
- How adequate are these resources?

Performance indicators:

- What are some of the indicators of success that the evaluation might try and measure? When could they be measured?
- Are there any questions or concerns about the program operation or results that you would like to see addressed by an evaluation?

SOURCE: Adapted from Midge F. Smith, *Evaluability Assessment: A Practical Approach* (Norwell, MA: Kluwer, 1989), p. 91.

in social conditions relate to program impact theory. Also associated with impact theory are any intermediate objectives that represent steps along the pathway leading from program services, on one end, to the improved social conditions that are the program's ultimate goal, at the other.

In contrast to the program goals and objectives related to effects on social conditions are those related to program activities and service delivery. These, in turn, are relevant to program process theory. For instance, "to provide case management" is a service objective but not an outcome goal because it describes action the program will take, not the effect of those actions on the social conditions the program aims to improve. An objective that might appear in the description of a program's service utilization plan, for instance, could be "all persons released from mental institutions are contacted and offered services" or "80% of the clients are retained in service for the full ten-week duration." These statements describe program accomplishments related to service delivery in terms of what happens to members of the target population, but do not address the benefits of those services for those persons. Similarly, the goals and objectives related to the program's organizational plan would deal with performing certain program functions, for instance, "to prepare curricular materials" or "to offer literacy classes four times a week."

One other consideration is important for the evaluator who is attempting to ascertain the various program goals and objectives and organize them into a description of program theory. The inquiry should attend to possible side effects and unintended outcomes that may be important for understanding the program as well as to the intended effects. Thus, a program "accomplishment" may be to have some impact that was not desired and may not be desirable. A mandatory job training program for women on welfare, for instance, may have the effect of increasing the number of children in substandard child care arrangements (Exhibit 5-F provides another example). Although such unintended effects cannot be said to be program goals, they follow from program activities in the same way as goal attainment and should be represented in program theory as possible outcomes.

Program functions, components, and activities. Program process theory mainly represents distinct program functions and how they relate to each other and to the participation of the targets in the program services. To properly describe this part of program theory, it is important for the evaluator to carefully identify each distinct program component, its functions, and the particular activities and operations associated with those functions. For this purpose, it is usually most instructive to view the program as a process rather than as an entity and describe it with verbs rather than nouns (Weick, 1982). An organization chart reveals little about how a program actually operates to achieve its objectives; that information appears in a description of what the program does. Thus, an essential part of describing program theory is to identify all the important program functions that must be performed for the program to operate as intended.

Program functions include such operations as "assess client need," "complete intake," "assign case manager," "recruit referral agencies," "train field workers," and the like. Viewed from the clients' perspective as part of a description of a service utilization plan, these functions appear in such forms as "receive referral for services," "contacted by case manager," "participate in group counseling sessions," and so forth. Each such function, whether represented from the program or the client perspective, will consist of various specific activities and will be associated with certain program personnel or components and resources. Full description of the program functions, therefore, also entails some level of description of the constituent activities and the program components and resources that support those activities (an example appears in Exhibit 5-G).

Logic or sequence linking program functions, activities, and components. A critical aspect of program theory is how the various steps and functions relate to each other. Sometimes those relationships involve only the temporal sequencing of key program activities and their effects; for instance, prison officials must notify the program that a convict has been released

EXHIBIT 5-F Unintended Effects of "Getting Tough" on Drunken Driving

California's drunk-driving laws were revised in 1982 to impose mandatory jail sentences and license suspension, even for first offenders, and to restrict plea-bargaining aimed at avoiding penalty. This new policy was intended to deter the practice of driving under the influence of alcohol (DUI) and reduce alcohol-related accidents.

However, subsequent research and anecdotal reports indicated that, whatever the positive effects of this policy, it also had a number of unintended and largely undesirable outcomes:

- An increase in court workloads. These changes resulted from an increased arrest rate for DUI and, also, because more defendants contested their arrest. There was a general decrease in the number of guilty pleas, an increase in the desire for attorney representation, an increase in the number of trials demanded by defendants (most noticeably for jury trials), and, because of the increased use of probation, an increase in probation revocation hearings.
- Increased cost for counties to provide defense and prosecuting attorneys. Because of the demands for more jury trials and the various avenues of postponement available to defendants, the cost of the time for publicly funded attorneys skyrocketed and some county boards of supervisors had to allot emergency funds to provide proper legal counsel to the influx of defendants.
- An increase in the need for new programs and facilities to deal with the DUI offenders. These offenders often served their sentences in areas or buildings apart from the mainstream jail populations or in special programs, for example, home monitoring systems to enforce house arrest or distinctive treatment, educational, or guidance programs.
- A strain on the correctional system. The increased numbers of DUI incarcerates caused a significant increase in the jail populations in all jurisdictions. Also, the DUI offenders occupied expensive space; due to overcrowding, many had to be housed in maximum and medium security space rather than minimum security. Probation populations also increased in the state.
- An upsurge in jail suicides. Individuals with drinking problems, who otherwise view themselves as law-abiding citizens, can feel stigmatized by incarceration; this apparently pushes some to take their own lives. Overcrowding exacerbates this problem by disallowing adequate prisoner supervision.

SOURCE: Adapted from Patrick T. Kinkade, Matthew C. Leone, and Wayne N. Welsh, "Tough Laws: Policymaker Perceptions and Commitment," *Social Science Journal*, 1995, 32(2):157-178.

before the program can initiate contact to arrange aftercare services. In other cases, these relationships have to do with activities or events that must be coordinated, as when child care and transportation must be arranged in conjunction with job training sessions, or with supportive functions, like training the instructors who will conduct in-service classes for nurses. Other relationships entail logical or conceptual linkages, especially those represented in program impact theory. Thus, the connection between mothers' knowledge about how to care for their infants and the actual behavior of providing that care assumes a psychological process through which information influences behavior (Exhibit 5-H describes some of the relationships that are basic to program logic).

Describing program theory, therefore, requires an understanding of how different events, persons, functions, and the other elements represented in the theory are presumed

EXHIBIT 5-G Program Functions for the Adapted Work Services Program

Exhibit 5-C earlier introduced the Adapted Work Services program, developed to provide low-stress, paid work and social interaction to patients in the early stages of Alzheimer's disease. The formulation of the program process theory that resulted from the procedures summarized in Exhibit 5-C was presented to stakeholders in the form of an "operational model" that described the following program functions:

Marketing	*Case Finding*	*Responding to Inquiries*	*Hosting Initial Contacts*
Compile data on community need	Develop referral sources	Answer questions from families, referral services, and participants	Introduction to staff, peers
Educate public and referral sources	Family members	Provide current cost schedule	Observe work in progress
Advertise program	EAPs	Complete screening	Staff observation of participant and family
	Health providers	Obtain referral form	Information sheet
	Assisted living facilities	Invite program visits	
	Hospital discharge		
	Church leaders		
	Senior citizen centers		
	Maintain referral network		

Conducting Assessments	*Intake and Assignment*	*Providing Service*	*Arranging Transitions*
Arrange trial period	Analysis of assessment	Transportation	Plan with caregiver for transitions
Observe	Testing	Coffee and socialization	Monitor criteria for discharge
Work competence	Family dynamics	Work periods	Self-select out
Motivation	Trial period	Training	Review transportation, health, family
Behavior	Enrollment	Support	Discuss with participant
Conduct testing	Application form	Case management	Feedback to
Mini-mental exam	Agreement form	Ongoing evaluation	Physician
Depression scale	Billing information	Functional status	Referral source
Assess family	Waiver form	Health	Families
Supportness	Nonenrollment	Communication	Follow-up
Ability to pay	Referral to other services	Progress reports	Family
Communicate with	Documentation	Caregivers	Next service provider
Referral source		Physicians	
Family		Referral sources	
Physician			

SOURCE: Adapted from Doris C. Quinn, "Formative Evaluation of Adapted Work Services for Alzheimer's Disease Victims: A Framework for Practical Evaluation in Health Care" (doctoral diss., Vanderbilt University, 1996), pp. 81-82.

EXHIBIT 5-H The Components of a Program Logic

A program logic consists of seven components, including

1. *An outcomes hierarchy.* This is a cause-effect hierarchy of desired outputs (e.g., the number of members of the target group serviced by a program), which lead to immediate impacts (e.g., changes in knowledge and skills of the target group), which in turn lead to outcomes (e.g., clients live an independent lifestyle, safer roads).

For each output, impact, and outcome in the hierarchy there should be

2. *Success criteria and definitions of terms* (e.g., what are the desired types of clients, what is meant by an independent lifestyle, what is meant by safer roads?).
3. *Factors that are within the control or influence of the program and are likely to affect the extent to which the outcome is achieved* (e.g., quality of service delivery, the way in which priorities are set).
4. *Factors that are outside the control or influence of the program and are likely to affect the extent to which the outcome is achieved* (e.g., the demographics of the target group, competing programs, past experiences of program clients).
5. *Program activities and resources used to control or influence both types of factors* (e.g., training given to staff to improve service quality, risk management strategies to respond to factors outside control).
6. *Performance information required to measure the success of the program in achieving desired outcomes* (e.g., the percentage of clients who show improved knowledge, information about the way in which the program is being implemented as a prerequisite for testing causal links between program activities and observed results).
7. *Comparisons required to judge and interpret performance indicators* (e.g., comparisons with standards to make judgments, comparisons with control conditions, pre-post comparisons to interpret performance and attribute it to the program).

SOURCE: Adapted from Sue Funnell, "Program Logic: An Adaptable Tool for Designing and Evaluating Programs," *Evaluation News and Comment: The Magazine of the Australasian Evaluation Society*, 1997, 6(1):5-17.

to be related. Because the number and variety of such relationships are often appreciable, evaluators typically construct charts or graphical displays to describe them (examples were shown in Chapter 3 and might be appropriately reexamined at this point). These may be configured as lists, flowcharts, hierarchies, or in any number of creative forms designed to identify the key elements and relationships in a program's theory. Such displays not only portray program theory but also provide a way to make it sufficiently concrete and specific for program personnel and stakeholders to engage. Working collaboratively with stakeholders to draft, differentiate, and discuss displays of program theory can be a very effective way for the evaluator to draw out the implicit knowledge of those informants.

Corroborating and Using Theory Description

Confirmation of a program theory description is chiefly a matter of demonstrating that pertinent program personnel and stakeholders endorse it as a meaningful account of how the program is intended to work. If it is not possible to generate a theory description that all relevant stakeholders accept as reasonable, this is diagnostic of a poorly defined program, conflicting perspectives among stakeholders about what the program is supposed to be doing and why, or competing program philosophies embodied in the same program. In such cases, the most appropriate response for the evaluator may be to take on a consultant role and assist the program in clarifying its assumptions and intentions to yield a theory description that will be acceptable to all key stakeholders.

Even when all pertinent stakeholders generally agree on a description of the program theory, they sometimes find parts of it questionable. Depicting the theory explicitly often surfaces assumptions and expectations inherent in a program that do not seem very plausible when laid out in black and white. This reaction may motivate program personnel and other stakeholders to pursue changes in program design. When this results from their involvement in the theory description process or from insights gained when the results of that process are reviewed, it demonstrates the utility of the theory description.

For the evaluator, the end result of the theory description exercise is a relatively detailed and complete statement of the program as intended that can then be analyzed and assessed as a distinct program evaluation function. Note that stakeholder agreement on the theory description serves only as confirmation that the description does, in fact, represent their understanding of how the program is supposed to work. This does not necessarily mean that the theory is a good one. To determine the soundness of a program theory, it must not only be described well but evaluated carefully. The procedures that evaluators use to assess program theory are described in the next section.

ASSESSING PROGRAM THEORY

Assessment of some aspect of a program's theory is relatively common in evaluation, often in conjunction with an evaluation of program process or impact. Nonetheless, outside of the modest evaluability assessment literature, remarkably little has been written of a specific nature about how this should be done, especially when the program design itself is the primary focus of the assessment. Our interpretation of this relative neglect is not that theory assessment is unimportant or unusual, but that it is typically done in an informal manner that relies on commonsense judgments, which, for most commentators, may not seem to require much explanation. Undeterred by the limited attention elsewhere, in this section we attempt to pull together a perspective on how to assess program theory, drawing from diverse sources and our own experience.

Frameworks for Assessing Program Theory

It is seldom possible or useful to individually appraise each distinct assumption and expectation represented in a program theory. But there are certain critical tests that can be conducted to provide general assurance that the program conceptualization is sound. Depending on how significant the questions about the

program theory are judged to be, a more or less stringent assessment may be appropriate. When there is little reason to believe that the program theory is problematic, its validity may be accepted on the basis of limited evidence or on commonsense judgment and "face validity." This is most likely to be the situation for programs whose services are directly related to straightforward objectives. A meals-on-wheels service, for instance, that brings hot meals to homebound elderly persons to improve their nutritional intake would be such a program.

Many programs are not based on presumptions as simple as the notion that delivering food to elderly persons improves their nutrition. A family preservation program that assigns case managers to coordinate community services for parents deemed at risk of having their children placed in foster care, for instance, involves many assumptions about exactly what it is supposed to accomplish and how. In such cases, the program theory might easily be faulty, and correspondingly, a rather probing evaluation of it may be warranted. The various procedures the evaluator might use for conducting that assessment are summarized below.

Assessment in Relation to Social Needs

The most important framework for assessing program theory builds on the results of needs assessment, as discussed in the previous chapter. Or, more generally, it is based on a thorough understanding of the social problem the program is intended to address and the service needs of the relevant target population, whether based on a formal needs assessment or not. A program theory that does not embody a conceptualization of program activities and outcomes that relate in an appropriate and effective manner to the actual nature and circumstances of the social conditions at issue will yield an ineffective program no matter how well implemented and administered. It is fundamental, therefore, to assess program theory in relationship to the needs of the target population the program is intended to serve.

There is no push-button procedure that an evaluator can use to assess program theory against social needs to determine if it describes a suitable conceptualization of how those needs should be met. Inevitably, this assessment requires a series of judgment calls. When the assessment is especially critical, its validity is strengthened if those judgments are made collaboratively with relevant experts and stakeholders to broaden the range of perspectives and expertise on which they are based. Such collaborators, for instance, might include social scientists knowledgeable about research and theory related to the intervention, administrators with long experience managing such programs, representatives of advocacy groups associated with the target population, and policymakers or policy advisers highly familiar with the program and problem area.

Whatever the nature of the group that contributes to the assessment, the crucial aspect of the process is *specificity*. When program theory and social needs are described in general terms, there often appears to be more correspondence than is evident when the details are examined. To illustrate, consider a curfew program prohibiting juveniles under age 18 from being outside their homes after midnight that is initiated in a metropolitan area to address the problem of skyrocketing juvenile crime. The program theory, in general terms, is that the curfew will keep the youths home at night and, if they are at home, they are unlikely to commit crimes. Because the general social problem the program addresses is juvenile crime, the program theory does seem responsive to the social need.

A more detailed problem diagnosis and service needs assessment, however, might show that the bulk of juvenile crimes are residential burglaries committed in the late afternoon when school lets out. Moreover, it might reveal that the offenders represent a relatively small proportion of the juvenile population who have a disproportionately large impact because of their high rates of offending. Furthermore, it might be found that these juveniles are predominantly latchkey youths who have no supervision during after school hours. When the program, in turn, is examined in some detail, it is apparent that it assumes that significant juvenile crime occurs late at night and that potential offenders will know about and obey the curfew. Furthermore, it depends on enforcement by parents or the police if compliance does not occur voluntarily.

Although more specificity than this would be desirable, even this much detail illustrates how program theory can be compared with need to discover shortcomings in the theory. In this example, examining the particulars of the program theory and the social problem it is intended to address reveals a large disconnect. The program blankets the whole city rather than targeting the small group of problem juveniles and focuses on late night activity rather than early afternoon when most of the crimes occur. In addition, it makes the questionable assumptions that youths already engaged in more serious lawbreaking will comply with a curfew, that parents who leave their delinquent children unsupervised during the early part of the day will be able to supervise their later behavior, and that the overburdened police force will invest sufficient effort in arresting juveniles who violate the curfew to enforce compliance. Careful review of these particulars alone would raise serious doubts about the validity of the program theory for addressing the social problem at issue (Exhibit 5-I presents another example).

One useful approach to comparing program theory with what is known (or assumed) about the respective social needs is to separately assess impact theory and program process theory. Each of these relates to the social problem in a different way and, as each is differentiated, specific questions can be asked about how compatible the assumptions of the theory are with the nature of the social circumstances to which it applies. We will briefly describe the main points of comparison for each of these theory components.

Program impact theory involves the sequence of causal links between program services and outcomes that improve the targeted social conditions. The key point of comparison between program impact theory and social needs, therefore, relates to whether the effects the program is expected to have on the social conditions according to the theory correspond to what the needs assessment indicates are required to improve those conditions. Consider, for instance, a school-based educational program aimed at getting elementary school children to learn and practice good eating habits. The problem this program attempts to ameliorate is poor nutritional choices among school-aged children, especially those in economically disadvantaged areas. The program impact theory would show a sequence of links between the planned instructional exercises and the children's awareness of the nutritional value of foods, culminating in healthier selections and improved nutrition.

Now, suppose a thorough needs assessment shows that the children's eating habits are, indeed, poor but their nutritional knowledge is not especially deficient. The needs as-

EXHIBIT 5-I The Needs of the Homeless as a Basis for Assessing Program Theory

Exhibit 4-M in the prior chapter on needs assessment described the responses of a large sample of homeless men and women to a needs assessment survey. The largest proportions identified a place to live and having a job or steady income as their greatest need. Fewer than half, but significant proportions, also said they needed help with medical, substance abuse, psychological, and legal problems. The evaluators reported that among the service delivery implications of the needs assessment were indications that this population needs interventions that provide ongoing support in a range of domains at varying degrees of intensity. Thus, to be responsive, programs must have the capacity to deliver or broker access to a comprehensive range of services.

These findings offer two lines of analysis for assessment of program theory. First, any program that intends to alleviate homelessness must provide services that address the major problems the homeless persons experience. That is, the expected outcomes of those services (impact theory) must represent improvements in the most problematic domains if the conditions of the homeless are to be appreciably improved. Second, the design of the service delivery system (process theory) must be such that multiple services can be readily and flexibly provided to homeless individuals in ways that will be accessible to them despite their limited resources and difficult circumstances. Careful, detailed comparison of the program theory embodied in any program for this homeless population with the respective needs assessment data, therefore, will reveal how sound that theory is as a design for effective intervention.

SOURCE: Daniel B. Herman, Elmer L. Struening, and Susan M. Barrow, "Self-Reported Needs for Help Among Homeless Men and Women," *Evaluation and Program Planning*, 1994, 17(3):249-256.

sessment further shows that the foods served at home and even those offered in the school cafeterias provide limited opportunity for healthy selections. Against this background, it is evident that the program impact theory is flawed. Even if the program successfully imparts additional information about healthy eating, the children will not be able to act on that information because they have little control over the selection of foods available to them. Thus, the proximal outcomes the program impact theory describes may be achieved, but they are not what is needed to ameliorate the problem at issue.

Program process theory, on the other hand, describes the interactions expected between the target population and the program (service utilization plan) and the functions the program is expected to perform (organizational plan). A sound process theory thus will make assumptions about the capability of the program to provide services accessible to the target population and compatible with their needs. These assumptions, in turn, can be compared with needs assessment information relating to the target population's opportunities to obtain service and the barriers that inhibit their service use.

As an example, consider an adult literacy program that offers classes in the evenings at the local high school. Its process theory incorporates significant instructional and advertising functions and it provides an appropriate selection of courses for the target population. The details of this scheme can be compared with needs assessment data that show what logistical and psychological support the target population requires to make effective use of the program. For instance, child care and transportation may be critical for many potential participants. Also, illiterate adults may be reluctant to enroll in courses without more personal encouragement than they would receive from advertising. Cultural and personal affinity with the instructors may be important factors in attracting and maintaining participation from the target population as well. The intended program process can thus be assessed in terms of how responsive it is to these dimensions of the needs of the target population.

Assessment of Logic and Plausibility

A thorough job of articulating program theory should reveal for inspection the critical assumptions and expectations inherent in the program's design. The program's goals and objectives will be specified and the primary program components and functions will be identified. The significant relationships among program functions and the nature of the expected interactions with the target population will be delineated. Most important, the description of the program's theory should lay out the cause-and-effect sequence through which program actions are presumed to ultimately produce the intended social benefits. One essential form of assessment is simply a critical review of the logic and plausibility of these various aspects of the program theory.

The appropriate questions to ask of the theory and its different aspects are basically of two sorts. First, "Is it well defined?" The theory in all its parts should be sufficiently specific, concrete, and clear to minimize ambiguity about what is supposed to be done and what is supposed to happen. Second, "Is it reasonable?" Informed reviewers should find it plausible that what is supposed to be done can be done and that what is expected to happen will happen. This judgment, in turn, will depend on an analysis of such matters as how logical the relationships are, what resources are available, what is viewed as realistic and practical within the organizational, political, and community context of the program, and assorted other such considerations. Exhibit 5-J describes such a review conducted as part of an evaluability assessment.

As should be apparent, assessing whether a program theory is well defined and reasonable requires considerable judgment and expertise. Although the evaluator will have a distinctive perspective on the issues and should be able to contribute importantly to the assessment, it would be rare for the evaluator to have the depth and breadth of knowledge about the program and its circumstances to be able to conduct a good assessment of its theory without assistance. As in the case of assessing the "fit" between a program's theory and the needs it addresses, discussed above, it may be appropriate to involve other knowledgeable persons in the review.

Commentators familiar with assessing program theory generally suggest that a panel of reviewers be organized for that purpose (Chen, 1990; Rutman, 1980; Smith, 1989; Wholey, 1994). This process may follow directly from the formulation of the program

EXHIBIT 5-J Assessing the Clarity and Plausibility of the Program Theory for Maryland's 4-H Program

An evaluability assessment of Maryland's 4-H youth program based on program documents and interviews with 96 stakeholder representatives included a review of key facets of the program's theory with the following results:

Question: Are the mission and goals clear?

Conclusion: There is a lack of clarity about the overall mission of 4-H and some lack of agreement among the stakeholders and between persons directly involved in implementing the program and those not. Among the statements of mission were "introduce youth to farm life," develop "sense of responsibility in agriculture and home economics," and "developing life skills."

Question: Is it clear who is to be affected, who is the audience?

Conclusion: There is some lack of agreement between 4-H faculty and the other stakeholders about the audience of 4-H. Written documents identified the audience as youth and adults; any youth between age 8 and 18 was viewed as the traditional audience for the program; recently, 6- and 7-year-olds have been targeted; some informants viewed the adult volunteers who assist with the program as one audience.

Question: Is there agreement about intended effects?

Conclusion: Social, mental, and physical development were listed as the program objectives in the state program direction document. There was agreement among all groups and in written documents that the effects of 4-H are primarily social in nature, for example, self-confidence/self-esteem, leadership, citizenship. There was less agreement about its effects on mental development and no agreement about its impact on physical development.

Question: Is it plausible that the program activities would achieve the intended effects?

Conclusion: Even if all the activities identified in the program model were implemented according to plan, the plausibility of these leading to the intended program effects is questionable. A link appears to be missing from the program logic—something like "Determine the Curriculum." Lack of such a link prevents plausible activities in the initial program events, that is, without a curriculum plan, how can county faculty know what types of leaders to recruit, what to train volunteers to do, and what they and the volunteers should implement?

SOURCE: Adapted from Midge F. Smith, *Evaluability Assessment: A Practical Approach* (Norwell, MA: Kluwer, 1989), p. 91.

theory and involve the groups organized for that purpose, such as the work group or policy group associated with an evaluability assessment. Certainly, an expert review panel should include selected representatives of the program staff and other major stakeholders as well as the evaluator. By definition, however, stakeholders have some direct stake in the program at issue. To balance the assessment and expand the available expertise, it may be advisable to bring in informed persons with no direct relationship to the program. Such outside experts might include experienced administrators of similar programs, social researchers with relevant specialties, representatives of advocacy groups or client organizations, and the like.

A review of the logic and plausibility of program theory will necessarily be a relatively unstructured and open-ended process. Many different aspects of the theory may be questioned in different ways, and there will be numerous particulars distinctive to the program and its context. Nonetheless, there are some general issues such reviews should be expected to address and that provide guidance for the assessment. These are briefly described below in the form of questions reviewers can ask. Additional useful detail can be found in Rutman (1980), Smith (1989), and Wholey (1994).

- Are the program goals and objectives well defined? The outcomes for which the program is to be accountable should be stated in sufficiently clear and concrete terms that it is possible to determine if they have been attained. One line of inquiry on this issue is to ask if there are observable implications of the goals and objectives such that meaningful measures and indicators of success could be defined. Goals such as "introducing students to computer technology" are not well defined in this sense whereas "increasing student knowledge of the ways computers can be used" is well defined and measurable.

- Are the program goals and objectives feasible; is it realistic to assume that they can actually be attained as a result of program action? Program theory should specify expected outcomes that are of a nature and scope that might reasonably follow from a successful program and should not be grandiose or represent unrealistically high expectations. Moreover, the stated goals and objectives should involve conditions the program might actually be able to affect in some meaningful fashion, not those that are largely beyond its influence. For instance, "eliminating poverty" is grandiose whereas "decreasing the unemployment rate" is not, but might be unrealistic for a program located in a chronically depressed labor market.

- Is the change process presumed in the program theory plausible? The presumption that a program will create benefits for the intended target population depends on the occurrence of some cause-and-effect chain that begins with the targets' interaction with the program and ends with the improved circumstances in the target population that the program expects to bring about (program impact theory). Every step of this causal chain should at least be plausible. Because the validity of this impact theory is the key to the program's ability to produce the intended effects, it is best if it is also supported by evidence that the assumed links and relationships do actually occur. A program based on the presumption that exposure to literature about the health hazards of drug abuse will motivate long-term heroin addicts to renounce drug use, for instance, does not present a plausible change process, nor is it supported by any research evidence.

- Are the program procedures for identifying members of the target population, delivering service to them, and sustaining that service through completion well defined and sufficient? The program theory should involve specific procedures and functions for locating potential service recipients, determining their eligibility, delivering service, and most important, handling all likely contingencies in this process. Moreover, those procedures should be

adequate for the purposes, viewed both from the perspective of the program's ability to perform them and the target population's likelihood of being engaged by them. Consider, for example, health screenings for high blood pressure among poor and elderly populations. It is reasonable to ask if these are provided in locations members of these groups frequent and if there is an effective means of locating those with uncertain addresses so that feedback can be provided. Absent these characteristics, it is unlikely that many persons from the target groups will receive the intended service.

- Are the constituent components, activities, and functions of the program well defined and sufficient? Program structure and process should be specific enough to permit orderly operations, effective management control, and monitoring using attainable, meaningful performance measures. Most critical, the program components and activities should be sufficient and appropriate to attain the intended program goals and objectives. Such functions as "client advocacy" have little practical significance if no personnel are assigned to it or there is no common understanding of what it means operationally. Similarly, providing a "supportive milieu" is not very convincing as the centerpiece of a program for emotionally disturbed adolescents.

- Are the resources allocated to the program and its various components and activities adequate? Program resources include funding, of course, but also personnel, material, equipment, facilities, relationships, reputation, and other such assets. There should be some reasonable correspondence between the program as intended that is described in program theory and the resources presumed (or known) to be available for operating the program. A program theory that calls for activities and outcomes that are unrealistic relative to available resources cannot be said to be a good theory. For a program to be conducted as expected and have the intended consequences, the assumptions made about how it will operate and what it will accomplish should be scaled to the resources available. For example, a management training program too short-staffed to initiate more than a few brief workshops cannot expect to have a significant impact on management skills in the organization.

Assessment Through Comparison With Research and Practice

Although every program is distinctive in some ways, few are based entirely on unique presumptions about how to engender change, deliver service to the target population, and perform major program functions. It follows that some information applicable to assessing the various components of program theory is likely to appear in the social science and human services research literature. In most program areas, there is also significant information available describing the experience of various programs with different practices, program approaches, and the like. One useful approach to assessing program theory once it is articulated, therefore, is to find out if it is congruent with research evidence and practical experience elsewhere (Exhibit 5-K summarizes one example of this approach).

There are several ways in which program theory might be compared with findings from research and practice. The most straightforward approach is to examine evaluations of programs based on similar concepts. The results will give some indication of the likelihood that such programs will be successful and perhaps identify some of the critical problem areas. Evaluations of very similar programs, of course, will be the most informative in this regard. However, evaluation results for programs that are similar only in terms of general theory, even if different in other regards, might also be instructive.

This approach can be illustrated by considering a mass media campaign in a metropolitan area to encourage women to have mammogram

EXHIBIT 5-K GREAT Program Theory Is Consistent With Criminological Research

In 1991 the Phoenix, Arizona, Police Department initiated a program with local educators to provide youths in the elementary grades with the tools necessary to resist becoming gang members. Known as GREAT (Gang Resistance Education and Training), the program has attracted federal funding and is now distributed nationally. The program is taught to seventh graders in schools over nine consecutive weeks by uniformed police officers. It is structured around detailed lesson plans that emphasize teaching youths how to set goals for themselves, how to resist peer pressure, how to resolve conflicts, and how gangs can affect the quality of their lives.

The program has no officially stated theoretical grounding other than Glasser's (1975) reality therapy, but GREAT training officers and others associated with the program make reference to sociological and psychological concepts as they train GREAT instructors. As part of an analysis of the program's impact theory, a team of criminal justice researchers identified two well-researched criminological theories relevant to gang participation: Gottfredson and Hirschi's self-control theory (SCT) and Akers's social learning theory (SLT). They then reviewed the GREAT lesson plans to assess their consistency with the most pertinent aspects of these theories. To illustrate their findings, a summary of Lesson 4 is provided below with the researchers' analysis in italics after the lesson description:

> Lesson 4. Conflict Resolution: Students learn how to create an atmosphere of understanding that would enable all parties to better address problems and work on solutions together. *This lesson includes concepts related to SCT's anger and aggressive coping strategies. SLT ideas are also present: Instructors present peaceful, nonconfrontational means of resolving conflicts. Part of this lesson deals with giving the student a means of dealing with peer pressure to join gangs and a means of avoiding negative peers with a focus is on the positive results (reinforcements) of resolving disagreements by means other than violence. Many of these ideas directly reflect constructs used in previous research on social learning and gangs.*

Similar comparisons showed good consistency between the concepts of the criminological theories and the lesson plans for all but one of the eight lessons. The reviewers concluded that the GREAT curriculum contained implicit and explicit linkages both to self-control theory and social learning theory.

SOURCE: Adapted from L. Thomas Winfree, Jr., Finn-Aage Esbensen, and D. Wayne Osgood, "Evaluating a School-Based Gang-Prevention Program: A Theoretical Perspective," *Evaluation Review*, 1996, 20(2):181-203.

screening for early detection of breast cancer. The impact theory for this program presumes that exposure to TV, radio, and newspaper messages will stimulate a reaction that will eventuate in increased rates of mammogram screening. Whatever the impact theory assumed to link exposure and increases in testing, its credibility is enhanced by evidence that similar media campaigns in other cities have resulted in increased mammogram testing. Moreover, if the evaluations or descriptive information for the campaigns in other cities shows that the program functions and scheme for delivering messages to the target population were similar to that intended for the program at issue, then the program's process theory also gains some support.

Suppose, however, that no evaluation results or practical descriptive accounts are available about media campaigns to increase rates of mammogram screening in other cities. It might still be informative to examine information about other media campaigns more or less analogous to the one at issue. For instance, reports may be available about the nature and results of media campaigns to promote immunizations, dental checkups, or other such actions that are health related and require a visit to a provider. The success of such programs, and different variations of such programs, might well be relevant to assessing the program theory on which the mammogram campaign is based so long as they involve similar principles.

In many program areas, numerous competing program approaches are directed toward accomplishing the same, or very similar, outcomes. Various different programs have been implemented for the treatment of alcoholism, for instance. In such cases, there are likely to be research reviews or meta-analyses that examine the existing evaluation research and summarize what has been learned about more and less promising program approaches. Consideration of how the program theory being assessed compares to those represented in the different program approaches covered in a research review often will support a convincing assessment.

In some instances, behavioral or social science research on the social and psychological processes central to the program will be available as a framework for assessing the program theory, particularly impact theory. From the perspective of the evaluation field, it is unfortunate that relatively little "basic" research has been done on many of the social dynamics that are common and important to intervention programs, because the results can be very useful. For instance, a mass media campaign to encourage mammogram screening, as in the example above, inherently involves persuasive messages intended to change attitudes and behavior. The large body of basic research on attitude change and its relationship to behavior in social psychology provides some basis for assessing the impact theory for any media campaign. One established finding, for instance, is that messages designed to raise fears are generally less effective than those providing positive reasons for a behavior. Thus, an impact theory based on the presumption that increasing awareness of the dangers of breast cancer will prompt increased mammogram screening may not be a good one.

There is also a large applied research literature on media campaigns and related approaches in the field of advertising and marketing. Although this literature largely has to do with selling products and services, it too may provide some basis for assessing the program theory for the breast cancer media campaign. Market segmentation studies, for instance, may show what media and what times of the day are best for reaching women with various demographic profiles. This information can then be used as part of the assessment of the program's service utilization plan to examine whether the media plan is optimal for communicating with women whose age and circumstances put them at the highest risk for breast cancer.

Fortunately, use of the research and practice literature to help with assessment of program theory is not limited to situations of relatively good overall correspondence between the programs or processes the evaluator is investigating and those represented in the literature. Often there is little or no literature dealing with programs or processes sufficiently similar to the ones under study to be applicable. An alternate approach for assessing program theory against existing research is to break the theory down into its component parts and linkages and search for research evidence relevant to each component.

Much of program theory can be stated as "if-then" propositions: If case managers are assigned, then more services will be provided; if school performance improves, then delinquent behavior will decrease; if teacher-to-student ratios are higher, then students will receive more individual attention. Frequently, research can be found that aids in appraising the plausibility of the individual propositions of this sort that are most fundamental to the program theory. The results of these appraisals, in turn, provide a basis for a broader assessment of the theory with the added advantage of potentially identifying any especially weak links. This approach was pioneered by the Program Evaluation and Methodology Division of the U.S. General Accounting Office as a way to provide rapid review of program proposals arising in the congress (Cordray, 1993; U.S. General Accounting Office, 1990).

Assessment Via Preliminary Observation

Program theory, of course, is inherently conceptual and not something that can be observed directly. However, program theory does involve many assumptions about how things are supposed to work that can be assessed by observing the program in operation, talking to staff and service recipients, and other such inquiries focused specifically on the program theory. Indeed, a thorough assessment of program theory description should incorporate some firsthand observation of the program and its circumstances and not rely entirely on logical analysis and similar "armchair" reviews. Direct observation provides something of a "reality check" to assess the concordance between the program theory and the program it is supposed to describe.

Consider a program theory that presumes that distributing brochures about good nutrition to senior citizens centers will influence the attitudes and eating behavior of elderly persons. Observations revealing that the brochures are rarely read by anyone attending the centers would certainly raise a question about a key assumption of the theory. In particular, this observation challenges the presumption that the target population will be exposed to the information in the brochures, which is a precondition for any attitude or behavior change. In this regard, it is the plausibility of the program theory, or a portion of it, that is being assessed, as discussed above. Rather than being assessed on the basis of informed judgment by the evaluator or other knowledgeable informants, however, it is checked directly through observation of the circumstances at issue or, perhaps, interviews with persons in close contact with those circumstances.

To assess program impact theory, observations and interviews might be focused on the intended program outcomes and the interactions between program services and the target population that are expected to produce those outcomes. This inquiry might look into the question of whether the intended outcomes are appropriate for the program circumstances and whether they are realistically attainable. For example, the presumption that a welfare-to-work program can enable a large proportion of welfare clients to find and maintain employment might be investigated by examining the local job market, the work readiness of the welfare population (number physically and mentally fit, skill levels, work histories, motivation), and the relative economic benefits of work to gauge how realistic the intended program outcomes are. At the service end of the change process, the job training activities might be observed and interviews with participants conducted to assess the plausibility that the intended changes would occur.

Inquiry aimed at testing the service utilization component of a program's process theory would examine the circumstances of the target population to better understand how and why they might become engaged with the program and, once engaged, continue until the intended services had been received. This information would permit some assessment of the quality of the program's service delivery plan for locating, engaging, and serving the intended target population. To assess the service utilization plan of a midnight basketball program to reduce delinquency among high-risk youths, for instance, the evaluator might observe the program in action and interview participants, program staff, and neighborhood youths about who participates and how regularly. The program's service utilization assumptions would be supported by indications that the most delinquent-prone youths participate regularly in the program.

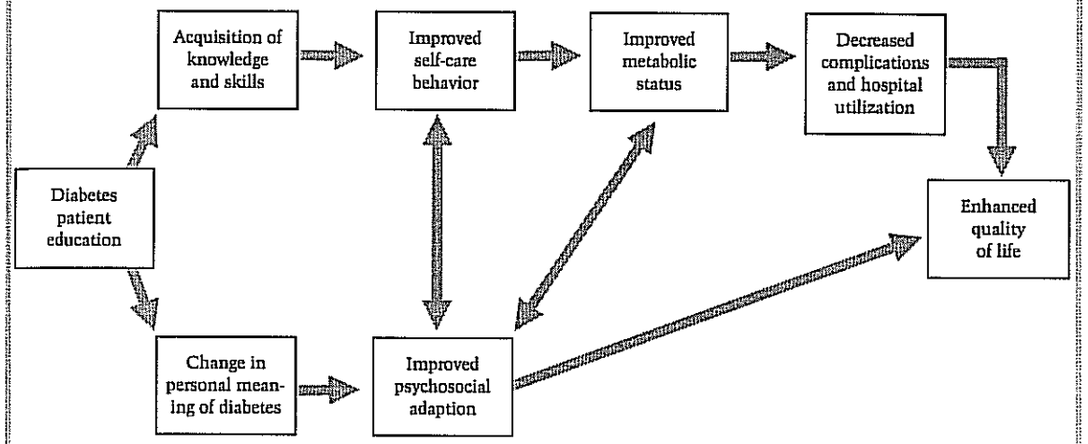
Indications of the plausibility of the organizational component of the program's process theory might be developed through observations and interviews relating to program activities and the supporting resources. Critical here is evidence that the program can actually perform the intended functions. Consider, for instance, a program plan that calls for the sixth-grade science teachers throughout the school district to take their classes on two science-related field trips per year. The evaluator could probe the presumption that this would be done by interviewing a number of teachers and principals to find out if this was broadly feasible in terms of scheduling, availability of buses, funding, and other such matters.

Any assessment of the practicality of program theory or its appropriateness for the program circumstances that involves the collection of new data could easily turn into a full-scale investigation of the program aimed at determining if what was presumed in the theory actually happened. And, indeed, an empirical "theory testing" study is one obvious approach to assessing program theory, an approach that emphasizes descriptive and predictive accuracy (see, e.g., Bickman, 1990; also, Exhibit 5-L gives an example). As later chapters in this volume will discuss, many aspects of the evaluation of program process and impact take on this character when the evaluation design is guided by a detailed analysis of program theory, as we advocated in Chapter 3.

In this chapter, however, our focus is on the task of assessing the soundness of the program theory description as a plan, that is, as a statement of the program as intended rather than as a statement of what is actually happening (that assessment comes later). By recognizing the role of observation and interview in the process, therefore, we are not suggesting that theory assessment as a distinct evaluation activity necessarily requires a full evaluation of the program. Instead, we are suggesting that some appropriately configured contact with the program activities, target population, and re-

EXHIBIT 5-L Testing a Model of Patient Education for Diabetes Self-Care Management

The daily management of diabetes involves a complex interaction of metabolic variables, self-care behaviors, and psychological and social adjustments to having the disease. An important component of treatment for diabetes, therefore, is the instruction of patients so that they have the skills and knowledge required to do their part. A team of university medical researchers with a particular interest in the personal meaning to patients of having diabetes formulated an impact theory for the effects of patient education, which they diagramed as follows:



The researchers investigated this model by examining the correlations representing some of the key hypotheszed relationships on survey data collected from a sample of 220 people with diabetes recruited from clinics in several states. The data were analyzed using a structural equation analysis which showed only an approximate fit to the model. The relationships between the "personal meaning of diabetes" variables and "psychosocial adaptation" were strong, as were those between knowledge and self-care behavior. However, other relationships in the model were eqivocal. The researchers' conclusion: "While the results showed that the data did not fit the proposed model well enough to allow for definitive conclusions, the results are generally supportive of the original hypothesis that the personal meaning of diabetes is an important element in the daily management of diabetes and the psychosocial adjustment to the disease."

SOURCE: Adapted from George A. Nowacek, Patrick M. O'Malley, Robert A. Anderson, and Fredrick E. Richards, "Testing a Model of Diabetes Self-Care Management: A Causal Model Analysis With LISREL," *Evaluation & the Health Professions*, 1990, 13(3):298-314.

lated situations and informants can provide the evaluator with valuable information about how plausible and realistic the program theory is.

Outcomes and Responses to the Results of Program Theory Assessment

A program whose conceptualization is weak or faulty has little prospect for success even if it adequately operationalizes that conceptualization. Thus, if the program theory is not sound, there is little justification for attempting to assess other evaluation issues, such as program implementation, impact, or efficiency. Within the framework of evaluability assessment, finding that the program theory is poorly defined or seriously flawed indicates that the program is not evaluable.

When assessment of program theory reveals deficiencies in the program theory, one appropriate response is for the responsible parties to redesign the program. This would involve carefully working out a well-defined impact theory, program process theory, or whatever components of those theories are deficient. Program reconceptualization may include (a) clarification of goals and objectives and identification of the observable implications of attaining them; (b) restructuring program components for which the intended activities are not happening, are not needed, or are not reasonable; and (c) working with stakeholders to obtain consensus on program objectives and the logic that connects program activities with the desired outcomes. The evaluator may help in this process as a consultant.

If an evaluation of program process or impact goes forward without articulation of a well-defined and credible program theory, then a certain amount of ambiguity will be inherent in the results. This ambiguity is potentially twofold. First, if program process theory is not well defined and explicit, there is ambiguity about what the program is expected to be doing operationally. This complicates the matter of identifying criteria by which to judge how well it is implemented. Such criteria must then be established individually for the various key program functions through some piecemeal process. For instance, administrative criteria may be stipulated regarding the number of clients to serve, the amount of service to provide, and the like, but they will not be integrated into an overall plan for the program.

The second form of ambiguity is introduced into an evaluation when there is no adequate specification of the program impact theory. Impact evaluation may be able to determine whether certain outcomes were produced (see Chapters 7-10), but without a guiding theory it will be difficult to explain why or—often more important—why not. Poorly specified impact theory limits the ability to identify or measure the intervening variables on which the outcomes may depend and, correspondingly, the ability to explain what went right or wrong in producing the expected outcomes. If program process theory is also poorly specified, it will not even be possible to describe very distinctly the nature of the program that produced, or failed to produce, the outcomes of interest. Evaluation under these circumstances is often referred to as *black box evaluation* to indicate that assessment of outcomes is made without much insight into what is causing those outcomes.

If program theory is well defined and well justified, the articulation of that theory permits ready identification of critical program functions and effects and defines what the program is expected to do and what is supposed to happen as a result. This structure provides meaningful benchmarks against which actual program performance can be compared by both managers and evaluators. The framework of program theory, therefore, gives the program a blueprint for effective management and gives the evaluator guidance in designing process, impact, and efficiency evaluations, as the subsequent chapters in this volume will discuss.

SUMMARY

- Every program embodies a program theory, a set of assumptions and expectations that constitute the logic or plan of the program and provide the rationale for what the program does and why. These assumptions may be well formulated and explicitly stated, representing an *articulated* program theory, or they may be inherent in the program but not overtly stated, thus comprising an *implicit* program theory.
- Program theory is an aspect of a program that can be evaluated in its own right. Such assessment is important because a program based on weak or faulty conceptualization has little prospect of achieving the intended results.
- The most fully developed approaches to evaluating program theory occur in the context of *evaluability assessment*, a preevaluation appraisal of whether a program's performance can be evaluated and, if so, whether it should be.
- Evaluability assessment involves describing program goals and objectives, assessing whether the program is well enough conceptualized to be evaluable, and identifying stakeholder interest in evaluation findings. Evaluability assessment may result in efforts by program managers to better conceptualize their program. It may indicate that the program is too poorly defined for evaluation or that there is little likelihood that the findings will be used. Alternatively, it could find that the program theory is well defined and plausible, that evaluation findings will likely be used, and that a meaningful evaluation could be done.
- To assess program theory, it is necessary for the evaluator to articulate the theory, that is, state it in a clear, explicit form acceptable to stakeholders. The aim of this effort is to describe the "program as intended" and its rationale, not the program as it actually is, although, of course, some resemblance is expected.
- The evaluator describes program theory by collating and integrating information from program documents, interviews with program personnel and other stakeholders, and observations of program activities. It is especially important that clear, concrete statements of the program's goals and objectives be formulated as well as an account of how the desired outcomes are expected to result from program action. Also, the relationships expected among program functions, components, and activities must be described.

- The most important assessment of program theory the evaluator can make is based on a comparison of the intervention specified in the program theory with the social needs the program is expected to address. Examining critical details of the program conceptualization and the social problem indicates whether the program represents a reasonable plan for ameliorating the target problem. This analysis is facilitated when a needs assessment has been conducted to systematically diagnose the problematic social conditions (Chapter 4).
- A complementary approach to assessing program theory uses stakeholders and other informants to appraise the clarity, plausibility, feasibility, and appropriateness of the program theory as formulated. This review can often be usefully supplemented with direct observations by the evaluator to further probe critical assumptions in the program theory.
- Program theory also can be assessed in relation to the support for its critical assumptions found in research or documented program practice elsewhere. Sometimes findings are available for similar programs, or programs based on similar theory, so that an overall comparison can be made between a program's theory and relevant evidence. If the research and practice literature does not support overall comparisons, however, evidence bearing on specific key relationships assumed in the program theory may still be obtainable.
- Assessment of program theory yields findings that can help improve the conceptualization of a program or, possibly, affirm its basic design. Such findings are an important evaluation product in their own right and can be informative for program stakeholders. In addition, a sound program theory provides a basis for evaluation of how well that theory is implemented, what outcomes are produced, and how efficiently they are produced, topics to be discussed in subsequent chapters of this volume.