Shadish, W.R. Jr., Cook, T.D., & Leviton, L.C. (1991). Chapter 1: Social program evaluation: Its history, tasks, and theory. *Foundations of Program Evaluation: Theories of Practice.* Newbury Park, CA: Sage.

1

Social Program Evaluation: Its History, Tasks, and Theory

We can evaluate anything—including evaluation itself. But this book is about evaluating social programs. Social programs, and the policies that spawn and justify them, aim to improve the welfare of individuals, organizations, and society. Hence it is useful to assess how much any social program improves welfare, how it does so, and how it can do so more effectively.

The need for such feedback is pressing because we have few clear, agreed-upon criteria for judging the worth of social activities. What, after all, is the "improvement of welfare"? Whatever it is, it is not as clear or widely accepted as the profit criterion we use to evaluate business activity in the private sector. Even in the rare social program where most of the interested parties think the same criterion is most important, it will rarely have the convenient properties of profit—its simple metric in dollars and cents, or its plasticity for combining outcomes as disparate as the number of people employed, the average hours worked, and the volume of goods shipped. Finally, most of us have an intuitive sense of what money "means," and of the value of different sums of money. This is less often true with social program outcomes. The intuitive meaning and comparative worth of different amounts of family cohesion, social mobility, or relative deprivation are not readily apparent. Moreover, in day-care centers, school reform, or job training, complete monetarization of important outcomes is rarely possible. Criteria other than profit must be justified, measures other than money must be constructed, and means of finding improvements must be developed.

For the last 20 years, practitioners and scholars in program evaluation have addressed these and related tasks. At first, they borrowed concepts and practices from other fields, especially from the academic disciplines in which they were trained. But as experience with program evaluation accumulated, evaluators adapted those concepts and methods, invented others, and combined them in new ways of practicing evaluation. More

than anything else, social program evaluation is a practice-driven field. Its greatest need is for advice about how to perform its special professional tasks.

Program evaluation is like medicine or engineering in emphasizing the practical. Nonetheless, all three fields rely heavily on theories that are not immediately practical. Physicians learn about basic anatomy and physiology not for tools of practice, but to understand the systems in which their practice occurs, influencing what is possible or efficacious. Engineers learn about physics not for rules about building rockets but because physics gives them concepts they need to understand and solve problems in their work. Similar observations hold in other practice-driven fields. They draw much basic theory from other disciplines, but use this knowledge and their own problem-solving experiences to develop specialty theories specifically adapted to the practical demands of their work. It would be a serious mistake to suggest that medicine or engineering could do without such basic theories. It is just as serious a mistake to overlook the importance of theory in program evaluation. The major purposes of this book are to suggest some necessary basic theories underlying practical program evaluation and to review some existing approaches to evaluation with these theories as a comparative standard.

Most current books on evaluation give theory short shrift. Instead, they mostly tend to be atheoretical listings of methods rather than theory-based prescriptions about how and when various methods should be used in practice. Such books can be misleading in the practices they prescribe. They can overlook important options that a complete theory would contain, recommend practices that are ill suited to situations evaluators face, and fail to tell evaluators about why certain practices are worth adopting over alternatives. Of course, evaluation has moved toward greater theoretical breadth and sophistication over the last two decades, and many topics raised in this book have been debated by others. But few evaluation texts contain detailed theoretical rationales for their practice prescriptions. Even the best of them are not as comprehensive as they might be in the theoretical matters they address (Cronbach, 1982a; Cronbach et al., 1980; Rossi & Freeman, 1985). We hope this book lays the groundwork for developing the next generation of evaluation theory.

THE CONTEXT AND HISTORY OF SOCIAL PROGRAM EVALUATION

Intrinsic to evaluation is an idealized problem-solving sequence for (a) identifying a problem, (b) generating and implementing alternatives to

reduce its symptoms, (c) evaluating these alternatives, and then (d) adopting those that results suggest will reduce the problem satisfactorily. For millennia, humans have been doing these tasks to identify options for improving their lives. From an evolutionary perspective, a noncognitive form of evaluation must have been involved in "inventing" *Homo sapiens.* To survive changes in their environment, species must (a) detect the change, (b) generate mutations to reduce any danger, (c) evaluate whether a given mutation reduces the problem, and (d) store genetic information about options that work in the new environment. In evolutionary biology, evaluation is not a cognitive process, as it is in social affairs. In evolutionary biology the criterion for success is clear: Does a morphological change raise the likelihood of passing on the genes responsible for the change? In social affairs we rarely have such clear and urgent criteria.

In this evolutionary conception, the third task is most explicitly evaluative. But its efficacy depends on other tasks being done well. What good is a fine evaluation of an important attempt to solve a trivial problem? What good is a fine evaluation of a puny attempt to solve a serious problem? What good is a fine evaluation of a program that solves a serious problem if the results are not stored and used to ameliorate the problem? Evaluation is just one part of a complex, interdependent, nonlinear set of problem-solving activities. Such evaluations have always been with us and always will be, for problems will always occur, solutions will always need to be generated, tests of their efficacy will need to be done, and the test results will have to be stored if they are to help.

In social life most decisions involve more foresight, planning, and reflection than in biological evolution. Planful social evaluation has been noted as early as 2200 B.C., with personnel selection in China (Guba & Lincoln, 1981; Wortman, 1983b; but see Bowman, 1989). Evaluations during the last 200 years have also been chronicled (Cronbach et al., 1980; Madaus, Stufflebeam, & Scriven, 1983; Rossi & Freeman, 1985; Weiss, 1978). Our concern is with evaluation theories and practices after about 1960, especially concerning the assessment of social policy. Those efforts have intellectual roots in earlier work, especially by Tyler (1935) in education, Lewin (1948) in social psychology, and Lazarsfeld (Lazarsfeld & Rosenberg, 1955) in sociology. They also have roots in the rapid economic growth in the United States after World War II, in the interventionist role that the U.S. federal government took on in social policy during the 1960s, and in the increasing numbers of social science graduates who became interested in policy analysis, survey research, field experiments, and ethnography. All this set the stage for modern program evaluation.

The 1960s and the Great Society

Modern social program evaluation emerged in the 1960s. Its growth is largely due to the social programs initiated under President Kennedy and expanded under Presidents Johnson and Nixon. Social programs were launched in education, income maintenance, housing, health, and criminal justice, primarily in the hope of protecting Americans against the negative effects of poverty. Most programs were launched with high hopes, great dispatch, and enormous financial investment. Bell (1983) estimates that the dollars spent on all public social programs, from social security through public aid programs, increased from $23.5 billion in 1950 to $428.3 billion in 1979—an increase of 1,800% before inflation is considered, or 600% after inflation. As a percentage of the gross national product, funds for social welfare programs doubled during that time. Social insurance programs, primarily social security, accounted for most of this, going from $5 billion to $160 billion between 1950 and 1979. Spending on all public aid programs, such as Aid to Families with Dependent Children (AFDC) and food stamps, increased from $2.5 billion to $40 billion; spending on public health and medical programs such as Medicaid and Medicare increased from $3 billion to $87 billion. Even adjusted for inflation, this financial investment is large by any standard.

These huge investments raised important issues. Congress is concerned with holding the recipients of federal funds accountable for their disbursement, especially with estimating expenditure patterns and preventing inappropriate payments. But Congress also cares that programs bring about some of the expected effects while avoiding important negative side effects. Until the social programs of the 1960s, these congressional functions existed more on paper than in practice. The rapid growth of federal funding for social programs, media reports of fraud, abuse, and mismanagement, and popular apprehensions about the legitimacy of social welfare programming led many in Congress to want more teeth in these oversight and accountability functions. These concerns increased over time, as did defense budgets, inflation rates, and deficits. In such an environment, proponents of social programs faced more calls to show that program funds had been spent as intended and in ways that caused desirable results.

Other political concerns also pushed toward evaluation. Some observers thought that some local projects were being implemented in ways that did not respond to federal intents (Cronbach et al., 1980; Cumming, 1976; House, 1980). Federal and state administrators and legislators wanted more leverage over these projects to ensure that the federal will held more sway. Conversely, others saw the new federal initiatives as threatening local control; they wanted to document beneficial effects from locally

controlled social programs (Feeley & Sarat, 1980). Both groups assigned evaluation a role – to control local variation in one case and to glorify it in the other.

Other concerns were managerial. Massive federal involvement in social welfare was a new phenomenon; few managers had experience with it. It was a formidable task just to learn which activities were actually implemented under a new social program. Since such implementation is a precondition for program effectiveness, managers wanted such data to manage programs better and to respond to varied information requests from Congress, administration executives, local constituencies, and media professionals (Wholey, 1983).

Other concerns were intellectual. Social critics quickly saw that some social programs had problems. They sought to discuss the process of solving social problems that undergirded the design of these programs, and to critique old assumptions and develop new ones. To do this they wanted assessments of how well programs were doing, data about why successes and failures occurred, and identification of nongovernment paths through which the goals of social programs might be obtained. These aims transcended the evaluation of individual programs, but evaluation of each program was necessary input.

Who Would Respond to These Concerns?

The public sector lacked sufficient established mechanisms to respond to these concerns. Social programs lacked the well-developed roles and professions that provide this feedback in the private sector. Independent accountants from certified firms regularly check company books to determine profits and tax liabilities. Their results are made public so that stockholders and IRS officials can judge how well corporate obligations are met. Sometimes they provide such data for individual offices or divisions in a company, to diagnose problems and achievements so as to improve operations. Management consulting firms also try to improve organizational functioning; their work is broader than accounting and not restricted to auditing financial records. Many corporations have their own research and development departments charged with basic research, product development, and test marketing. In the private sector, these three functions – summarizing achievements, improving operations, and designing new products – are generally recognized as important, evaluation-dependent activities. It is no accident that professional accountants, management consultants, and research and development specialists evolved to do them.

The public sector employs some specialists to serve these three functions—auditors at the Internal Revenue Service and the U.S. General Accounting Office, economists in executive branch agencies, planning and systems analysts in the Department of Defense, and budget specialists everywhere. But they were overwhelmed when it came to providing feedback about social programming. Too few personnel existed in government to meet the demands for evaluation; many were often in the wrong departments. For example, planning and systems analysts were mostly in the Department of Defense prior to the Johnson administration. The government either had to move these people—hard to do when demands on Defense were also increasing—or look outside itself to managers of private sector corporations for organizational expertise; to financial experts for accounting, planning, budgeting, and auditing; and to academicians for scientific methods for evaluation.

In addition, skills of existing government employees were partly irrelevant to giving feedback about social programs. Planning and systems analysts could forecast the effects of new initiatives, but had little training for providing retrospective, grounded evaluations of the operations and consequences of existing social programs (Wholey, Scanlon, Duffy, Fukumoto, & Vogt, 1970). Economists could easily analyze the cost-benefits of a water project, but had less idea how to measure social outcomes such as increased family stability or education. Consequently, existing methodologies in accounting, auditing, surveying, and forecasting yielded less confident conclusions in the social sector. Moreover, strong theories in physical science aid problem solving in engineering programs, and similar theories in biological sciences aid medicine, but few strong theories exist in the social sciences. This made the design of social programs for social problem solving more difficult.

By the late 1960s, demand for feedback about social programs exceeded the supply of personnel with appropriate skills. That demand swept into evaluation many graduates of professional schools and social science departments. Graduate education rapidly expanded during this period. From 1955 to 1974, annual U.S. production of M.A. and M.S. degrees rose 479%, from 58,000 to 278,000; the number of doctorates rose 375%, from 8,800 to 33,000 (Keller, 1983). Increased doctoral production in social sciences (mostly economics, education, political science, psychology, and sociology) was even more dramatic. Between 1960 and 1970 alone, U.S. doctoral production in these disciplines increased 333%, from 2,845 to 9,463; between 1950 and 1986 doctoral production in social science increased 895%, from 1,469 to 13,153 (U.S. Bureau of the Census, 1951, 1962, 1972, 1989). Employment in academia did not keep pace with this rapid increase. From 1955 to 1974, faculty in U.S. colleges and

universities increased 238%, from 266,000 to 633,000 (Keller, 1983). Data about increases in social science faculty during this period are more difficult to locate, because early Census Bureau reports did not always identify social scientists as a separate category. But we do know that from 1965 to 1985, social science faculty increased 228%, from 42,283 to 96,300, with most of this increase taking place by 1980 (U.S. Bureau of the Census, 1972, 1989). Professional evaluation became a viable career alternative to academic employment. Thus evaluation met a need of the day, and a supply of labor existed to conduct its tasks, which led to a profession of evaluation.

The Structural Base of the Profession of Evaluation

Evaluation is a profession in the sense that it shares certain attributes with other professions and differs from purely academic specialties such as psychology or sociology. Although they may have academic roots and members, professions are economically and socially structured to be devoted primarily to practical application of knowledge in a circumscribed domain with socially legitimated funding (Austin, 1981). Professionals have somewhat greater constraints than academics about which tasks they can undertake, and they tend to develop standards of practice, codes of ethics, and other professional trappings. Program evaluation is not fully professionalized, like medicine or the law; it has no licensure laws, for example. But it tends toward professionalization more than most disciplines.

Probably the key impetus to the establishment of professional evaluation was federal legislation mandating it and supplying funding for it. We can trace this development only inexactly. Weiss (1987b) claims that "the first Federal program to require evaluation was the juvenile delinquency program enacted by Congress in 1962" (p. 40), but funding for this evaluation was modest by later standards. Wholey (1986b) says that "program evaluation has been an important component of Federal employment and training programs since the Manpower Development and Training Act of 1962, [and] an important component of Federal antipoverty programs since the Economic Opportunity Act of 1964" (p. 6). Wholey and White (1973) describe the educational evaluation requirement generally referred to as "Title I" as the "grand-daddy" of them all – referring to Senator Robert Kennedy's 1965 evaluation rider to the Title I (compensatory education) section of the Elementary and Secondary Education Act (House, 1980). This act (now called Chapter 1) provided major funding for evaluation. Another oft-noted milestone was the 1968 federal RFP

(Request for Proposal) for evaluating Head Start (Wortman, 1983b); tl Office of Economic Opportunity (OEO) supported 13 university-base evaluations of Head Start as early as 1966 (Wholey et al., 1970). Al in 1968, Stanford Research Institute got funds to evaluate the Follo Through program (Wholey et al., 1970). Boruch and Cordray (1980) four 119 federal education statutes using the word *evaluation* between 19 and 1978—and their sample did not exhaust all education statutes.

Evaluation quickly spread to other social sectors, especially at tl federal level (Comptroller General, 1980a, 1980b, 1980c). Wholey et a (1970) note that Congress earmarked funds for evaluations, or otherwi authorized them, in (a) the 1967 amendments to the Economic Opportuni Act for Community Action Programs and for Job Corps, (b) the 1967 Chi Health Act, (c) the 1967 Elementary and Secondary Education Amen ments, (d) the 1967 Partnership for Health Amendments, and (e) the 19 Vocational Rehabilitation Amendments. In 1967, the assistant secreta for planning and evaluation in the Department of Health, Education a Welfare (HEW; now Health and Human Services, or HHS) was giv responsibility for departmentwide improvements in evaluation. In tl same year, the Community Action Program funded an impact evaluati of its own programs. By 1968, Congress had funded evaluation in HE under 11 pieces of legislation, and OEO had a well-developed evaluati plan to evaluate Head Start, Upward Bound, the Community Action Pr gram, neighborhood health centers, and family planning and legal servic programs.

Many of these mandates were accompanied by funds specifically a propriated for evaluation. How many dollars were allocated and spent federal, state, and local levels is unclear. Wholey et al. (1970) report th in fiscal year 1969, $17 million was obligated to federal evaluati contracts and grants (as opposed to demonstration projects) for 12 pr grams in four federal agencies. By fiscal year 1972 this figure increas to roughly $100 million (Buchanan & Wholey, 1972): $8 million in t Manpower Administration of the Department of Labor, $11 million f education in HEW, $2.5 million for child development in HEW, $2 million for health services in HEW, $6-7 million for drug abuse a alcoholism in HEW, $11 million for crime control in HEW, $25 milli for income maintenance in HEW, and $10 million for housing in t Department of Housing and Urban Development. HEW alone spent abc $40 million a year on evaluation between 1978 and 1980 (Wholey, 198 and more than $50 million in 1980 (Abramson & Wholey, 1981). Over federal agencies, $177.8 million was obligated for evaluation in fis years 1975-1977 (Abramson, 1978; cited in Cronbach et al., 1980). Final a U.S. General Accounting Office (1982) survey of federal evaluati

activities (excluding both Defense and all funds for audits and routine management information) found "228 separate evaluation organizations, employing about 1,400 highly trained professionals and spending about $180 million in fiscal year 1980" (p. ii), producing 2,362 evaluations. These figures are not always mutually consistent. What is clear is that a good deal of federal money was being spent on evaluation.

In addition to federal dollars, state and local funds finance evaluations, but how much is not clear. Wholey and White (1973) claim local evaluation accounted for most Title I evaluation expenditures, and exceeded federal Title I evaluation funds. Similarly, in mental health, the National Institute of Mental Health spent $11 million between 1969 and 1981 evaluating community mental health center (CMHC) programs (Williams & Light, 1982); local CMHCs supplemented this with between $3.7 million and $24.3 million per year on self-evaluation between 1976 and 1980 (Neigher, 1982).

These funds were earmarked for evaluation. If one adds federal investment in demonstration projects and applied research, some of which is evaluative, federal investment in evaluation for 1975-1977 exceeded $3 billion (Abramson, 1978; cited in Cronbach et al., 1980). Clear financial, legislative, and administrative incentives were present for people to do evaluations.

The 1980s saw declines in evaluation funding and activities. Cronbach claims the decline began in the late 1970s and was in full swing by the 1980s, given momentum by budget cuts of the Reagan administration. Federal evaluation offices were hit hard. Evaluation-related "fiscal measures" at the Department of Education's Office of Planning, Budget, and Evaluation (OPBE) declined 62% in constant dollars between 1980 and 1984 (Cordray & Lipsey, 1987). OPBE conducted 114 evaluation studies in 1980, but only 11 in 1984. Funding for statistics in OPBE was down 8% in 1984 compared with 1980, and personnel declined by 12% (Cordray & Lipsey, 1987). In 1984, $111 million was available to assess the results of all domestic programs. The decline slowed in the late 1980s. Funds for evaluation decreased 37% from 1980 to 1984 (in constant dollars), but fell only 6% from 1984 to 1988. Manpower dropped proportionally. Professional evaluation staff decreased 52% from 1980 to 1988 in 15 evaluation units, 12% of which occurred between 1984 and 1988 (U.S. General Accounting Office, 1988). Some signs of decline continue today. However, lively evaluation activity continues on many fronts, often linked to substantive disciplines rather than to organizations with the word *evaluation* in their names. These include the growing use of passive longitudinal data systems like the National Assessment of Educational Progress, evaluations of large-scale health promotion and disease prevention programs

(Braverman, 1989), international evaluations (Conner & Hendricks, 1989), and evaluations of poverty and labor force participation programs. Despite declines, therefore, much evaluation is probably still occurring.

The funding and activity described above gave evaluation credibility. Wholey and White (1973) say of the effects of mandated local evaluation in education, "The major impact of all this local Title I evaluation activity has been an increasing acceptance of evaluation; 'Evaluation' is now a household word among educators" (p. 73). Federal initiatives did more than give credibility. Much as clinical psychology became a profession through federal initiatives after World War II (Sarason, 1981), so evaluators received the necessary institutional support for a new profession responsible for certain tasks (feedback about social programs) in a sector (public social programming), with a funding base (reimbursement for evaluative functions) and social legitimation (through government need for evaluation) (Austin, 1981). Without this support, evaluation might exist only as a small, applied academic discipline, like community psychology (Heller & Monahan, 1977) or clinical sociology (Glassner, 1981).

Responses by Evaluators

Professionalization of evaluation. Professions require a unique and transmittable knowledge base. The early knowledge base of evaluation borrowed heavily from existing methods and theories in nearly all social sciences. More specialized knowledge bases emerged as experience with evaluation increased. Explicating these evaluation-specific knowledge bases is the core of this book. To develop and transmit this knowledge, some universities started evaluation centers or degree-conferring programs (Evaluation Research Society, 1980). Wortman (1983a) claims the first of these programs began admitting students in 1973—presumably referring to the Northwestern University doctoral and postdoctoral evaluation training program, where he was a faculty member. Other universities trained evaluators in related doctoral programs. For example, between 1966 and 1986 the Center for Instructional Research and Curriculum Evaluation (CIRCE) at the University of Illinois produced 49 doctorates specializing in measurement and evaluation. Many professional evaluators do not have doctorates, but most have at least master's degree training in related fields, such as public administration (Shadish & Epstein, 1987).

Another indicator of professionalization is creation of yearbooks, societies, journals, and professional standards (Austin, 1981). The field's yearbook was the *Evaluation Studies Review Annual,* first published in 1976 (Glass, 1976) and each year after until 1987 (Shadish & Reichardt, 1987). Evaluation journals exist for the field at large (*Evaluation Review,*

Evaluation Practice, Evaluation and Program Planning), and for specific sectors (*Evaluation and the Health Professions, Evaluation and Educational Policy*). In 1976 two professional societies were founded—Evaluation Research Society and Evaluation Network; in 1985 they merged to form the American Evaluation Association, with about 3,000 members and annual meetings attended by 500-1,000 people (American Evaluation Association, 1986). Finally, the field developed standards for practice that imply minimal levels of competence for evaluators (Rossi, 1982b).

Diversity of professional practice. Evaluators responded to government requests for evaluation in three ways (Cook & Buccino, 1979). First, some contract research firms quickly specialized in bidding for evaluations. Some grew to include 800 Ph.D.-level professionals. In 1970 more than 300 firms were qualified to receive federal evaluation RFPs (Wholey et al., 1970). Second, university researchers won evaluation contracts and grants, consulted with contract research firms, and developed evaluation theory and methods. Third, specific evaluation offices were established in federal, state, and local agencies to respond quickly and pointedly to managers' information needs.

Across these settings, evaluators did an enormous number of studies. The U.S. General Accounting Office compiled a three-volume index of completed federal evaluations, and found 5,610 of them between 1973 and 1979 (Comptroller General, 1980a, 1980b, 1980c). Aaronson and Wilner (1983) found 3,027 local mental health center evaluations—a nonexhaustive sample from the Databank of Program Evaluations at UCLA. The validity of these figures is unclear, and depends on how evaluation is defined. But clearly tens of thousands of evaluations may have been done.

These evaluations were diverse in many ways. They were diverse in the level of government to which they responded. Some responded to federal mandates; an example in education is the evaluation of the national Follow Through program (House, 1980). Others responded to such state mandates as Chapter 328 of the Laws of 1976 of the State of New York authorizing mental health evaluation (Landsberg, Neigher, Hammer, Windle, & Woy, 1979). Still others responded to local project managers; an example is the locally controlled community mental health center evaluations mandated by the 1975 amendments to the Community Mental Health Centers Act (Cook & Shadish, 1982). Evaluators were also diverse in their substantive areas, including education, public health, criminal justice, medicine, labor force participation, income support, nutrition, traffic safety, international aid, mental health, and many other sectors. Indeed, a major task for evaluators is keeping up with the accomplishments in these areas (Light, 1983; Shadish & Reichardt, 1987).

But more than anything, evaluators were so diverse in the activities they conducted that it is often hard to see what those activities share. Some evaluators constructed management information systems (MIS) to provide timely data about program operations; mental health evaluators were even told that "an MIS is prerequisite to undertaking formal program evaluation" (Landsberg et al., 1979, p. 5). But other evaluators ignored this supposed prerequisite entirely. Others conducted case studies and participant observation in the traditions of sociology and anthropology (Guba & Lincoln, 1981). Some evaluators did huge social experiments in which units were randomly assigned to different treatments, as with the New Jersey Negative Income Tax Experiment (Rossi & Lyall, 1978) or the Manhattan Bail Bond Experiment (Ares, Rankin, & Sturz, 1963; Botein, 1965). All this diversity led one reviewer to say, "Evaluation—more than any science—is what people say it is; and people currently are saying it is many different things" (Glass & Ellett, 1980, p. 211).

Diversity was also fostered by funding agencies that demanded different activities under the rubric of evaluation. Evaluators responded with different disciplinary frameworks and methods, and the programs studied had substantive theoretical ties to many different social science and professional fields. This diversity continues to the present day, making it difficult for evaluators to agree on what the practice of evaluation should be, and why. Such differences of opinion are reflected in lively debate about evaluation definitions, models, and methods, debates that have intensified as experience in evaluation has grown (Cook & Reichardt, 1979; Glass & Ellett, 1980; House, 1980; Lewy & Shye, 1978; Madaus, Scriven, & Stufflebeam, 1983; Stufflebeam & Webster, 1981). Gradually, comprehensive theories of social program evaluation tried to integrate this diversity into a coherent whole to help practitioners understand the field and improve their practice.

From Diversity to Integration in Evaluation Theories

What do we mean by theory? No single understanding of the term is widely accepted. *Theory* connotes a body of knowledge that organizes, categorizes, describes, predicts, explains, and otherwise aids in understanding and controlling a topic. Theories do this many ways, such as searching for invariant laws, using definitions and axioms to deduce testable propositions, and describing the causal processes that mediate a relationship (Reynolds, 1971). Our conceptualization is closest to this last kind (Bhaskar, 1979, 1982). The ideal (never achievable) evaluation theory would describe and justify why certain evaluation practices lead to

particular kinds of results across situations that evaluators confront. It would (a) clarify the activities, processes, and goals of evaluation; (b) explicate relationships among evaluative activities and the processes and goals they facilitate; and (c) empirically test propositions to identify and address those that conflict with research or other critically appraised knowledge about evaluation.

Toward unique evaluation theories. As evaluation matured, its theory took on its own special character that resulted from the interplay among problems uncovered by practitioners, the solutions they tried, and traditions of the academic discipline of each evaluator, winnowed by 20 years of experience (Shadish & Reichardt, 1987). Out of this developed a specialty-centered theory. As a specialty, evaluation is most like methodological specialties—ethnography, psychometrics, experimental design, or survey research. But even the narrow specialty of psychometrics uses many kinds of theory—not just statistical theory but also theory about the nature of data (Coombs, 1964) and the role of measurement in applied decisions (American Psychological Association, 1985). Ethnography aspires to broader goals, so it needs a broader theory base about what questions to ask, how to implement studies, and what the researcher's role is in those studies. Evaluation may be the broadest methodological specialty. Its theory includes a vast array of decisions about the shape, conduct, and effects of an evaluation. Evaluation theory is about methods, but not just methods. To inform evaluators about choosing methods, it needs to discuss philosophy of science, public policy, value theory, and theory of use.

Without its unique theories, program evaluation would be just a set of loosely conglomerated researchers with principal allegiances to diverse disciplines, seeking to apply social science methods to studying social programs. Program evaluation is more than this, more than applied methodology. Program evaluators are slowly developing a unique body of knowledge that differentiates evaluation from other specialties while corroborating its standing among them. Evaluation is diverse in many ways, but its potential for intellectual unity is in what Scriven calls "the logic of evaluation" (see Chapter 3), which might bridge disciplinary boundaries separating evaluators.

Developmental trends in evaluation theory. Early theories were based on little experience and so were naive about how research fits into social policy. For example, early theories were more concerned with methods than with the politics of applying methods in field settings, partly because the methodological problems were so pressing that it took time for political factors to make their full force known. As experience with evaluation accumulated, however, this kind of craft knowledge was gradually incor-

porated into the theoretical literature. Hence some early theories in this book may strike the reader as naive given what is known today — unavoidable in an account of the intellectual history of a field.

The field's first theoretical integration was by Suchman (1967), whose ideas overlapped with the more influential views of Campbell (1969, 1971). Both were more interested in summarizing achievements of existing programs in the public sector for policymakers than in collecting information to help practitioners at the local level. Their greatest interest was in evaluating demonstrations of new ideas that might be incorporated into existing or new programs. Reform, testing bold new approaches, was the watchword; evaluating marginal change in existing programs was less prized; evaluating local practice for local reasons was mostly ignored.

Over time, evaluation theories changed and diversified to reflect accumulating practical experience. Exclusive reliance on studying outcomes yielded to inclusive concern with examining the quality of program implementation and the causal processes that mediated any program impacts (Sechrest, West, Phillips, Redner, & Yeaton, 1979). Exclusive reliance on quantitative studies yielded to including qualitative methods (Guba & Lincoln, 1981). Using policymakers as both the source of evaluation questions and the audience for the results yielded to consideration of multiple stakeholder groups (stakeholders are those who have a stake in the program or its evaluation) (Weiss, 1983a, 1983b). Concern with methodology gave way to concern with the context of evaluation practice and to fitting evaluation results into highly politicized and decentralized systems (Cronbach et al., 1980). Today, modern evaluation theories cover more topics, have a better sense of the complexities that plague evaluation practice, and better integrate the diverse concepts, methods, and practices that span the field (see, e.g., Cronbach, 1982b; Rossi & Freeman, 1985).

Our review of theory in this book. This book documents and analyzes these kinds of changes in evaluation theory. We intentionally show the field's theoretical development by describing and analyzing seven theories that were constructed between about 1965 and 1990. Through this vehicle we show how assumptions and prescriptions have changed about five fundamental issues that undergird practical program evaluation:

(1) *social programming:* the ways that social programs and policies develop, improve, and change, especially in regard to social problems
(2) *knowledge construction:* the ways researchers learn about social action
(3) *valuing:* the ways value can be attached to program descriptions
(4) *knowledge use:* the ways social science information is used to modify programs and policies
(5) *evaluation practice:* the tactics and strategies evaluators follow in their professional work, especially given the constraints they face

We refer to these five fundamental matters repeatedly throughout this book. They justify the need for five theoretical bases essential to a good theory of evaluation. Consequently, these five theoretical bases are the core topics of a critical analysis of evaluation theory presented in Chapter 2.

We apply this critical framework to the work of seven evaluation theorists whose writings are scattered across journals and books in many areas. Consequently, many evaluators have not read most of these works, and so may not be as well grounded as they could be in the scholarly thinking of the field (Shadish & Epstein, 1987). By summarizing each author's major points, often verbatim, we offer evaluators the chance to study a broad sample of theoretical writings in the field. More important, our critical analysis of each theory helps clarify its strengths and weaknesses and why it differs from other theories.

Why do this? First, many agreements may emerge across theorists concerning the logic and practice of evaluation. For example, most theorists profess dedication to using evaluation for social problem solving, and all imply that their evaluative practices are improvements over gossip and other informal means of evaluating programs. Such agreements are especially important because they have emerged despite radical disagreement among theorists on many matters. Also, critically appraised commonalities have helped set standards for practice, such as the Standards for Evaluation Practice (Rossi, 1982b). Careful analysis may reveal other commonalities.

Second, disagreements among theorists invite attempts to explain and resolve them, and help identify ambiguities in current knowledge and practice. An example is the dispute between advocates of qualitative and quantitative methods for evaluation. One group feels that qualitative methods best serve evaluators in most cases (Guba & Lincoln, 1981; Stake, 1978); they criticize those they believe prefer quantitative methods (Boruch & Cordray, 1980). Both groups include reasonable and informed scholars; attempts to resolve their differences have advanced understanding of method choices in evaluation (Cook & Reichardt, 1979).

Third, critical analysis reveals areas not addressed well by any theorist. Cronbach, for example, found that few extant theories of evaluation emphasized generalizability or cogently discussed the relationship between evaluation and policy-making (Stanford Evaluation Consortium, 1976). In his effort to resolve this problem, Cronbach began by surfacing these underemphasized concerns (Cronbach et al., 1980). We hope this book will identify other omissions.

Fourth, we hope to demonstrate how evaluation theory evolved in response to experience in doing evaluation. Experience is the major means by which theorists ground their writings. Early theorists were disadvantaged compared to later ones because there was less experience to form

the empirical content of their theories. In the 1960s, Campbell (1969) could reasonably think that experimental methods should be widely adopted in field settings to identify effective solutions to social problems. Subsequent experience, however, led to frank acknowledgment of difficulties with field experiments. So other theorists developed theories that place lower value on causal statements (e.g., Cronbach, 1982a). Early theorists are also influenced by experience, and their pronouncements are not cast in stone. Campbell's current theory of evaluation places less stress on experiments and more on mutual criticism of knowledge claims, which he thinks is inadequately practiced in social science (Campbell, 1979b, 1986b, 1987a). Through reflections on such changes in practice and theory, we vicariously benefit from others' experience without recreating their mistakes.

Why write about evaluation theory? Why not write another book on evaluation methods? The most widely used textbooks (e.g., Cook & Campbell, 1979; Rossi & Freeman, 1982, 1985, 1989; Rossi, Freeman, & Wright, 1979), kits (e.g., Herman, 1987; Morris, 1978), and sourcebooks (e.g., Brinkerhoff, Brethower, Hluchyj, & Nowakowski, 1983) in program evaluation deal primarily with methods. Books on evaluation theory—as this one is—have never been as popular (Cronbach et al., 1980). The popularity of methodological books is no surprise, given that program evaluation is a pragmatic activity. Evaluation practitioners need to act and need tools to use in their daily work. Methods texts are their essential references.

Why, then, write about evaluation theory? We do so because there is an imbalance in evaluation between the great attention to methods and the small attention to theoretical issues that guide method choice. No method is appropriate always and everywhere. Always using just one method, such as experiments or case studies, leads to such problems as producing less useful data or reporting inaccurate findings. Evaluation theory tells us when, where, and why some methods should be applied and others not, suggesting sequences in which methods could be applied, ways different methods can be combined, types of questions answered better or less well by a particular method, and benefits to be expected from some methods as opposed to others. Evaluation theories are like military strategy and tactics; methods are like military weapons and logistics. The good commander needs to know strategy and tactics to deploy weapons properly or to organize logistics in different situations. The good evaluator needs theories for the same reasons in choosing and deploying methods. Without thorough grounding in evaluation theory, the evaluator is left to trial and error or to professional lore in learning about appropriate methods. Such on-the-job training is partly feasible for evaluators who remain many years

in one place (say, a local mental health center evaluator who stays in the same position for years), but can be fatally impractical for evaluators whose responsibilities change rapidly and dramatically (for example, an evaluator in a private sector research firm with contracts from diverse sources to evaluate different programs).

At the same time, however, all evaluation practitioners are nascent evaluation theorists. They think about what they are doing, make considered judgments about which methods to use in each situation, weigh advantages and disadvantages of choices they face, and learn from successes and failures in their past evaluations. In fact, the pragmatic concepts developed in practice probably constitute the most important basis for academic theories. This book is meant to encourage the theoretical dispositions of practitioners by expanding their repertoire of methods, challenging the assumptions behind their methodological and strategic decisions, and creating a broader conceptual framework for them to use in their work. Readers should finish this book with increased ability to ask and answer key questions such as these:

(1) *social programming:* What are the important problems this program could address? Can the program be improved? Is it worth doing so? If not, what is worth doing?

(2) *knowledge use:* How can I make sure my results get used quickly to help this program? Do I want to do so? If not, can my evaluation be useful in other ways?

(3) *valuing:* Is this a good program? By which notion of "good"? What justifies the conclusion?

(4) *knowledge construction:* How do I know all this? What counts as a confident answer? What causes that confidence?

(5) *evaluation practice:* Given my limited skills, time, and resources, and given the seemingly unlimited possibilities, how can I narrow my options to do a feasible evaluation? What is my role – educator, methodological expert, judge of program – worth? What questions should I ask, and what methods should I use?

Books on methods rarely answer these questions, because the questions are mostly not methodological. These questions, and attempts to answer them, are the stuff of theory. These questions do not always have one correct answer. The answers often vary from situation to situation, and depend on factors that evaluators can only partly control, such as resources, skills, constraints imposed by funding agencies, and time frame. But even in such cases, by the time readers have finished this book they should better understand the contingencies that bear upon the answers to these questions.