Shadish, W.R. Jr., Cook, T.D., & Leviton, L.C. (1991). Chapter 2: Good theory for social program evaluation. *Foundations of Program Evaluation: Theories of Practice*. Newbury Park, CA: Sage.

2

Good Theory for Social Program Evaluation

Judging the merits of evaluation theories requires specific description of the things such theories ought to do and the issues they ought to address competently. The linchpin of our argument is that the fundamental purpose of program evaluation theory is to specify *feasible practices that evaluators can use to construct knowledge of the value of social programs that can be used to ameliorate the social problems to which programs are relevant.* This description has five components—practice, knowledge, value, use, and social programming—and implies that an evaluation theory should have a knowledge base corresponding to each component.

This chapter describes these components in detail, but they are briefly described as follows. The *social programming* component concerns the nature of social programs and their role in social problem solving. It deals with the internal structure and functioning of programs, their relationship to other institutions, and the processes through which programs and their components can be changed to improve program performance. The *knowledge* component is concerned with what counts as acceptable knowledge about the object being evaluated, with methods to produce credible evidence, and with philosophical assumptions about the kinds of knowledge most worth studying. The *value* component concerns the role that values and the process of valuing play in evaluation. It deals with which values ought to be represented in an evaluation, and how to construct judgments of the worth of social programs. The *use* component concerns how social science information can be used in social policy and programming. It deals with possible kinds of use, relative weight to be given to each kind of use, and what evaluators can do to increase use. The *evaluation practice* component concerns the things evaluators do as they practice their profession. It deals with the role of evaluators in relating to program stakeholders; how to decide which questions to ask; where one gets questions from; and what methods to use given priorities among questions, the issues about which uncertainty is greatest, and constraints of time, financial resources,

staff skills, and procedural standards (Cook & Shadish, 1986; Shadish & Reichardt, 1987).

Every comprehensive evaluation theory will be better if it explicitly describes and justifies each of these five components. But the fifth component—evaluation practice—is most important because evaluators have to practice in a context where leisurely reflection about theoretical alternatives must yield to action within constraints. Even so, in using existing theories of practice, practitioners implicitly accept the assumptions built into them about social programming, knowledge construction, valuing, and use. It is better that this be explicit. Some theorists are explicit about particular components, such as Weiss (1977b) about use, Campbell (1977) about knowledge construction, and Scriven (1980) about valuing. Yet no theory we review deals explicitly and in detail with all five components. Hence we will describe and analyze the assumptions about social programming, use, valuing, knowledge construction, and evaluation practice that each theorist has made or assumed. We turn now, however, to an extended discussion of the five components, since they are our critical tool for analyzing evaluation theory. Each section begins with a summary table that is then discussed in detail.

THE SOCIAL PROGRAMMING COMPONENT OF EVALUATION THEORY

Program evaluation assumes that social problem solving can be improved by incremental improvements in existing programs, better design of new programs, or terminating bad programs and replacing them with better ones. If these conditions do not hold, evaluation cannot achieve its purpose. Theories of social programming, therefore, must show if and how these things can be done.

Based on our experience and some of our past writing (Cook, 1981; Cook et al., 1985; Shadish, 1987b), we can divide the relevant territory into three elements:

- internal program structure and functioning
- external constraints that shape and constrain programs
- how social change occurs, how programs change, and how program change contributes to social change

The internal structure of a program includes its staff, clients, resources, outcomes, administration, internal budget allocations, social norms, facilities, and internal organization. Internal structure also involves how these

TABLE 2.1 A Summary of the Social Programming Component

The issues: How can social programs contribute to social problem solving, and how can programs be improved in this task?

Knowledge bases: This component describes

(1) how programs are structured internally, what functions they fulfill, and how they operate;
(2) how the external context shapes and constrains the program; and
(3) how social change occurs, how programs change, and how program change contributes to social change.

A better theory of this component

(1) discusses all three elements;
(2) recognizes that
  (a) authority for starting, changing, and ending programs is diffuse,
  (b) heterogeneity results from implementing multiple programs under any social policy, multiple projects in each program, and multiple elements in projects, and
  (c) the value of evaluation depends on how well other problem-solving activities are done—clear definition of an important problem, generation of a bold set of potential solutions, faithful implementation of any particular solution, and targeted dissemination of information about results; and
(3) identifies key concepts or leverage points for improving program capacity to address social problems.

things are combined in a program model that relates inputs to activities to outputs (Chen & Rossi, 1987; Wholey, 1979). Knowing internal structure helps evaluators to ask questions about such matters as whether actual structure meets intended structure, or about strengths and weaknesses in ongoing structure and functioning.

Programs do not operate in a vacuum. They affect and are affected by other social, political, and economic institutions and activities. Among these are external funding sources, pressure from political constituencies and program stakeholders, availability of local resources such as transportation systems, and the political and economic values of society. Context plays an enormous role in shaping programs, particularly during implementation, as local interests change program design in ways that we are only beginning to understand (Pressman & Wildavsky, 1984). Even in older programs, administrators respond to contextual matters when considering which program improvements are desirable from political, economic, administrative, and logistical perspectives.

Knowing internal and external structures tells one little about changing them; in fact, structure can be so stable that conscious attention to change is required. Programs can change by introducing incremental improvements in small practices, by adopting or adapting demonstration projects that are more effective than existing ones, and by radical shifts in

values and priorities. Evaluation needs to understand these matters if it is to contribute to change in a purposeful way.

A theory can deal well or poorly with these three elements. Better theories should be more *comprehensive*, more *accurate in their content*, and should *prioritize* more acutely those things worth attending to. Is current evaluation theory about social programming comprehensive? We think not. Most evaluators attend more to internal structure and functioning of programs than to external context or to program change (Bickman, 1987). Evaluation practice can suffer as a result. The program itself may be well understood, but its role in society and policy may not be. For example, a problem in the adoption of Fairweather's (1980) Lodge—an effective project for treating persons with chronic mental disorders—was that the evaluation focused mostly on the Lodge rather than on the fit of the Lodge to its political-economic environment. Consequently, it has not been widely adopted relative to other settings for these patients (Shadish, 1984).

Better program theory also has descriptively accurate content. At a minimum, it must accommodate the following. First, in the United States, beginning, improving, or terminating social interventions occurs in a political system and economic marketplace in which authority to act is diffuse. Changes in social programs result from thousands of accumulated small inputs. No single authority can radically change a program. Second, social programs are characterized by considerable heterogeneity of local implementation. "Programs" (e.g., the Community Mental Health Center Program) are funding and regulatory umbrellas for diverse local "projects" (e.g., local mental health centers) that provide service. Projects are composed of service and administrative "elements" (e.g., intake, day treatment, psychotherapy, prevention). Homogeneously implemented programs, such as social security money that is distributed relatively uniformly across the nation or a public health program that delivers inoculations, are rare. But the diverse structure of most social programs, coupled with diffuse authority for action, produces heterogeneous program implementation even when central authorities want uniform implementation. Third, the worth of evaluation depends on how well other social problem-solving activities are executed—problem definition, solution generation, solution implementation, and dissemination of results. Successful completion of each activity is thwarted in a political system where problem definition is more a political than a technical exercise, where radical ideas about solutions occur less often than minor variations in the status quo, and where adopting successful solutions is dependent on practical feasibility and professional discretion as much as on efficacy (Cook et al., 1985; Shadish, 1984).

Good program theory also *prioritizes* by identifying and justifying key leverage points for improving programs. Wholey (1979), for example, says that programs can improve when managers are ready to manage for results; he assesses this readiness and helps those managers who meet this criterion. Rossi and Freeman (1985) say that there is no one key to meaningful change in social programs; rather, different evaluations will be useful at different stages of program development. Hence their recommendations depend on whether the target is an innovation, an adjustment to an existing program, or an established program. The justification for such priorities is crucial to making the case that the priorities make a real difference in social problem solving, and to suggesting which priorities are better and worse, because different authors have different priorities.

Does a better theory of social programming make a difference? "Good" program theory makes a real difference to evaluation theory and its practice. Consider, for example, the implicit program theory in Scriven's work. In discussing why a curriculum developer prepares a new book, Scriven (1972a) says, "He is presumably doing what he is doing because he judges that the material being presented in the existing curriculum is unsatisfactory" (p. 126). In Scriven's theory, society is as devoted to such rational problem solving as is Scriven himself:

> Business firms can't keep executives or factories when they know they are not doing good work and a society shouldn't have to retain textbooks, courses, teachers, and superintendents that do a poor job when a good performance is possible. The appropriate way to handle anxiety of this kind is by finding tasks for which a better prognosis is possible for the individuals whose positions or prestige are threatened. (p. 125)

Weiss (1981b) presents a different view that challenges whether Scriven gives a descriptively accurate account of how programs work. She faults evaluators who assume

> that organizations make decisions according to a rational model: define problems, generate options, search for information about the relative merits of each option, and then, on the basis of the information, make a choice. As our colleagues who study organizations tell us . . . this is a patently inaccurate view of how organizations work. When we implicitly adopt this as our underlying theory of organizations . . . we inevitably reach distorted conclusions. (pp. 25-26)

Scriven's theory predicts that the best new textbook would be chosen. If so, the evaluator designs the evaluation to find the best book. Weiss's theory says other motives also drive textbook adoption, including profit,

perks, desire to keep a used-book market viable, and personal contact with authors. Weiss would study these contextual factors more intently than would Scriven. The evaluator who follows Weiss's advice will produce an evaluation that is more realistic about how programs change.

Early evaluators assumed evaluation would identify solutions to be widely disseminated and uniformly adopted in local projects across the nation. But this proved difficult. For example, a "planned variations" evaluation was once done to estimate the relative efficacy of each of three types of Follow Through programs at multiple sites around the country. When evaluators examined actual program implementation, they found more variability within the three types than between them. Had evaluators known that programs are so heterogeneous, they might have planned for heterogeneous rather than homogeneous implementation, or focused on better describing and explaining implementation. Later evaluation theories took such considerations into account (Cook & Walberg, 1985), developing finer pictures in which heterogeneity plays a greater or lesser role depending on the political system in a country (some centrally planned systems might have more homogeneous implementation), the policy sector (heterogeneity is less an issue with SSI payments compared with education programs), and relevant practitioners (Public Health Service personnel are trained toward uniform implementation of medical technology).

For all these reasons, evaluation theory must deal with the historical and political origins of a program; its structure, governance, and funding; the ways it is implemented; its context; and available leverage for changing it. Otherwise, the theory is deficient and may lead evaluators to assume unrealistic or illogical links between social programs and problems (Bickman, 1987; Shadish, 1987b) or to assume mistakenly that certain parts of programs can be changed (Cook et al., 1985).

THE KNOWLEDGE COMPONENT OF EVALUATION THEORY

Why employ professional evaluators to study social programs? Why not rely on gossip, cronyism, newspaper reports, or lobbyists? Do evaluators offer different or better knowledge than what is already available? A theory of knowledge addresses such questions. Evaluators claim to provide knowledge that is especially worth having, often characterized as scientific. All evaluators seem to share this assumption and related ones—that conflicts between common sense and the systematic observations of evaluators should generally favor the latter, or that some methods for constructing knowledge are better than others. But much dispute lurks

Table 2.2 A Summary of the Knowledge Component

The issues: Is anything special about the knowledge evaluators construct, and how do they construct such knowledge?
Knowledge bases: This component describes
(1) ontology, the study of the ultimate nature of reality;
(2) epistemology, the study of the nature, origins, and limits of knowledge; and
(3) methodology, the study of techniques for constructing knowledge.
A better theory of this component
(1) addresses all three elements;
(2) recognizes that
  (a) no paradigm of knowledge construction is best because significant difficulties plague all epistemological and ontological approaches,
  (b) in methodology, all methods are not equally good for all tasks, so the task is to sort out strengths and weaknesses of methods for different purposes, and
  (c) no method is routinely feasible and unbiased, so no study is ever free of flaws;
(3) helps evaluators prioritize the kinds of knowledge to construct, how much uncertainty reduction is needed, and what methods to use given available tasks and resources.

behind this united front. We will find many disagreements in this book about the kinds of knowledge evaluators should construct, and how they should do so. Theory of knowledge helps us understand and resolve such disputes.

Theories of knowledge make three kinds of assumptions:

- *about ontology:* What is the nature of reality?
- *about epistemology:* What are the justifications for knowledge claims?
- *about methodology:* How do we construct knowledge?

Terms like *ontology* and *epistemology* bore many evaluators, because they conjure up images of sterile philosophical debates. Yet evaluators probably know more about theories of knowledge than about any other theories because of methodology. Evaluators care deeply about such matters as experimental and quasi-experimental design, assessment, case study methods, survey and sampling techniques, data analysis, and methods for question generation. They often know about these methods in detail—for example, that quasi-experimentation deals with interrupted time-series designs, regression discontinuity designs, and nonequivalent control group designs; and that variations on nonequivalent control group designs include nonequivalent dependent variables and reversed treatment implementations (Cook & Campbell, 1979). All evaluators must be familiar with such methods to do their work, which is why so many evaluation books are about methods. By comparison, most evaluators see debates about ontology and epistemology as tangential to their work. Yet many key

arguments in evaluation—like debates between Campbell and Cronbach about the kinds of causal knowledge that are possible and worthwhile—are as much epistemological and ontological as they are methodological. So most evaluators are already familiar with these topics in the implicit sense that they continually have to resolve such debates in their daily work.

Ontology concerns common and pervasive attributes of being, such as whether things we experience are real or not. The recent popularity of qualitative evaluation (Guba & Lincoln, 1981; Lincoln & Guba, 1985; Patton, 1980) acquainted many evaluators with ontology's more conspicuous issues. For instance, most evaluators know of Heisenberg's uncertainty principle: Measuring either the position or the motion of atomic particles precludes simultaneous accurate measurement of the other (Davies & Brown, 1986). Some theorists use such controversies to make radical constructivist claims that question the existence of reality (Lincoln & Guba, 1985). If reality did not exist, concepts such as causality and validity would have meanings that are radically different from those we now give them. Validity might be what each individual decides it is.

Epistemology is the study of the characteristics of and standards for knowledge. Knowledge, for example, can be about different topics. Knowledge about generalizability concerns extrapolating from specific observations to constructs, places, people, and times with similar and dissimilar characteristics. Causality concerns whether A caused B (causal inference) and how that happened (causal mediation). Construct validity speaks to the accuracy of labels for causes, outcomes, and the things in between. Evaluators also want to know about the certainty they can place in knowledge, and so require standards for what is to count as acceptable or exemplary knowledge. An example is Campbell and Stanley's (1963) standard that "*internal validity* is the *sine qua non*" (p. 5) of good research. Campbell and Stanley outline a host of standards to use in judging internal validity.

Better theories of knowledge comprehensively address their ontological, epistemological, and methodological assumptions. Few theorists do so for ontology and epistemology (Campbell, 1969, 1971, 1977, 1988; Guba & Lincoln, 1981; Lincoln & Guba, 1985). We prefer comprehensive theories to a narrow focus on methods. Narrow theories run two risks. First, by their silence they leave evaluators to their own wits in sorting out confusing ontological, epistemological, and methodological debates, increasing the likelihood that no resolutions occur. Second, they may commit sins of omission or commission, especially in methodology, as when narrow advocacy of one method limits the applicability of a theory.

Good knowledge theory also needs accurate content, and should probably take into account the following. First, in epistemology and ontology

there are many losers but few clear winners. Logical positivism probably counts in the loss column; even its proponents stopped advocating it by the 1940s (Meehl, 1986). But there is no consensus replacement for it (Brown, 1977; Machan, 1977; Morgan, 1983), since significant difficulties plague all epistemologies and ontologies. Any theorist who claims to have the answer is almost certainly wrong. Second, not all methods are equally good for all tasks. So it is folly to prescribe one method for all evaluations, and evaluation theory must sort out the relative strengths and weaknesses of different methods for specific tasks. For example, the original quasi-experimental evaluation of Head Start (Cicirelli & Associates, 1969) used a covariance analysis to adjust for selection biases between treatment and nonequivalent control groups. Campbell and Erlebacher (1970) point out why this is incorrect. Subsequent analysts (e.g., Rindskopf, 1981) have used linear structural modeling techniques to remedy some but not all of the problems (Cronbach, 1982a; LaLonde, 1985; Murnane, Newstead, & Olsen, 1985; Reichardt & Gollob, 1986). This sorting out of strengths and weaknesses is an essential theoretical job. Third, no method is routinely feasible and unbiased, so no study is ever free of flaws. Hence evaluation theories rely more on research syntheses or systematic programs of research than on single studies, hoping that many heterogeneous studies will minimize the biases due to any one study (Campbell, 1987a).

Good knowledge theory also identifies key priorities for knowledge construction. Early evaluation theorists, for example, thought it most important to assess program effects as opposed to, say, implementation or needs. But over the years this prioritization diminished. Similarly, early theorists placed a high premium on rigorous scientific standards for judging the certainty of results; later evaluators thought it more important to produce useful knowledge for policymakers. Finally, theorists differ in the methods they recommend. For example, Stake tells evaluators to use observation, inspection of records, and open-ended interviewing most of the time; Campbell prefers experimental methods. Both give epistemological reasons for their preferences. Good knowledge theory is partly about justifying such priorities.

Does good theory of knowledge make a difference? When Campbell and Stanley (1963) say that internal validity is the sine qua non of good research, they highlight the priority they assign to causal inference: "Internal validity is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatment make a difference in this specific instance?" (p. 5). They contrast this with external validity—whether the causal inference could be generalized to other populations, settings, treatments, and measures. They want to be certain first that

a program works before generalizing it to a place where it might have no effects or harmful ones. Campbell's thinking had great impact on evaluation in the 1960s, so his preference was widely adopted in evaluation.

Cronbach (1982a) directly challenges Campbell on this: "Evaluators need not and ought not sort inferences into a more honored category of causal statements and a less honored category of correlational statements" (p. 140). Cronbach argues "that internal validity is of only secondary concern to the evaluator" (1982a, p. 112). For him, relevance to decisions in circumstances not studied—one part of Campbell's external validity—is most important. Cronbach objects to the apparent triviality of research emphasizing internal validity, and resists applying the word *cause* to the results of such research at all:

> I consider it pointless to speak of causes when all that can be validly meant by reference to a cause in a particular instance is that, on one trial of a partially specified manipulation *t* under conditions A, B, and C, along with other conditions not named, phenomenon P was observed. To introduce the word *cause* seems pointless. Campbell's writings make internal validity a property of trivial, past-tense, and local statements. (1982a, p. 137)

Cronbach offers a different notion of cause: "Progress in causal knowledge consists partly in arriving gradually at fuller formulations" (1982a, p. 139) of relationships among variables, the nature of the manipulation, the conditions surrounding it, and the characteristics of outcomes. "The effect of an intervention depends on initial conditions, and . . . without a close analysis of those conditions one's experiment teaches nothing" (1982a, p. 128). Because it is so hard to specify all these conditions accurately, Cronbach saw little reason to place causality on a pedestal in the pantheon of evaluation gods.

Cronbach also questions the authoritativeness of the randomized experiment, pointing out that its key benefits often are thwarted in implementation. To make his point he assembled information from often obscure sources to criticize a favorite example of those advocating experiments, the Manhattan Bail Bond Experiment—an experimental group of subjects recommended for pretrial release without bail and a control group not recommended. Cronbach claims that the mystique of randomization made readers overlook serious flaws in the experiment's execution; in his opinion, citing it as an exemplar is justified only by overlooking the weakness (perhaps falsity) of the evidence it produced. Despite the severe biases Cronbach found, he agreed the study had an impact despite its flaws as an experiment: since "internal validity cannot be claimed for its main conclusions, that the randomized design added little force to the conclusions"

(1982a, p. 144). And "inferences from loose, biased, and poorly described contrasts loomed as large in reports and must have carried as much weight in the bail reform movement as those meeting Campbell's strict standard" (p. 149). Cronbach prefers to gather policy-relevant information by scientific methods as much as possible, but with randomization as one of many standards of science, not always the highest priority.

More extremely, Lincoln and Guba (1985, 1986) premise their radically different theory of evaluation almost entirely on arguments that there is no reality beyond what we each construct, so causality, generalizability, and truth have little useful meaning for them. They advocate nearly exclusive use of case study methods as best suited to capturing the idiosyncratic richness of each person's reality. If Lincoln and Guba's (1985) arguments are accurate, they have important implications for evaluation practice.

Theories of knowledge make another difference that is closely tied to practice. Methods with different strengths and weaknesses will often yield different or contradictory results in evaluations of a single program. Trend (1979) describes the contradictory outcomes of evaluations of a housing project yielded by quantitative and qualitative methods. The program was effective or ineffective, depending on how it was evaluated, leaving Trend to explain the discrepancy. The resolution of such common discrepancies is partly a theoretical matter concerning general strengths and weaknesses of methods.

In summary, if evaluators claim to construct knowledge that is special or authoritative, then a theory of knowledge construction must be explicit in evaluation theory. This component helps evaluators to place abstract epistemological debates into practical context, to reach conclusions about epistemological assumptions in their own work, to see connections between those assumptions and the methods they use, and to assess the worth of various methods for constructing particular kinds of knowledge.

THE VALUE COMPONENT OF EVALUATION THEORY

Early evaluators mostly ignored the role of values in evaluation—whether in terms of justice, equality, liberty, human rights, or anything else. Scriven (1966, 1983a, 1983b) suggests that such evaluators believed their activities could and should be value-free. But it proved to be impossible in the political world of social programming to evaluate without

Table 2.3 A Summary of the Valuing Component

The issues: Values are omnipresent in social programming. How can evaluators make value problems explicit, deal with them openly, and produce a sensitive analysis of the value implications of programs?

Knowledge bases:

(1) metatheory, the study of the nature of and justification for valuing;
(2) prescriptive theory, theories that advocate the primacy of particular values; and
(3) descriptive theory, theories that describe values without claiming one value is best.

A better theory of this component

(1) describes all three of these elements;
(2) recognizes that
  (a) no prescriptive theory is widely accepted as best,
  (b) all prescriptive ethics are underjustified, and selecting one involves trade-offs about which few stakeholders agree, and
  (c) descriptive theories are more consistent than prescriptive theories with the social and political organization of the United States, which is based upon fostering a pluralism of values that compete against each other in social and political arenas;
(3) clearly states its priorities about which kinds of values to attend to, and why.

values becoming salient. Social programs are themselves not value-free. For example, Fairweather's (1980) Lodge program was not widely implemented partly because its implicit values were inconsistent with profit in the free marketplace and professional control over mental health (Shadish, 1984). Second, evaluative data about program effects relative to needs, about expenditures and implementation relative to congressional intent, or about fraud and abuse enter policy debates to influence decisions about programs. Those decisions, involving distribution of social resources, are matters of values and ethics. Third, data do not speak for themselves, but are interpreted in terms that invoke values. For instance, Cook et al. (1975) found that *Sesame Street* teaches some alphabet skills to economically disadvantaged children who view the show regularly. But the disadvantaged watch the show less often than affluent cohorts, who learned even more skills. So *Sesame Street* increased the gap in skills between these two groups. Are the gains of disadvantaged children worth widening the gap between them and advantaged children? There is no easy answer to such value dilemmas. Fourth, an implicit notion that evaluation serves the "public good" runs through most evaluation, yet the concept is rarely explicated. Little agreement exists among evaluators or others about what the concept means. Since it is implicitly central to evaluation, it deserves more attention. Theories of valuing help evaluators to make such value problems explicit, deal with them openly, and produce a sensitive analysis of the pros and cons of results.

Theories of valuing can have three elements (Beauchamp, 1982):

- *metatheory:* the study of the nature of and justification for valuing
- *prescriptive theory:* theory that advocates the primacy of particular values
- *descriptive theory:* theory that describes values without advocating one as best

Metatheory describes how and why value statements are constructed, for example, analyzing the meaning of key terms, the structure or logic of valuing, and the nature of justifications for values. In this book, Scriven has the only metatheory. He argues that evaluation ought to be about constructing value statements; he analyzes the meaning of key terms having to do with values; and he describes a logic for constructing value statements about anything. The logic involves (a) selecting criteria of merit that something must do to be good, (b) setting standards of performance about how well it must do on the criteria, (c) measuring performance on each criterion and comparing it to standards, and (d) synthesizing results into a value statement. The logic allows evaluating such diverse entities as curricula, word processors, teachers, and social programs. It is a technology, a tool that is neutral with respect to any particular ethic or morality. Scriven claims it unites all evaluative work, and we tend to agree. But even if he is wrong, without a metatheory for constructing value statements, it is unclear how evaluators can state or justify a conclusion that a program is good or better than something else.

Some theorists promote particular values—a prescriptive theory of valuing. House (1980) does this when he concludes that evaluators should follow Rawls's theory of justice, prioritizing in the interests of the economically and politically disadvantaged. Bunda (1985) takes the prescriptive approach a step further and examines what variables would be studied in three kinds of educational programs depending on which system of ethics one used. Depending on the theory, evaluators would have different conceptions of programs, and would use different dependent variables.

The justification and internal explication of prescriptive theories usually have been carefully worked out over centuries of philosophical thinking. For example, Rawls premises the priority he gives to the disadvantaged on an initial assumption that the ideal decision maker is blind to his or her own economic, social, and political conditions (for example, blind to whether or not his or her parent will leave a large inheritance). Evaluators can assess those assumptions prior to using them in evaluation. The relationships of prescriptive theories to each other and to issues in philosophy have also been explored, so the relative strengths and weaknesses of each approach are known. For example, those who know

Rawls's theory of justice will detect that Scriven's definition of "need"—central to his theory—implies endorsing assumptions of egalitarian theories of ethics that are not obvious. Prescriptive theories give evaluators a critical perspective and intellectual authority that descriptive theories cannot match. They broaden evaluators' understanding of good social programs by broadening their understanding of what is good for the human condition generally.

Most evaluators use descriptive valuing: They describe values held by stakeholders, determine criteria they use in judging program worth, find out if stakeholders think the program is good, and see what they think should be done to improve it. The claim is not that these values are best, but that they are perceptions of program worth that are grist for the mill of decision making. Descriptive valuing is implicit in most evaluation theory, even though the word *values* may never be used. Stakeholder-based approaches are descriptive approaches because they solicit information about stakeholder interests in the program and its evaluation (Bryk, 1983). Similarly, Wholey uses a version of this approach even though he never mentions values. Other theorists who discuss values explicitly, such as Stake, say evaluators should study values descriptively because we do not have a correct prescriptive theory, and because the evaluator should not impose one ethical view on a program in a political system characterized by value pluralism. Descriptive values are easily constructed by contacting stakeholders; no special training in ethics is needed. All this makes descriptive valuing more practical than prescriptive valuing.

Good value theory comprehensively discusses its stance on these things. Only Scriven gives extensive attention to values, and then only to metatheories. No evaluation theorist does descriptive, prescriptive, and metatheorizing explicitly and systematically. Yet evaluators acknowledge that values deserve more attention (House, 1980). Nearly all the theorists in this book agree that evaluation is about determining value, merit, or worth, not just about describing programs. But few theorists do more. They rarely even deal with related topics, such as cost-benefit and cost-effectiveness analyses, to develop statements about utility based on scaling outcomes in monetary terms. These analyses deal explicitly with valuing, and most theorists ignore them.

Good value theory has descriptively accurate content. Since most evaluation theories have little content about values, they fail this test straight away. But some theorists struggle with valuing. For them, the following minimum criteria ought to be part of a good theory about valuing in evaluation. To begin with, just as in epistemology, in prescriptive value theory there are no clear winners. In ethics, utilitarian theories (roughly, theories aiming to produce the greatest good for the greatest number)

compete with deontological theories (doing one's duty), and both compete with virtue theories (emphasizing characterological propensities to act properly). Egalitarian theories of justice (emphasizing equality) compete with libertarian theories (emphasizing liberty), and the list goes on (Beauchamp, 1982). Philosophers simply disagree on which theory is best. A theory about prescriptive values that fails to mention this fact is deficient. In addition, the choice of prescriptive ethic involves trade-offs. For example, Rawls's theory of justice focuses on the material needs of the disadvantaged, which libertarians will object might require them to sacrifice resources their theory says they can keep (for example, giving up more or all of an inheritance to taxes). Any theory advocating a particular ethic should outline such gains and the losses, and consider if it is worth alienating stakeholders who may object to the losses. Finally, the social and political organization of the United States reflects a conscious effort to foster a pluralism of values that compete against each other in social and political activity. When evaluators describe this plurality of values, and provide results that bear on those values, they increase the chances that the information will be perceived as fairly reflecting the interests being debated. Conversely, advocating a prescriptive ethic, and gathering data on that basis, will not reflect this plurality well, and the likelihood that the information will be perceived as fair will be decreased, thus making it less credible in policy. Descriptive theories are more practical than prescriptive theories in this sense.

Good value theory clearly states its priorities concerning which kinds of values to study, and why. One choice is whether descriptive or prescriptive theories have priority. Since descriptive theories are more politically and socially practical in a system of pluralistic interests, we assume that descriptive values ought to have priority most of the time in evaluation. Any theory that prioritizes on prescriptive theories has the heavy burden of justifying why. In either case, the theory must justify selections within descriptive or prescriptive theories. Within prescriptive theories, Rawls's theory of justice is only one of many theories; one alternative, for example, is Nozick's (1974) procedural theory of justice. Justice is just one important moral issue raised in social policy; others are human rights, liberty, freedom, equality, and utility. Value theories must tell the general orientation to be taken (descriptive or prescriptive), the particular values preferred in each orientation, and justification of those choices.

There are many different ways to do descriptive valuing. Stake tries to give all values a hearing by consulting with all stakeholders. This uses resources that could be spent elsewhere in evaluation, so other evaluators choose only a few stakeholders from which to construct descriptive values. Cohen (1983) says that each stakeholder group should have its own

evaluator to represent it, but this is not very practical. Wholey gives priority to program managers. Many evaluators give priority to elected representatives, because no other group is so politically validated to represent wide interests.

Does good theory of valuing make a difference? The differences between prescriptive and descriptive valuing are driven home by House's (1980) claim that prescriptive ethics are integral to evaluation. He begins his argument by saying:

> Evaluation is by its nature a political activity. It serves decisionmakers, results in reallocations of resources, and legitimizes who gets what. It is intimately implicated in the distribution of basic goods in society. It is more than a statement of ideas; it is a social mechanism for distribution, one which aspires to institutional status. Evaluation should not only be true; it should also be just. Current evaluation schemes, independently of their truth value, reflect justice in quite varying degrees. And justice provides an important standard by which evaluation should be judged. (p. 121)

He examines different prescriptive ethics for their suitability to evaluation, and settles on Rawls's theory of justice. From Rawls, he draws implications for how evaluations ought to be done, especially regarding the needs of the disadvantaged.

House's position has elicited strong reactions from evaluators who were probably advocating descriptive approaches. Kenny (1982) complains:

> I am very suspicious of those who say they are speaking for the poor or disadvantaged when they themselves are not poor or disadvantaged. It strikes me that the highest form of elitism occurs when persons unchosen by the disadvantaged say that they speak for the disadvantaged or they say that they take the disadvantaged's interests into account. Let us be concerned, but let us remember that we can speak only for ourselves. (pp. 121-122)

Having the disadvantaged speak for themselves is descriptive valuing.

Wortman (1982) responds to House:

> Most modern evaluators are concerned with the fairness, justice, and ethical conduct of the government. However, most of this work is done by contract from the government and for the government. There is little opportunity to negotiate agreements, and all audiences are assumed to be represented by this legitimate authority. For the evaluator to challenge this except in clear legal and ethical violations of conduct would be foolhardy. (p. 124)

Wortman endorses the political legitimacy of government authority given the political process – also descriptive valuing.

Consider the practical payoff of descriptive valuing in more detail. If the values of a particular stakeholder group are not considered, that group may feel morally and politically slighted and may be uncooperative with the work and critical in subsequent debates. If a group's values are misunderstood, group members may see the evaluation as less relevant than otherwise. House, Glass, McLean, and Walker (1978) claim that this happened with the Follow Through evaluation. Program developers said evaluation measures did not tap the constructs they thought the program would change. Parents said they had been excluded from decisions about the evaluation, so it did not reflect their interests. Consequently, the evaluation had less credibility with these stakeholders.

In all of these forms, value theory is necessary to evaluation theory. It helps evaluators understand what steps to undertake to make value statements about programs, to see value judgments implicit in their work, to place recommendations about ethics and values in a common perspective within which to contrast and compare them, and to make decisions about implementing those recommendations in their work. Without this component, evaluators may not understand, or even detect, the values that permeate their work.

THE USE COMPONENT OF EVALUATION THEORY

Evaluators hope that their work is useful in social problem solving. By contrast, basic researchers often disclaim any intention to be useful to decision makers, policymakers, program managers, or other stakeholders. Society justifies large funding for evaluation partly expecting that some more or less immediate payoffs will accrue. If evaluation is not useful, those funds could be used for more programs, for reducing the deficit, or for other alternatives that might yield more immediate results. Thus evaluators need a theory to tell how, when, where, and why they can produce useful results.

Theories of use have three elements:

- a description of possible kinds of use
- a depiction of time frames in which use occurs
- an explanation of what the evaluator can do to facilitate use

Consider different kinds of use. Early evaluations aimed to show which programs worked. The implicit assumption about use was that policymakers would eliminate ineffective interventions and replace them with better

Table 2.4 A Summary of the Use Component

The issues: How can evaluators produce results that are useful for social problem solving?
Knowledge bases:
(1) a description of possible kinds of use;
(2) a depiction of the time frames in which use occurs;
(3) an explanation of what the evaluator can do to facilitate use under different circumstances.
A better theory of this component
(1) addresses all three of these elements;
(2) recognizes that
  (a) use of evaluative results can threaten entrenched interests,
  (b) certain types of information are harder to use than others (for example, it is hard to use results to start or end social programs because those programs as a whole rarely start or end),
  (c) the slow, incremental nature of policy change implies that instrumental use is also slow and incremental,
  (d) policymakers often give ideology, interests, and feasibility a higher priority than evaluation results,
  (e) using evaluation results is not a high priority for many practitioners who assume the efficacy of what they do, and
  (f) different activities facilitate different kinds of use, but limited time and resources make it hard to do them all;
(3) identifies key choices that evaluators must make in deciding how, when, where, and why to try to produce useful results.

ones. This is instrumental use: making direct decisions about changing programs based on evaluation results. But over time evaluators found that other kinds of use also occurred. Sometimes results were not used instrumentally to make changes, but they still affected how people thought about an issue. Theorists call this "conceptual use" (Leviton & Hughes, 1981), "enlightenment" (Weiss, 1977b), or "demystification" (Berk & Rossi, 1977). All believe that this form of use is legitimate. Sometimes evaluation results were used to persuade people of a position already taken. At first evaluators dismissed such use as less scientifically legitimate – it seemed more like lobbying than scientific reasoning. But presenting evaluative data in policy debates is always partly an exercise in persuasion. Evaluators themselves engage in this kind of use when they argue that their approaches to particular studies are compelling, or that their interpretations are authoritative. Eventually all three of these uses (instrumental, conceptual, persuasive) have found a place in evaluation.

Second, evaluators began to see that different kinds of use occur in different time frames. The early hope was that use would occur quickly, leading to immediate program changes and improvements. But evaluators found that sometimes changes in programs did not show up for years, until policy circumstances allowed adoption of a new intervention (Polsby,

1984). They found that people who read evaluation results were conceptually influenced, but did not have much latitude to make changes that later proved feasible (Weiss & Bucuvalas, 1980). Hence evaluators granted legitimacy to long-term use, even though it moved away from the hopes for short-term use that initially drove the field.

Finally, because use is central to justifying the field, evaluators quickly realized that they needed plans to facilitate use. The early hope was that use would happen with little effort because evaluation results were compelling, because stakeholders eagerly awaited scientific data about programs, or because policy-making was a rational problem-solving endeavor. If those hopes were realistic, the evaluator's job would be simple: Produce results and wait for users to arrive at the doorstep. But those hopes were dashed, because evaluation results were seldom compelling relative to the interests and ideologies of stakeholders, stakeholders usually regarded scientific input as minor in decision making, and problem solving is far from a rational endeavor. If evaluations were to be used in such an environment, evaluators had to take active steps toward that end. Largely through trial and error, evaluators developed ways to do so.

A good theory of use comprehensively discusses possible kinds of use, time frames in which use occurs, and things evaluators do to facilitate use. Debates about use occurred early in evaluation's history, so comprehensive discussions of use have appeared (e.g., Leviton & Hughes, 1981) that have been adopted by some evaluation theorists (e.g., Rossi & Freeman, 1985). But most theorists discuss their favorite kinds of use in detail while minimizing alternatives. Some theorists have no explicit theory of use at all, and have flawed implicit theory.

Good theory of use has descriptively accurate content. Evaluators have learned much about obstacles to short-term instrumental use (Weiss, 1972a, 1972b)—with implications for accomplishing the purpose of the field and for the credibility of the claim that evaluation provides society with short-term return on its investment. One obstacle is that evaluative results often threaten entrenched interests. Evaluation always has an adversarial relationship with some parties to public debates; its findings do not enter debates as uncontested nuggets of truth. A second obstacle is that social programs *as a whole* (as opposed to projects or elements) rarely die or get replaced. When they do, it is mostly for political or economic reasons, not because they were evaluated negatively. Hence evaluative information about *program* effects does not affect policy quickly. A third obstacle is that service deliverers engage in practices for reasons besides efficacy, such as convenience, habit, and security. A fourth obstacle is that political decision making is a slow process, and change is almost always incremental. A fifth obstacle is that policymakers and managers use information in many ways. Although political decision makers appreciate information about program performance (Weiss & Weiss, 1981), they must attend to conflicting values, interests, and expediency; and they have less power to modify programs than outside observers might think. Constraints are imposed upon them by past decisions and current fiscal realities, and by the political realities of social programming. A sixth obstacle is that even if an innovation is highly successful, local personnel may still not adopt it. Potential adopters of an innovation do not learn about it through journals, books, or in-service training, since not all professions require in-service training and few practitioners cite scholarly journals as influencing their practice (Barlow, 1981; Barrom, Shadish, & Montgomery, 1988). Textbooks for practitioners in social service fields contain information about major evaluations, but these take time to be known (Leviton & Cook, 1983).

Evaluators have learned much about what to do to facilitate use. Helpful activities for instrumental use include identifying users early in the evaluation; having frequent contact with users, especially during question formation; studying things that users can control; providing interim results; translating findings into actions; and disseminating results through informal meetings, oral briefings, media presentations, and final reports with brief and nontechnical executive summaries. Each of these activities also aids conceptual use even when the user cannot act on results. Conceptual use can also be facilitated by challenging fundamental assumptions about problems and policies, and by circulating results through the network of scholars, policymakers, and interest groups concerned with the issue.

Good theory of use identifies key choices that evaluators must make about how, when, where, and why to produce useful results. Although some conceptual use accompanies instrumental use, evaluations that produce the latter will sacrifice some of the former. For example, studying interventions that policymakers control increases instrumental use but restricts the range of options to those that most resemble what already exists. These may be the least likely to challenge fundamental assumptions about social problems and solutions. A good theory makes such trade-offs explicit, and justifies the choices made.

Does good theory of use make a difference? In his seminal article "Reforms as Experiments," Campbell (1969) proposes the short-term instrumental use of evaluation results, advocating 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" (p. 409). At the time, this seemed reasonable to many evaluators. Evaluation would provide the data about effects, and policymakers would use those data to decide.

Weiss (1973b) counters that, contrary to Campbell's hopes, "devastating evidence of program failure has left some policies and programs unscathed, and positive evidence has not shielded others from dissolution" (p. 40). Even if evaluation shows a program is ineffective,

> a considerable amount of ineffectiveness may be tolerated if a program fits well with prevailing values, if it satisfies voters, or if it pays off political debts. What evaluation research can do is clarify what the political trade-offs involve. It should show how much is being given up to satisfy political demands and what kinds of program effects decision-makers are settling for or foregoing when they adopt a position. (p. 40)

All this led Weiss to emphasize conceptual use and activities that facilitate it.

Neither Weiss nor Campbell is as comprehensive as Rossi, who finds a place for all kinds of use in his theory. Rossi contends that sometimes Weiss is right to stress enlightenment, but that evaluation need not give up instrumental use entirely just because some of Campbell's implicit assumptions were not realistic. Rossi points to Wholey as a model for instrumental use, and he outlines conditions under which evaluators should pursue Wholey's approach. Rossi's theoretical comprehensiveness, coupled with solid grounding in empirical findings about use, gives evaluators better advice than a narrower theory could.

Evaluation has some examples of use gone awry that we could better predict with good theory of use. Salasin (1980) describes locally controlled community mental health center (CMHC) evaluations that were intended partly to increase local short-term instrumental use by tying evaluation to local CMHC managers. The premise was that managers would request information that they were particularly likely to use and to trust because they supervised its gathering. This approach did produce short-term instrumental use, but the information needs of managers revolved around billing, reporting, and public relations, so this approach underemphasized the effect of the local program in addressing social problems (Cook & Shadish, 1982). Too much emphasis on local instrumental use, undisciplined by knowledge of political and organizational constraints that shape the questions asked, risked trivializing evaluation by destroying the distinction between evaluating a program and constructing a management information system within it.

Such experiences led evaluators to increased humility about the use of their work, and about the small role social science information plays in social problem solving (Lindblom & Cohen, 1979). It also prompted them to explore the problem in more detail, leading many theorists to set the goal of facilitating use as the major agenda in their work. All these theorists are skeptical that major instrumental use occurs often for changing programs, but they are exploring circumstances that foster it without sacrificing a connection to social problem solving (Leviton & Boruch, 1983). They study how social science information is used by policymakers (Weiss & Bucuvalas, 1977; Weiss & Weiss, 1981) to learn how to take advantage of those uses. They explore how use is promoted by a focus on questions about program implementation to help managers target the program effectively, and about the effects of robust practices that can be adopted during practitioner training and development.

Hence a theory of use is a necessary component of evaluation theory. It helps evaluators assess the role of new information in decision making and in shaping practice, to know what information is more or less effective for such purposes, to understand the role scientific information plays in an interest group democracy where certain freedoms are inconsistent with naive conceptions of how science should influence policy, and to decide what kinds of use they want their work to facilitate. Without such a component, evaluators risk producing information that cannot be used.

THE PRACTICE COMPONENT OF EVALUATION THEORY

Practitioners are action oriented. If they rely on theories of evaluation at all in their work, it is to find pragmatic concepts to orient them to their task and to suggest general strategies and some practical methods to implement those strategies. They want useful advice about how to make their decisions given the constraints under which they work. Theories of practice—the most essential component in theories of evaluation—address this need.

Theories of practice discuss the essential decisions in any evaluation (Shadish, 1986b; Shadish & Epstein, 1987). They have elements that address the following:

- whether or not an evaluation should be done at all
- what the purpose of the evaluation should be
- what role the evaluator ought to play

Table 2.5 A Summary of the Practice Component

The issues: Practicing evaluators need pragmatic concepts to orient them to their task and to suggest general strategies; they need practical methods to implement those strategies within the constraints under which they work.

Knowledge bases: This component addresses

(1) when an evaluation should be done,
(2) what the purpose of the evaluation should be,
(3) what roles the evaluator ought to play,
(4) what types of questions should be asked,
(5) what design will be used, and
(6) what activities will be carried out to facilitate use.

A better theory of this component

(1) addresses all these elements;
(2) prescribes various options that can be implemented, given various constraints under which evaluators work;
(3) clarifies the logical and logistical contingencies among these decisions; and
(4) suggests priorities in general and in different situations.

- what questions will be asked
- what design will be used
- what activities will be carried out to facilitate use

Theories of practice depend on the other four components of evaluation theory, because the content of the above list is drawn partly from the other components. Theory of knowledge covers methodology; choice of method depends on assumptions about the strengths and weaknesses of methods, and about the kinds of knowledge most worth constructing. Theory of practice is similarly dependent on theory of valuing. Choice of a particular descriptive or prescriptive approach to valuing influences the questions asked and the variables measured. With a descriptive approach, an evaluator will let stakeholders heavily influence question formation and decisions about dependent variables; this might not be done with a prescriptive approach. Theory of use has implications for how to facilitate use in practice. Similar connections hold between theory of social programming and theory of practice. Evaluators who believe that short-term incremental change can lead to worthwhile improvement in a program's capacity to address social problems may undertake evaluations with that purpose.

Yet theory of practice is more than just the sum of content from the other components. Its primary purpose is to prioritize that content given the limited time, resources, and skills that constrain any given evaluation—*which* methods to choose from those possible, or *which* questions to ask. Theory of practice deals with the following decisions. First, an evaluator must decide whether to do an evaluation at all. Some evaluators have little

choice, being told by an employer to do so. Other evaluators have more latitude in making the decision. Some academic evaluators decide to evaluate (or not) out of intellectual curiosity, out of interest in a basic research question addressable by evaluation, or out of a desire to publish in the area. Other considerations also enter the decision. Scriven tells evaluators to consider if the benefits of the evaluation exceed its costs. Wholey asks if management will make changes if findings are provided. Nearly everyone asks if the evaluation can be done in a worthwhile form given available resources. Those who fund evaluation consider whether the money would be better spent on other things, such as more social programming.

Second, if an evaluation is to be done, it could be done for different purposes. Scriven identifies two alternatives—formative and summative evaluation. Formative evaluations improve program performance by influencing immediate decisions about the program, especially about how its component parts and processes could be improved. Summative evaluations judge program worth by assessing program effects in light of relevant problems. This dichotomy oversimplifies, as any simple dichotomy is bound to do. It emphasizes short-term rather than longer-term or conceptual uses, for example.

Third, evaluators can take on different roles. Most early evaluators were social scientists who were used to pursuing questions out of personal curiosity rather than to satisfy needs of clients or interests of stakeholder groups, to having freedom of inquiry in topics and methods, to being responsible to scientific peers rather than to evaluation funders, and to letting the marketplace of ideas use results rather than taking an active part in promoting use. These role assumptions were not always functional in the world of evaluation. Even those evaluators who remained in academia modified the traditional role somewhat, seeing themselves as methodological experts giving competent consultation to clients, as educators of evaluation clients about social programs, as servants of the "public interest" who owe no allegiance to the interests of stakeholders, or even as judges of program worth who use the security of academia to safeguard against being co-opted by stakeholders. Evaluators in policy or program settings, or in the private sector, began to see themselves as servants of program managers or program stakeholders, as facilitators of program change, or as part of the program team. In the latter roles, responsibility to a specific program, its staff, and stakeholders often takes precedence over traditional role behaviors of scientists.

Fourth, evaluators can ask questions about (a) intended or actual clients, their needs and desires; (b) program inputs, such as budget and staff; (c) internal structure and processes of the program, including activities,

internal organization, and program model outlining the relationship among these things; (d) variables outside the program that affect it, such as legislation, regulations, or local political and organizational support; (e) changes in clients or society; and (f) costs and benefits of the program. The specifics of each of these questions can be drawn from such sources as stakeholder information needs, past research and theory, and pending decisions or legislation.

Fifth, the evaluator must design the study. This includes selecting methods—inspection of program records, on-site observation, sample surveys, interviews with stakeholders, program monitoring, client needs assessments, experimentation and quasi-experimentation, metaevaluation, secondary analysis, causal modeling, and meta-analysis, to name a few. Design also includes identifying variables to be measured—inputs, program implementation and activities, or outcome. Breaking this down even further, measures of outcome can be drawn from program goals, anticipated side effects, relevant legislation, social science theory, ethical theory, or management expectations.

Finally, the evaluator must decide how to facilitate use. Minimally, the evaluator places results in public forums such as books, journals, or popular media. Extensive user contact also facilitates use by allowing the evaluator to identify users early, to ask about how information could be used, to keep in contact during the evaluation, and to provide interim results. Use can be aided by final reports that include executive summaries and action recommendations, and by using oral briefings and other supplements to final reports.

Few theories of evaluation practice are comprehensive. Since few of them are equally explicit about all of the other four components of theory, their prescriptions for practice often borrow heavily from one component but are poorly grounded in another. For example, Scriven takes a metatheory of valuing and combines it with a theory of knowledge emphasizing critical realism and questions about needs and program effects, but leaves out most theories of use and social programming. Stake uses descriptive valuing, emphasizes local program improvement from theory of social programming, stresses conceptual use, and says little about theory of knowledge. No theorist builds explicitly on strong analysis of all four of components.

Comprehensiveness has another implication for theories of practice: They must include advice about *what* to do and *how* to do it. When Campbell (1969, 1971; Campbell & Stanley, 1963) discusses methods for causal inference, he tells the evaluator how to use experiments or selected quasi-experiments to do this. When Wholey (1979, 1983) tells the evaluator to provide quick, sequentially better evaluation results to managers,

he gives them the details of evaluability assessment, rapid feedback evaluation, and performance monitoring to implement that advice. Other theorists err by not elaborating how to implement their prescriptions. The mistake is not serious if the method is common and easily accessible; but it is often fatal for novel prescriptions. Scriven's (1976b) goal-free evaluation has been widely criticized for lack of operations by which to conduct it; it may be less widely used as a result. He tells you what to do but not how to do it. Creative practitioners may invent means of doing it, but it is doubtful this occurs often.

Good theory of practice has descriptively accurate content. This is largely ensured if the theory has accurate content in the four component theories—use, valuing, knowledge, and social programming. In theories of practice, the term *accurate* also takes on the connotation of *realistic.* Some theories of practice fail to give operations that can be implemented given the constraints under which evaluators work. Many evaluators, for example, do not have the skills or background to implement linear structural modeling programs. A similar criticism of randomized experiments focuses on the practicality of denying access to eligible clients of a program. Practicing evaluators need to know the resources required to implement different methods, especially relevant methods, and the quality of information yielded by each.

A descriptively accurate theory of practice also explicates connections among the decisions that evaluators face. A decision to evaluate for one particular client—say, program management—influences the kinds of questions asked. Management is more responsible for program improvement than for policy decisions about programs, and so is more interested in questions about program improvement than in whether the program as a whole is effective. Similarly, choice of question might lead to selecting some methods over others. If the question has to do with problem prevalence in a population, most methodologists might recommend a sample survey. Even the choice of role constrains practice; evaluators who see themselves as methodological experts may be less likely to worry about whether or not results get used.

Good theory of practice—more than anything else—is about *setting priorities* and the trade-offs that go with doing so. The very point of the theory is that practice cannot be comprehensive because of limited time, resources, staff, interests, and skills. Practice is about making constrained choices with a realistic understanding of losses and gains. Theory of practice must clearly identify those constraints and help practitioners sort through options to find the feasible few. This is usually done through heuristic devices to guide choices. These heuristic devices have qualities we associate with the notion of a schema from cognitive science. Schemas

are organized, relatively stable frameworks of concepts, connections among concepts, slots for categorizing input, and prescriptions for actions. Once learned, schemas organize and make sense of experience. When a situation does not fit well into the schema, it is adjusted or a new schema is learned. As with schemas generally, evaluators have a limited number of heuristic devices to use in practice. They continue to apply those concepts routinely until the fit between them and the situation is so poor that a change must be made.

Heuristic devices in theories of practice are of two kinds: focusing devices and contingency devices. Focusing devices reduce the scope of possible activities by focusing attention on some things rather than on others. They tell evaluators to use some subset of roles, questions, methods, or practices. An example is Scriven's goal-free evaluation; he tells the evaluator to ignore program goals, to avoid contact with management, and to identify program effects relative to social needs. Stake's responsive evaluation tells evaluators to be responsive to the program by observing it passively and in its natural state, and to avoid introducing treatments, raters, or questionnaires into it. Campbell's experimenting society focuses attention on cause-and-effect questions, experimental methods, and a direct problem-solving view of social change. The advice does not vary over evaluators, programs, or situations; rather, the evaluator's attention is focused on a limited, tractable set of tasks.

Focusing devices always succeed partially because their recommendations are always realistic for some evaluators in some settings. But they are inherently limited because they tell evaluators about only part of the total picture. If a more complete theory of practice is better than a less complete one, then focusing devices are not a good answer to the problem of setting priorities. A more desirable solution is the use of contingency devices. Focusing devices were common during the 1960s and 1970s, but recent theory makes more use of contingency devices. Contingency devices specify practices that vary depending on situations. Rossi's tailored evaluation attends to level of program development, telling evaluators how to vary their practices depending on if the program is a new innovation, an ongoing program, or a modification to an ongoing program. Cronbach agrees, but presses evaluators toward giving more attention to those uncertainties with greatest significance for policy, if they can be investigated cost-effectively. In these cases, the questions asked, methods used, and anticipated uses vary depending on the situation. Contingency devices allow evaluators to exercise a wider array of skills in a wider array of situations than focusing devices allow. But contingency devices have problems, too. They are more complex than focusing devices, and require evaluators to possess more skills. Further, no single heuristic device

captures all the complexities of practice, so it helps to use more than one device in a theory of practice—increasing complexity more. Finally, in simplifying complex input, heuristic devices risk oversimplifying and so painting an inaccurate picture of practice.

Cronbach, faced with this array of possibilities and constraints, claims that evaluation practice is more art than science. The possible choices are so vast as to require seasoned judgment and virtuosity in combining them, and each evaluation occurs within so many unique circumstances that the combinations are rarely the same over time or place. Hence even the best theories of practice do not try to specify rules for good practice completely. Most use a limited number of heuristics to organize the options into a tractable few.

Does good theory of practice make a difference? Yes; it helps evaluators sort through the discrepant advice that different theorists give about practice. Wholey (1979), for example, tells evaluators to use the following four steps of his sequential purchase of information:

> Evaluability assessment clarifies the extent to which evaluation information is likely to be useful and suggests changes in program activities and objectives which could improve program performance. Rapid feedback evaluation provides preliminary evaluations as by-products of the data collection and analysis needed in designing full-scale evaluations that will be worth their cost. Performance monitoring provides the program performance information that program managers need to set performance targets, to test alternative ways to meet or exceed performance targets, and to document the feasibility and cost of improving program performance. Intensive evaluation provides more conclusive evidence on the effectiveness of program activities when such information is needed for specific policy or management decisions. (pp. 202-203)

But Stake (1978) prescribes a quite different approach that relies heavily on case study methods:

> Most case studies feature: descriptions that are complex, holistic, and involving a myriad of not highly isolated variables; data that are likely to be gathered at least partly by personalistic observation; and a writing style that is informal, perhaps narrative, possibly with verbatim quotation, illustration, and even allusion and metaphor. Comparisons are implicit rather than explicit. Themes and hypotheses may be important, but they remain subordinate to the understanding of the case. (p. 7)

For logical or logistical reasons, both sets of advice cannot be followed concurrently. One example is logically contradictory advice. Wholey (1983) suggests working with managers to specify goals and objectives;

Scriven (1976b) says to avoid both managers and goals. Logically, the evaluator cannot do both. Another example concerns discrepancies that are practically inconsistent. Campbell (1969) says to use experimental methods or some high-quality quasi-experiments; Stake says to do nonexperimental case studies. Limited resources generally preclude the evaluator from doing both in any given study.

Surveys of evaluators suggest that their practices are influenced by theory. Although they learn from on-the-job experience, discussions with colleagues and clients, observations of other evaluations, graduate school training, reading, and workshops, they rate new ideas in the field as the most influential factor in the changes they have made in their work (Shadish & Epstein, 1987). The theoretical orientations these evaluators endorse are also statistically related to the use of practices consistent with those theories.

Of all the components, evaluation theory is better if it contains a complete, accurate, and realistic theory of practice. Such a theory lists the tasks that evaluators must do, the options for doing them, the resources required for each, the trade-offs among them, and the justifications for choosing one over another in particular situations. Practicing evaluators need such a discussion to avoid doing evaluations that are incomplete, impractical, or technically inferior.

THE REST OF THIS BOOK

Our vehicle for discussing evaluation theory is the work of seven evaluation theorists from the last 25 years. We devote one chapter each to Michael Scriven, Donald Campbell, Carol Weiss, Joseph Wholey, Robert Stake, Lee Cronbach, and Peter Rossi, in that order for reasons discussed shortly. The first part of each chapter uses extensive quotation from each theorist's original work to present his or her theory as faithfully as we can to the spirit of the work. We identify the concepts each explicitly uses in dealing with each component, or suggest those that may be implicit in his or her theory, although we try to label it as such when we do so. The second half of each chapter examines how each theorist deals with the five theoretical components.

Our Selection of Theorists

It is worth commenting on how we chose these seven theorists, and why, because the reader might have made other choices. We selected theorists who had written broadly about issues in evaluation, excluding those who

mostly write about, say, methodology, as do many econometricians who do evaluations (Stromsdorfer & Farkas, 1980). We tended to include theorists who had been in evaluation for a while, mostly because they have experienced the diversity of growth in approaches in evaluation as it tried to cope with the harsh lessons of field research over the last decades. We selected authors in part to illustrate our view of the historical development of evaluation theory. We also chose theorists to reflect diverse positions in evaluation, omitting some theorists who elaborate positions initially outlined well by earlier theorists. All these decisions were partly arbitrary, so we omit from this book a number of good theorists whom others might have included.

This latter point deserves more comment. Program evaluation is a changing, dynamic field. New theorists constantly emerge whose work could be included in a book like this, such as Chelimsky (1987a, 1987b, 1987c), Hausman and Wise (1985), Haveman (1987), Lincoln and Guba (1985), Nathan (1989), and Patton (1988). If a second volume like this were written, they would be prime candidates for inclusion. But omitting such theorists does less *conceptual* harm than one might think. The issues raised by our theory—about use, valuing, social programming, knowledge, and practice—are broader than the work of any one theorist. The reader who learns to think in these terms, who can criticize any theory (or any evaluation!) from this perspective, is better equipped to understand the merits of new and emerging theories than is a reader who is simply coached about who is currently saying what. This is not just a book that describes evaluation theory; it is about *how to think about* evaluation theory. Moreover, new theorists do not always provide new answers to fundamental problems in the field. Their contribution is sometimes to combine old answers in new ways, to popularize an established point, or to show how an old idea applies in a new context. Such theorists make novel contributions, of course—indeed, we regret we cannot include all the novel contributions of the many bright theorists in evaluation. But still, newer theorists inevitably have a conceptual debt to the ideas of past theorists. There may be diminishing overall yield from new theories as time passes and an area matures.

It is probably also worth mentioning criteria we did not use in selecting theorists. We did not select these seven theorists as being the *most* influential in the field. Some of these theorists have been more influential and others less so, both relative to each other (Shadish & Epstein, 1987) and relative to theorists we did not include (Smith, 1981). Finally, these seven people might not necessarily see themselves primarily as either theorists or evaluators. One reviewer of a draft of this book said he did not think of some of them as being evaluators as much as "public administra-

tors, interventionists, problem-solvers, solution-seekers," nor so much theorists as people with "a philosophy, orientation, bag of lessons learned over the years." Yet whether or not they intended to be evaluation theorists is beside the point. Along with a few other people whose work we could have included, these seven are viewed as evaluation theorists in the field. We cannot imagine writing a book on evaluation theory that did not include them or colleagues like them.

Order of Theory Presentation: Three Stages of Evaluation Theory

The order in which we present these theorists illustrates our perception of the evolution of evaluation theory in three stages. Scriven and Campbell represent the first stage. As two of the earliest theorists in the 1960s, they provided concepts and methods particularly for valuing and knowledge construction. They advocated a rigorous scientific search for effective social interventions to solve social problems. But discontent with first-stage approaches led to a search for alternatives, represented by Weiss, Wholey, and Stake. These theorists illustrate an explosive growth of alternatives in the 1970s, and a special concern with being more realistic about the nature of social programs and about how social science concepts and findings are used in policy. Finally, a third stage of theory was devoted to integrating these alternatives into a coherent approach to evaluation, represented by Cronbach and Rossi. We describe these stages in more detail in material preceding the chapters themselves.

Stage theories oversimplify complex arguments. They imply continuous progress that belies the starts and sputters that actually characterize the development of evaluation theory. Also, the stages are not mutually exclusive; later theorists absorbed parts of previous theorists' work, rejected other parts, invented new parts, and made their own errors. Some theorists resist classification into one stage. For example, Wholey and Weiss briefly endorsed Campbell's approach, but most of their subsequent efforts to generate useful, politically realistic theories have dramatically modified or departed from Campbell. Similarly, when first-stage theories were dominant, Cronbach (1963) foreshadowed his contributions 20 years later, though the latter contributions were more comprehensive, better informed, and might not have been possible during the first stage. In some respects Cronbach and Weiss are closer in spirit than Cronbach and Rossi. We categorize Cronbach and Weiss in different stages to acknowledge Cronbach's explicit appeal to the defining features of third-stage theorists – use of contingency devices, and specific efforts to incorporate the work of preceding theorists. Finally, it would be wrong to infer

that the third stage is the ultimate possible achievement in evaluation theory. Third-stage theorists disagree with each other, so that more integration is still needed; alternatives (additional second-stage theories) continue to be generated that existing syntheses must take into account; and even the best third-stage theories need improvement, especially in the degree to which each is buttressed by empirical evidence – the latter being mostly lacking in all theories of evaluation.

Despite the oversimplifications, these stages accurately reflect general trends in the field:

- Evaluation started with theories that emphasized a search for truth about effective solutions to social problems.
- It next generated many alternatives predicated on detailed knowledge of how organizations in the public sector operate, aimed at producing politically and socially useful results.
- It then produced theories that tried to integrate the alternatives generated in the first two stages.