1. Method and Theory in Program Evaluation: A Question of Balance

THEORY IN PROGRAM EVALUATION: A NEGLECTED ISSUE

Theory is a frame of reference that helps humans to understand their world and to function in it. Theory is crucial in research (e.g., Rudner, 1966). Theory provides not only guidelines for analyzing a phenomenon but also a scheme for understanding the significance of research findings. Traditionally, however, theory has been neglected in the discipline of program evaluation. Until very recently, evaluation literature has rarely been concerned with the importance of theory in evaluating a program or with how to incorporate theory into evaluation processes. Influential evaluation books such as Guttentag and Struening's (1975) *Handbook of Evaluation Research* and Reicken and Boruch's (1974) *Social Experimentation* hardly focused upon or discussed the concepts and implications of theory. The *Evaluation Thesaurus* (Scriven, 1981) and the section on "Evaluation of Programs" in *Encyclopedia of Educational Research* (Talmage, 1982) have gathered the concepts and terminology commonly used in program evaluation, yet theory or related concepts are not included in these collections.

Furthermore, Lipsey et al. (1985) intensively reviewed a sample of 175 evaluation studies from a variety of journals. They found that most

1. Method and Theory in Program Evaluation: A Question of Balance

THEORY IN PROGRAM EVALUATION: A NEGLECTED ISSUE

Theory is a frame of reference that helps humans to understand their world and to function in it. Theory is crucial in research (e.g., Rudner, 1966). Theory provides not only guidelines for analyzing a phenomenon but also a scheme for understanding the significance of research findings. Traditionally, however, theory has been neglected in the discipline of program evaluation. Until very recently, evaluation literature has rarely been concerned with the importance of theory in evaluating a program or with how to incorporate theory into evaluation processes. Influential evaluation books such as Guttentag and Struening's (1975) *Handbook of Evaluation Research* and Reicken and Boruch's (1974) *Social Experimentation* hardly focused upon or discussed the concepts and implications of theory. The *Evaluation Thesaurus* (Scriven, 1981) and the section on "Evaluation of Programs" in *Encyclopedia of Educational Research* (Talmage, 1982) have gathered the concepts and terminology commonly used in program evaluation, yet theory or related concepts are not included in these collections.

Furthermore, Lipsey et al. (1985) intensively reviewed a sample of 175 evaluation studies from a variety of journals. They found that most

of those studies failed to integrate a prior theory into evaluations in terms of formulating program elements, rationale, and causal linkages. Lipsey et al. (1985) also noted that this lack of theory development appeared to be widespread throughout the evaluation community regardless of the authors' academic affiliations or program areas. The insensitivity to theoretical questions and issues and the prevalence of input/output types of evaluation has led Wortman (1983, p. 224) to comment that "program evaluation is a multi-disciplinary and (unfortunately) a largely atheoretical activity."

Chen and Rossi (1983) argue that, at the extreme, an atheoretical approach to evaluation is characterized by adherence to a step-by-step cookbook method of doing evaluations. Program evaluation in this situation can become a set of predetermined research steps that are uniformly and mechanically applied to various programs without concern for the theoretical implications of program content, setting, participants, implementing organizations, and so on.

The atheoretical view tends to result in a simple input/output or black box type of evaluation (Lipsey, 1987; Bickman, 1987b; Chen and Rossi, 1983). This type of evaluation is characterized by a primary focus on the overall relationship between the inputs and outputs of a program without concern for the transformation processes in the middle. Such simple input/output or black box evaluations may provide a gross assessment of whether or not a program works but fail to identify the underlying causal mechanisms that generate the treatment effects, thus failing to pinpoint the deficiencies of the program for future program improvement or development. A black box evaluation is usually not sensitive to the political and organizational contexts of input and output, and it neglects issues such as the relationship between the delivered treatment and the planned treatment, between official goals and operative goals, or between intended and unintended effects.

Whatever the results, the simple input/output or black box type of evaluation often will generate conclusions that are less than satisfactory. On one hand, an unqualified claim of program success that is based upon a black box evaluation may be difficult for policymakers or practitioners to apply. To use a medical example, if a black box evaluation shows a new drug to be capable of curing a disease without providing information on the underlying mechanisms of that cure, physicians will have difficulty in prescribing the new drug because the conditions under which the drug will work and the likelihood of negative side effects will not be known.

On the other hand, a claim of program failure based upon a black box evaluation may be even more misleading. Does failure imply that the theory on which the program is based is incorrect? Or is the failure due to a problem with implementation? Or is the strength of treatment too low? Or is the measurement of the treatment effect not sensitive enough? And so on. The finding of program failure in black box evaluations is vague and ambiguous. At most, little information is provided to assist in improving the program. At times, true program effects may be concealed by a gross estimation of the treatment effect in the black box evaluation, as demonstrated in the TARP Experiment (Rossi, Berk, and Lenihan, 1980) described later in this book.

Mitroff and Bonoma (1978) also argue for the importance of specifying prior theoretical assumptions in program evaluation in order to meet policy needs. They object to the belief held by some proponents of the experimental paradigm that overcoming methodological difficulties alone can make the data collected in an evaluation valid and scientifically precise without consideration of why the data were collected. They argue that "validity cannot be even approached until one learns to question his or her assumptions as closely as he or she questions the rigor with which data was [*sic*] generated" (1978, p. 256). Because scientific data can only be unearthed with the prior assumption of theory, Mitroff and Bonoma believe that evaluation requires a level of openness and flexibility in methodology that allows for inquiry into the background assumptions of the program and the data collection process.

Similarly, Wholey (1979, 1987) insisted that a program is not ready to be evaluated unless the theoretical basis of the program has been developed and carried out. An evaluation, according to Wholey (1979, 1987), is less likely to be useful under the conditions of unclear program objectives, lack of testable assumptions linking program components, and lack of sufficient resources and implementation efforts.

HISTORICAL DEVELOPMENTS AND THE NEGLECT OF THEORY IN EVALUATION

An intriguing question to raise is this: Why has program evaluation developed as an atheoretical activity? Taking the viewpoint of the philosophy of science (Kuhn, 1970), Shadish and Reichardt (1987a) argue that action and practice tend to precede theory development in any discipline. Program evaluation also falls within this trend. However, there

are other interwoven reasons related to the historical conceptualization and focus of program evaluation that have particularly oriented program evaluation toward an atheoretical stance.

In the very early efforts to build program evaluation as a discipline, scientific research methods were greatly emphasized in many pioneer works in their attempts to define and conceptualize program evaluation. Perhaps this helped to promote evaluation as a new science that was clearly distinct from mere casual or arbitrary judgments in assessing program worthiness. For example, Suchman (1967, p. 7) viewed evaluation as "the utilization of scientific research methods and techniques for the purpose of an evaluation." Similarly, Scriven (1967, p. 40-41) defined evaluation as "a methodological activity which combines performance data with a goal scale." To underline the scientific status of this new discipline, the application of the scientific method has been emphasized. This emphasis is characteristic of what has been given the traditional and common label of the discipline of "evaluation research" (e.g., Weiss, 1972; Caro, 1977; Guttentag and Struening, 1975). Yet with such a great emphasis on research methods in conceptualizing and defining program evaluation, the implications of program theory tended to be ignored.

The focus on methodological issues has been further reinforced by the fact that, in spite of their many important contributions, the major evaluation perspectives have mainly been method oriented. For proponents of the experimental paradigm (e.g., see Campbell and Stanley, 1963; Reicken and Boruch, 1974; Cook and Campbell, 1979), an evaluation is best carried out by exactly following the classic randomized experimental design. In the classic experimental design, the treatment is manipulated in a controlled setting where subjects are randomly assigned to the experimental and control groups, variables are objectively measured, and experimental results are precisely analyzed with rigorous statistical methods. The merits of other designs or methods, such as quasi-experiments or preexperimental designs, are sometimes judged by the degree to which they approximate the experimental design.

In contrast, advocates of naturalistic approaches propose that qualitative or ethnographic methods serve best in doing an evaluation. According to Lincoln and Guba (1985), Guba and Lincoln (1981), and Patton (1980), evaluations are best carried out with minimum constraints on the antecedent conditions (independent variables) and the outputs of a research inquiry. They propose that qualitative or ethnographic investigations are superior to the more structured approaches usually

involved in quantitative studies because the research format of naturalistic inquiry can be relatively free or fluid.

In the naturalistic approach, sensitive observers record behavior in "natural settings," and researchers analyze the resulting protocols with due regard for the humanity of the subjects and the desires of program operators, administrators, and other stakeholders. The full complexity of human behavior is thus given due attention, free of the constraints imposed by the research designs of the positivistic approaches.

Although the major concern in Cronbach's approach (Cronbach et al., 1980; Cronbach, 1982) is to incorporate program evaluation into the context of political and organizational processes to facilitate pluralistic decision making, his approach is closely related to qualitative methods. Cronbach maintains that flexible, qualitative methods are useful both in achieving the generalizability of evaluation results and in serving to enlighten program stakeholders.

Economists have developed sophisticated modeling processes in their evaluations of large-scale economics-related programs (e.g., Ferber and Hirsch, 1982; Hausman and Wise, 1985), but the major contribution of economists to the mainstream of program evaluation has been their introduction of econometric methods. The methodology of evaluation has been expanded considerably by the economists' demonstration that econometric methods are useful in dealing with nonexperimental data (e.g., Heckman, et al., 1987; Barnow et al., 1980) and by their intensive debates on the relative merits of econometric versus experimental methods.[1]

These major evaluation perspectives have made many important contributions to the development of program evaluation. However, in relying so heavily upon research methods as cornerstones of their approaches, they have also contributed to the traditional emphasis upon methodological and research issues to the neglect of program theory in doing evaluations.

The popularity of method-oriented evaluations has been further strengthened by the long and intensive debates between qualitative and quantitative camps about which research method is most suitable for evaluations (e.g., see Cook and Reichardt, 1979; Lincoln and Guba, 1985). On one hand, advocates of the experimental paradigm (e.g., Cook and Campbell, 1979) insist that rigorous methods such as randomized experiments—or at least strong quasi-experiments—should be used to assess program effectiveness. They believe that the use of preexperimental

designs or qualitative methods tends to provide misleading information that confuses the decision-making process. Economists agree with the basic tenet of the experimental paradigm that internal validity is crucial in an evaluation, yet they disagree among themselves as to whether econometric methods are as effective as experimental methods in ensuring an unbiased estimate of treatment effect.

On the other hand, naturalists such as Lincoln and Guba (1985) and Patton (1980) argue that the rigidity of rigorous methods tends to alienate evaluators from stakeholders (those persons most directly affected by evaluation results) and prevents evaluators from understanding the stakeholders' perspectives and concerns and/or program operation processes. Similarly, Cronbach (1982) argues that the use of rigorous methods not only makes an evaluation rigid and narrow in scope but also exhausts scarce resources in dealing with trivial issues. Cronbach and many naturalists believe that flexible qualitative methods, rather than quantitative methods, best serve the needs of an evaluation. Guba and Lincoln (1981) and Lincoln and Guba (1985) even predict that in the future the naturalistic paradigm will gradually replace the experimental paradigm in program evaluation.

This conflict between the experimental and qualitative camps was dramatized in the debate between Robert F. Boruch and Milbrey W. McLaughlin (Davis, 1982). The controversy was concerned with the recommendation to Congress and the Department of Education that the federal government should mandate the use of experimental methods whenever appropriate. The debate was held at the 1981 annual meeting of the Evaluation Network and the Evaluation Research Society in Austin, Texas (Davis, 1982).

In the debate, Boruch (Davis, 1982, p. 12) argued that "if one wants statistically unbiased estimates of effects, then well designed field tests are warranted." He advocated that, if the experimental methods he recommended were implemented at federal or state levels, the quality of evidence of an evaluation would improve. This would eliminate the experience of many programs where real experiments had to be conducted after poor quasi-experiments or preexperimental research designs failed to evaluate the program appropriately.

However, McLaughlin (Davis, 1982, p. 14) argued just the opposite. She pointed out that the rigidity of experimental methods tends to lead an evaluator to "ask the wrong questions, use the wrong measures, and fails to provide information that validly informs policy and practice." Problems such as rigidity, high cost, and poor fit to the local program

operations and program changes can make the use of experimental methods a waste of time and money in evaluation efforts.

The intensive conflicts and debates on research methods have generated information that is interesting and insightful for methodological development. However, this discussion may also create the impression that many or most problems in evaluation result mainly from methodological shortcomings and that the refinement of research methods alone will lead to the solution of many difficulties and problems in program evaluation.

FOCI AND PROBLEMS OF METHOD-ORIENTED EVALUATIONS

Approaching an evaluation from the viewpoint of a research tradition provides the advocates with a shared view of the aims of program evaluation as well as a set of established techniques and procedures to carry it out. The use of a particular method allows the advocates of that perspective to intensively explore and deal with the particular evaluation issues with which they are primarily concerned. For example, the emphasis upon experimental designs and the development of various quasi-experimental designs not only has enhanced our understanding of issues relating to internal validity but has also helped prevent unsound and careless work in evaluation (e.g., Campbell and Stanley, 1963; Cook and Campbell, 1979). Campbell and his associates' checklist of threats to validity demystifies the sources of bias in applying a particular design in an evaluation and indicates the potential impact of these biases on the assessment of treatment effects.

Similarly, naturalistic approaches based upon qualitative methodology have helped us to explore and develop techniques to better understand multiple stakeholders'—especially program managers' and administrators'—needs and concerns. The use of qualitative methods provides evaluators with rich, firsthand information on questions such as how a program is implemented, what the patterns of interaction between stakeholders are, the kind of day-to-day problems that are confronted by program staff, and so on. This type of inquiry allows naturalists to work closely with stakeholders and to provide the timely information they need.

However, the overemphasis of both experimental and naturalistic perspectives on methodological issues also tends to narrow their focus.

These perspectives each tend to focus upon the area of program evaluation that corresponds to the strength of their methods. Other areas of program evaluation usually receive little attention and are sometimes argued to be less important. For example, the experimental paradigm (Campbell and Stanley, 1963; Cook and Campbell, 1979) deals mainly with issues relating only to internal validity, while other evaluation issues are given less priority by this perspective. Naturalistic approaches (e.g., Guba and Lincoln, 1981) mainly emphasize issues related to process evaluation, while the issues of outcome evaluation are not regarded from this perspective as very useful for stakeholders. Cronbach's (1982) approach places higher priority on achieving external validity, and issues of internal validity are regarded by him as trivial.

Currently, however, there is a growing consensus among evaluators that an evaluation must deal with multiple values and issues (e.g., Cook and Shadish, 1986). Adherence to a particular method prevents evaluators from developing strategies for dealing with more than a narrow range of issues. As a consequence, the traditional perspectives tend to be limited in scope and encounter difficulties in evaluation situations that require dealing with multiple issues.

Another problem of the individual method-oriented perspectives is that they are difficult to connect and seldom communicate with each other. It is well documented in research methods texts that each research method has its own strengths and weaknesses (e.g., see Babbie, 1986). Because this is the case, it is easy for an advocate of one particular research tradition to highlight only its strengths and attack other research traditions on their weaknesses. Accordingly, arguments for the replacement of one research method by another tend to generate only continuing debates, to further polarize proponents' positions, and to confuse the audience. An overly rigid adherence to any particular research method results in heightened differences that fuel continuing difficulties in communication.

The excessive advocacy of any one method might result in the exaggeration of that method's strengths and blindness to its weaknesses. For example, naturalists such as Parlett and Hamilton (1978) and Guba and Lincoln (1981) totally reject any merits of experimental or quasi-experimental methods. In the end, a "competition of research methods" may evolve that will hinder the development of a more comprehensive conceptual framework within which it would be possible for all research methods to make even greater contributions.

In spite of such debates, most evaluation studies still follow the framework provided by the experimental paradigm (Lipsey et al., 1985). As is reported from time to time, however, applications of the experimental paradigm in program evaluation are not always satisfactory. Among these complaints, Weiss and Rein (1969) report that using an experimental design to evaluate a broadly aimed and highly fluid program may lead to misleading and artificial results. Guttentag (1973) illustrates many difficulties involved in implementing experimental designs in the field, and Deutscher (1977) indicates that the inflexible structure of experimental methods can interfere with the detection of real program effects. Cronbach (1982) argues that evaluations that follow the experimental paradigm tend to focus on trivial issues that are not very useful for policy decisions. Evaluation studies following experimental designs have been found to have difficulties in maintaining the integrity of the design and have deficiencies in various methodological issues such as treatment integrity, statistical power, and the like (Lipsey et al., 1985).

Furthermore, the major purpose for doing evaluation is to provide timely and relevant feedback information for policy making. However, utilization studies indicate that many evaluation results are not used by decision makers (e.g., Weiss, 1977). Program stakeholders frequently report that evaluation studies fail to provide them with relevant and useful information (Chelimsky, 1977).

It might be thought that the naturalistic approaches are the perfect answer to this problem and that naturalistic evaluations would be popular. Judging from the evaluations reported in the major evaluation journals, however, and despite the optimism expressed by Lincoln and Guba (1985), the application of naturalistic approaches in program evaluations has been relatively limited in comparison with the experimental paradigm (Lipsey et al., 1985).

Part of the problem may be that the naturalistic approaches have not yet clearly demonstrated an ability to generate valid and generalizable information (Chen, 1988). Sadler (1981) points out that an evaluator's personal observations and intuitive analyses can result in biases in the naturalistic evaluation. Furthermore, Williams (1986b) points out that the application of naturalistic inquiries in the field may make it necessary to compromise evaluation standards and criteria. For example, naturalistic evaluators may insist on using an unstructured and inductive format in conducting an evaluation, while stakeholders want to know

from the beginning of the evaluation what is going to be done, how, and by whom.

With the growing awareness of the problems and difficulties associated with the traditional perspectives, there has been a concurrent interest in taking a more pragmatic view of research methods. Unlike hard-line naturalists such as Lincoln and Guba (1985), some naturalistic evaluators such as Williams (1986a) and Smith (1986) do not zealously reject quantitative methods; neither do they predict and advocate a new research era dominated by naturalistic approaches. They are willing to admit that both qualitative and quantitative methods have their strengths and weaknesses. As Williams (1986a, p. 85) suggests: "No single inquiry method will suffice for all evaluation needs."

Williams (1986a) does not believe that the naturalistic approach should be used blindly on every occasion. He suggests that naturalistic approaches may be most suitable under the following conditions: where the issues of the evaluation cannot be clearly defined before the evaluation begins, when a large group of stakeholders' informational needs and values must be dealt with, when the evaluation audience requires information about program processes, and so on.

Smith (1986) advocates the use of a combination of both qualitative and quantitative methods when the situation allows it. Smith admits that combining the qualitative and quantitative approaches can be expensive and requires not only the combination of many skills but also the accommodation of divergent viewpoints within the evaluation team. Nonetheless, she argues that the benefits outweigh the costs. Smith lists four circumstances under which a combined approach might prove most fruitful: when a complete description of the evaluation is necessary, when circumstances indicate that the results of a qualitative study can be generalized, when a combination of methods might enhance validity, and when qualitative feedback might be effective in influencing stakeholders' opinions. Smith (1986) believes the use of multiple methods can combine the best of the qualitative and quantitative worlds in one evaluation.

The idea of using multiple methods to overcome the shortcomings of any single research method is also shared by some evaluators from the quantitative camp (e.g., see Mark and Shotland, 1987). However, despite the potential advantages, the advocacy of multiple methods also presents some shortcomings of its own.

Reichardt and Cook (1979) point out factors that may prevent the use of multiple methods, such as the greater time and costs involved,

lack of training programs for combined methodologies, and the persisting rivalry between research traditions. Furthermore, Shotland and Mark (1987) point out that evaluation results generated by the use of multiple methods may be difficult to interpret because different methods may address different questions, may generate conflicting results, and could also suffer the same inadequacies.

Furthermore, the simple advocacy of multiple methods alone is not adequate for providing guidance in actual practice because qualitative and quantitative methods each have their own unique assumptions, logic, and research procedures. There are no self-evident logical connections between these opposing methods. This is perhaps the reason why the advocates of one method-oriented perspective tend to debate instead of work with the proponents of other method-oriented perspectives (e.g., see Cook and Campbell, 1979; Guba and Lincoln, 1981). In fact, any suggestion that the assumptions, logic, and research procedures of a specific method be modified tends to threaten the existence and functioning of that method-oriented perspective. Without efforts toward conceptual integration, the advocacy of multiple methods may simply be expedient and similar to advocating a shotgun marriage.

Because each method, or even multiple methods, involves its own strengths and weaknesses, there realistically is no one best method for evaluation that can universally apply to every evaluation situation. Which method or methods should be used in an evaluation may be contingent upon external factors such as evaluation purposes, the maturity of the program, the availability of time and resources, stakeholders' and evaluators' values, and the political and organizational environments of a program. For example, where a program is still in the development stage, the issue of program effectiveness is too early to be judged and the randomized experiment may not be very useful for providing information for program improvement. Similarly, qualitative methods may not be useful when a program involves a large number of stakeholder groups and each has differing values and views on the program's purposes. For program management purposes, decision makers want to specify a set of goal dimensions of the program that clearly represent these stakeholders' views and values. Under this condition, quantitative methods may be more useful than qualitative methods in uncovering the underlying goal dimensions among a large number of stakeholders.

If the appropriateness of a research method or methods for any given evaluation can only be judged within a specific context, then without linking the evaluation process to the context, further efforts to advance

research methods alone may not appreciably expand the focus and scope of program evaluation. The refinement of research methods is helpful, but what is most needed in the future for advancing program evaluation may be conceptual and theoretical efforts to systematically integrate these contextual factors and research methods. This is where incorporating program theory into evaluation processes becomes crucial. Program theory can provide guidelines for identifying which issues are most important in an evaluation, determining what method or methods are most relevant to address these issues, and suggesting how to apply the best method or methods for dealing with these issues. These kinds of concerns will be the fundamental force behind the new movement toward theory-driven evaluations.

THE MOVEMENT TOWARD THEORY-DRIVEN EVALUATIONS

Currently, there is a new movement to shift program evaluation from method-oriented evaluations to theory-oriented evaluations. Lipsey (1987) argues that the traditional method-oriented or black box type of evaluation underrepresents the complexities of the treatment circumstances. He argues that, in reality, the treatment can deviate greatly from a simple dichotomously coded group membership, that exogenous variables can accompany or interact with the treatment, and that a broader range of outputs can be produced. Lipsey (1987) urges the development of a theoretical framework that will differentiate more richly the details of causal processes that can serve as a basis for planning and organizing evaluation activities.

Similarly, Trochim (1986b) has also criticized the mechanical application of randomized experiments and quasi-experiments as the primary means of assessing program effectiveness. Trochim (1986b, p. 3) argues that "this *ceteris paribus* mentality is inherently atheoretical and noncontextual. It assumes that the same mechanism works in basically the same way whether we apply it in mental health or criminal justice, income maintenance or education." Trochim believes that the causal mechanisms of a program should be examined within the framework of the program's theory.

Cordray (1986) argues that the traditional input-output assessment leads an evaluator to provide an improverished version of causal inference. Instead, he proposes that evaluation should broaden the evidential basis by actively considering plausible rival explanations, by examining

implementation procedures, and by investigating mediating and contextual factors.

Chen and Rossi (1987) argue that method-driven evaluations tend to maximize one type of validity at the expense of others. To avoid this problem, Chen and Rossi (1987) point out the importance of program theory in simultaneously dealing with various types of validity.

The growing emphasis on the importance of program theory is also evidenced in some recent publications, such as the 1986 volume (edited by Cordray and Lipsey) and the 1987 volume (edited by Shadish and Reichardt) of *Evaluation Studies Review Annual* in which program theory is one of the major themes. Furthermore, a volume devoted to program theory in *New Directions for Program Evaluation* has been edited by Bickman (1987a). In this volume, Bickman (1987b) provides a list of benefits that can result from an articulation of program theory and its integration with program evaluation. Among other advantages, specifying the underlying theory of a program within the evaluation allows that theory to be tested in a way that reveals whether program failure results from implementation failure or theory failure. This will also help to clarify whether a program is being implemented under conditions in which it is appropriate. Program theory clarifies the connections between a program's operations and its effects, and thus helps the evaluator to find either positive or negative effects that otherwise might not be anticipated. It also can be used to specify intermediate effects of a program that might become evident and measurable before final outcomes can be manifested, which can provide opportunities for early program assessment in time for corrective action by program implementors. Finally, developing a program theory may be the best method of informing and educating stakeholders so that they can understand the limits of the program.

Given these positive functions, Bickman (1987b) is surprised that program theory has generally been slighted in the evaluation literature until recently and, in practice, has been largely ignored. In their chapters in the same volume, Conrad and Miller (1987), McClintock (1987), Scheirer (1987), Wholey (1987), and Shadish (1987) also share with Bickman (1987b) a desire to illustrate how incorporating program theory into an evaluation can enhance a program evaluator's sensitivity to planning, goal clarification, implementation, stakeholders' needs, and social change theories in general.

In addition, the current developments in this area are discussed in a special 1989 issue of *Evaluation and Program Planning* featuring the theory-driven approach. In this special issue, Chen and Rossi (forth-

coming) point out issues relevant to formulating and using theory-driven evaluations. Finney and Moos (forthcoming), Scott and Sechrest (forthcoming), and Palumbo and Oliver (forthcoming) discuss the implications of program theory in treatment processes and implementation. Costner (forthcoming), Patton (forthcoming), Lipsey and Pollard (forthcoming), Trochim (forthcoming), Shapiro (forthcoming), and Chen (forthcoming) provide alternative views and strategies for formulating program theory. Cordray (forthcoming) and Bickman (forthcoming) discuss potential problems in theory-driven evaluations and possible strategies to deal with them.

This new movement toward theory-driven evaluations is not meant to detract from the significant contribution of research methods; as in many disciplines, research methods are useful tools for obtaining empirical knowledge and verifying hypotheses. However, as will become clear in later chapters of this book, this new movement argues strongly that it is not appropriate to perceive program evaluation mainly as an array of methods and data collection techniques. As a discipline, program evaluation must emphasize and develop its own unique, systematic, and theoretically based body of knowledge. Instead of being treated as ends in themselves, methods should be considered to be the means for facilitating the development of knowledge. As Bickman (1987c, p. 1) argues, "Evaluation is often referred to as a practical science, but both as a practice and as a science it requires theory."

The movement toward theory-driven evaluations is related to several important past developments in program evaluation. First of all, as discussed in the previous sections, this new trend may result from the growing recognition that concentration on methods alone may not be sufficient either to solve the current problems and difficulties in evaluations or to further advance the field of program evaluation in the future (e.g., Lipsey, 1987; Bickman, 1987b).

The intensive debates and conflicts between qualitative and quantitative camps have indicated that there is no shortage of methods in program evaluation. In fact, there is an abundance of advanced and sophisticated qualitative and quantitative methods available (e.g., see Cook and Campbell, 1979; Guba and Lincoln, 1981). At this stage of evaluation development, the problems and discontent currently raging may result not so much from a lack of research methods as from a lack of comprehensive conceptual frameworks or theories to link or integrate evaluation activities.

The need for focusing on program theory for the future advancement of program evaluation has been illustrated in the current development of the postexperimental perspective (Cook, 1985). Due to dissatisfaction with the limitations and problems of the experimental paradigm, some of its original advocates are attempting to expand and transform their traditional framework into a more general and comprehensive perspective that, hopefully, can better adapt to the political aspects of program evaluation.

This postexperimental perspective of "critical multiplism" has been developed in the last few years by Cook and his associates (e.g., see Cook, 1985; Shadish, Cook, and Houts, 1986). Generally speaking, critical multiplism asserts that evaluators should plan to use multiple methodologies, investigate multiple issues, and consider the views of multiple stakeholders. Cook (1985) believes that critical multiplism has distinct advantages over the traditional experimental paradigm. He claims that this perspective reduces the possibility of misinterpretation, provides more comprehensive information for policy processes, and makes an evaluation more rational and conscious of its values.

In its early stages of development, however, critical multiplism raises a major difficulty that cannot be resolved methodologically. An evaluation is carried out within the constraints of available resources and obviously cannot pursue all of the multiple options that might possibly be studied. Choices and trade-offs among these options are necessary and inevitable. These trade-offs, as noted by Shadish and Cook (1986), require guiding principles or theories rather than methodologies. Critical multiplism has yet to develop these. Shadish et al. (1986) note that the development of theories or guiding principles is the most urgent challenge that the advocates of critical multiplism currently face.

While methodological advancements will continue in program evaluation, we seem to have reached a stage where theoretical rather than methodological efforts are most needed. Perhaps an emphasis on a balance between methods and theory in program evaluation can help evaluators in examining the basic assumptions and dilemmas of an evaluation, in facilitating the development of strategies to deal with trade-offs, and in encouraging expansion of the scope of program evaluation.

This new trend may also result from the wisdom and experience accumulated in the past few decades and from the realization that the conceptualization and assumptions made about social intervention or planned changes may be too simplistic. The earlier works by theorists

such as Campbell (1969) and Scriven (1967) promoted the view that the main purpose of an evaluation is to assess the overall effectiveness of a program. Based on this information on program effectiveness, decision makers then decide whether the program should continue or be canceled.

However, it has been found that social interventions usually do not work so predictably and that problems are not so straightforward. Planned changes are usually implemented in an incremental fashion (Lindblom and Cohen, 1979). Decision makers are mainly risk-avoiders who prefer to alleviate present problems rather than initiate precipitate changes. Because evaluation results are seldom applied to a go/no-go decision situation (Cronbach, 1982), the information provided from the traditional outcome evaluation tends not to be useful in decision-making processes.

Frustrated with the low utilization of evaluation results, Weiss (1972) has urged expansion of the scope of evaluation to include theoretical issues. Weiss noted that simple input/output evaluations provide no information on how and why a program worked or did not work and cannot identify which elements of the program "amalgam" are the essential ingredients for successful implementation. Weiss (1972) suggested that the utilization of evaluation results would be greatly enhanced if three basic elements were included: an analysis of the theoretical premises of the program; specification of the program process or linkages between inputs and program outcomes; and an analysis of the components of the program—which ones are the most effective—and possible alternative approaches that could enhance program effectiveness.

Another reason for this new movement may relate to the recognition of the need to understand how the treatment is implemented. The early conceptualization of program evaluation was heavily influenced by the laboratory experimental tradition in agricultural and other physical science research (e.g., Fisher, 1935). In these areas, because researchers often have full control of subjects and research conditions, the experimental treatment can usually be precisely manipulated. For example, in an agricultural assessment of the effect of a new fertilizer, the new fertilizer can be precisely distributed to plants in treatment areas, but not to those in control areas, by following the original plan.

However, a large body of implementation studies in intervention programs provide highly convincing evidence that program implementation is extremely complicated and difficult within the human service areas (e.g., see Williams and Elmore, 1976). For example, it has been

found that participating organizations may be receiving funds without committing themselves to implementing the program (e.g., McLaughlin, 1975). Even if it is implemented, the treatment may not be exactly the same as originally planned (e.g., Dobson and Cook, 1980). Furthermore, community political processes can easily hinder or deter the implementation of a program (e.g., Pressman and Wildavsky, 1973). Cooperative interorganizational relationships may be necessary for program implementation, but they may be so complicated that a program can hardly move ahead (Derthick, 1972). Also, with the discretion they have, implementors tend to proceed with a program according to their views and interests, and these are not necessarily consistent with the intentions of decision makers or program designers (Lipsky, 1980).

The difficulties and complications involved in implementation provide a serious challenge for evaluators to develop their own strategies and conceptual frameworks with which to deal with these problems, and they clearly cannot simply borrow from the traditions of physical science. Critics such as Sechrest and Redner (1979) and Leithwood and Montgomery (1980) have made a strong argument that evaluators should pay attention to implemented treatment rather than simply to planned treatment. A few studies that have examined this issue have revealed that the implemented treatment was not exactly identical with the planned treatment, regardless of the rigorousness of the designs used (e.g., Dobson and Cook, 1980). The traditional assumption is that, if the treatment is properly planned, then a coherent and proper implementation will follow. This assumption may have to change.

The revelation of the extent of the difficulties and complications present in implementation studies indicates that program stakeholders require considerable help from evaluators in order to improve program implementation. Argyris (1980) argues that the applicability of social research will be enhanced by understanding the ecological context within which the social action occurs. It is important for developing a new theoretical framework to integrate program implementation into the overall evaluation activity. Under this broad conceptual framework, evaluators not only can track the implementation and report on its progress, but they can also work jointly with program stakeholders to improve the program implementation in the course of their evaluation activities.

The need for an expansion of the conceptual framework from method-oriented to theory-oriented may also result from evaluators recognizing the importance of examining their evaluation activities from a broader

perspective (e.g., Chen and Rossi, 1980; Mitroff and Bonoma, 1978). The emergence of program evaluation as a discipline relates to the general trend of growing rationalization and accountability in modern society. However, Weber (1947) cautioned that it is important to see the distinction between two types of rationality: formal rationality and value or substantive rationality. Formal rationality concerns efficiency in attaining specific short-term goals, while value rationality refers to the substantive purposes and long-term ends of individuals and groups. Because formal rationality is impersonal and bureaucratic, Weber saw an element of irrationality connected with it. Where formal rationality leads to the narrow specification and segmentation of social activities, overall purposes and goals are lost sight of and the attainment of narrow, short-term goals often results in unforeseen and unintended consequences.

Weber's argument is relevant to the conceptualization of program evaluation. When it is conceptualized narrowly as simply the measurement of goal attainment, as is common within the experimental paradigm, the emphasis is mainly on formal rationality. In such a case, evaluators are concerned only with the efficiency of the treatment in attaining given goals. The question of whether or not the goals are appropriate for the effectiveness of the program, or whether these rational goals and procedures could lead to unintended consequences, might not be considered. Evaluators may sometimes be serving only bureaucratic interests and neglect the broader implications of the program for human needs and purposes from the perspective of other stakeholders. To avoid such problems, a new conceptual framework for evaluation should be concerned with value rationality and should provide more insights into the real purposes of a program and its implications for wider social interests.

Generally speaking, the current developments of program evaluation clearly indicate a need for a theoretical perspective that not only is comprehensive enough to be sensitive to important evaluation issues in areas such as program implementation, underlying causal mechanisms, treatment designs, and program outcomes but also is sophisticated enough to provide guidance in dealing with multiple or even conflicting options. These concerns provide a basis for the development of a conceptual framework for the theory-driven perspective, which is the main focus of the next three chapters.

Before discussing the theory-driven perspective, however, it is important to point out that this emphasis on theory in an evaluation does not represent a rejection of using appropriate research methods.

As will be made clear in the later chapters of this book, a theory-driven evaluation requires the use of appropriate methods for data collection and empirical verification. The theory-driven perspective developed in this book disagrees with the current major evaluation perspectives in regard to their focus and conceptualization of evaluation, but recognizes their important contributions for devising various useful techniques and tools for conducting evaluations. The theory-driven perspective may be viewed as an expansion of previous contributions to program evaluation made by the traditional perspectives. This expansion can provide an agenda for the systematic integration of theory and methods.

NOTE

1. Their disagreements are highlighted in two journals that have recently devoted special issues to reviewing the use of state-of-the-art econometric methods for estimating treatment effects: *Evaluation Review* (1987, Volume 2, Number 4) and *Journal of Human Resources* (1987, Volume 21, Number 6).