
3. FOUNDATIONAL MODELS FOR 21st CENTURY PROGRAM EVALUATION*

DANIEL L. STUFFLEBEAM

INTRODUCTION

Evaluators today have available many more evaluation approaches than in 1960. As they address the challenges of the 21st century, it is an opportune time to consider what 20th century evaluation developments are valuable for future use and which ones would best be left behind. I have, in this chapter, attempted to sort 22 alternative evaluation approaches into what fishermen sometimes call the “keepers” and the “throwbacks.” More importantly, I have characterized each approach; assessed its strengths and weaknesses; and considered whether, when, and how it is best applied. The reviewed approaches emerged mainly in the U.S. between 1960 and 1999.

20th Century Expansion of Program Evaluation Approaches

Following a period of relative inactivity in the 1950s, a succession of international and national forces stimulated the expansion and development of evaluation theory and practice. The main influences were the efforts to vastly strengthen the U.S. defense system spawned by the Soviet Union's 1957 launching of Sputnik I; the new U.S. laws in the 1960s to equitably serve minorities and persons with disabilities; Federal government evaluation requirements of the Great Society programs initiated in 1965; the U.S. movement begun in the 1970s to hold educational and social organizations accountable for both prudent use of resources and achievement

*This chapter is a condensed version of a manuscript prepared for the Western Michigan University Evaluation Center's Occasional Papers Series.

of objectives; the stress on excellence in the 1980s as a means of increasing U.S. international competitiveness; and the trend in the 1990s for various organizations, both inside and outside the U.S., to employ evaluation to ensure quality, competitiveness, and equity in delivering services. In pursuing reforms, American society has repeatedly pressed schools and colleges, health-care organizations, and various social welfare enterprises to show through evaluation whether or not services and improvement efforts were succeeding.

The development of program evaluation as a field of professional practice was also spurred by a number of seminal writings. These included, in chronological order, publications by Tyler (1942b, 1950, 1966), Campbell and Stanley (1963), Cronbach (1963), Stufflebeam (1966, 1967), Scriven (1967), Stake (1967), Suchman (1967), Alkin (1969), Guba (1969), Provas (1969), Stufflebeam et al. (1971), Parlett and Hamilton (1972), Weiss (1972), Eisner (1975), Glass (1975), Cronbach and Associates (1980), House (1980), and Patton (1980). These and other authors/scholars began to project alternative approaches to program evaluation. Over the years, a rich literature on a wide variety of alternative program evaluation approaches developed (see, for example, Cronbach [1982]; Guba and Lincoln [1981, 1989]; Nave, Miech, and Mosteller [1999], Nevo [1993]; Patton [1982, 1990, 1994, 1997]; Rossi and Freeman [1993]; Schwandt [1984]; Scriven [1991, 1993, 1994a, 1994b, 1994c]; Shadish, Cook, and Leviton [1991]; Smith, M. F. [1989]; Smith, N. L. [1987]; Stake [1975b, 1988, 1995]; Stufflebeam [1997]; Stufflebeam and Shinkfield [1985]; Wholey, Hatry, and Newcomer [1995]; Worthen and Sanders [1987, 1997]).

Evaluation Models And Approaches

The chapter uses the term *evaluation approach* rather than *evaluation model* because, for one reason, the former is broad enough to cover illicit as well as laudatory practices. Also, beyond covering both creditable and noncreditable approaches, some authors of evaluation approaches say that the term *model* is too demanding to cover their published ideas about how to conduct program evaluations. But for these two considerations, the term *model* would have been used to encompass most of the evaluation proposals discussed in this chapter. This is so because most of the presented approaches are idealized or “model” views for conducting program evaluations according to their authors’ beliefs and experiences.

Need to Study Alternative Approaches

The study of alternative evaluation approaches is important for professionalizing program evaluation and for its scientific advancement and operation. Professionally, careful study of program evaluation approaches can help evaluators legitimize approaches that conform with sound principles of evaluation and discredit those that do not. Scientifically, such a review can help evaluation researchers identify, examine, and address conceptual and technical issues pertaining to the development of the evaluation discipline. Operationally, a critical view of alternatives can help evaluators consider, assess, and selectively apply optional evaluation frameworks. The review also provides substance for evaluation training. The main values in studying

alternative program evaluation approaches are to discover their strengths and weaknesses, decide which ones merit substantial use, determine when and how they are best applied, and obtain direction for improving these approaches and devising better alternatives.

The Nature of Program Evaluation

The chapter employs a broad view of program evaluation. It encompasses assessments of any coordinated set of activities directed at achieving goals. Examples are assessments of ongoing, cyclical programs, such as school curricula, food stamps, housing for the homeless, and annual influenza inoculations; time-bounded projects, such as development and dissemination of a fire prevention guide; and national, regional, or state systems of services, such as those provided by regional educational service agencies. Such program evaluations both overlap with and yet are distinguishable from other forms of evaluation, especially evaluations of students, personnel, materials, and institutions.

Previous Classifications Of Alternative Evaluation Approaches

In analyzing the 22 evaluation approaches, prior assessments regarding program evaluation's state of the art were consulted. Stake's (1974) analysis of nine program evaluation approaches provided a useful application of advance organizers (the types of variables used to determine information requirements) for ascertaining different types of program evaluations. Hastings' (1976) review of the growth of evaluation theory and practice helped to place the evaluation field in a historical perspective. Guba's (1976) presentation and assessment of six major philosophies in evaluation was provocative. House's (1983) analysis of approaches illuminated important philosophical and theoretical distinctions. Scriven's (1991, 1994a) writings on the transdiscipline of evaluation helped to sort out different evaluation approaches; it was also invaluable in seeing the approaches in the broader context of evaluations focused on various objects other than programs. The book *Evaluation Models* (Madaus, Scriven, & Stufflebeam, 1983) provided a previous inventory and analysis of evaluation models. All of the assessments helped sharpen the issues addressed.

Program Evaluation Defined

In characterizing and assessing evaluation approaches, the various kinds of activities conducted in the name of program evaluation were classified on the basis of their degree of conformity to a particular definition of evaluation. In this chapter, evaluation means *a study designed and conducted to assist some audience to assess an object's merit and worth*. This definition should be widely acceptable as it agrees with common dictionary definitions of evaluation; it is also consistent with the definition that underlies published sets of professional standards for evaluations (Joint Committee on Standards for Educational Evaluation, 1981b, 1994). However, it will become apparent that many studies done in the name of program evaluation either do not conform to this definition or directly oppose it.

Classification and Analysis of the 22 Approaches

Using the above definition of evaluation, program evaluation approaches were classified into four categories. The first category includes approaches that promote invalid or incomplete findings (referred to as pseudoevaluations), while the other three include approaches that agree, more or less, with the definition (i.e., Questions/Methods-Oriented, Improvement/ Accountability, and Social Agenda/Advocacy). Of the 22 program evaluation approaches that are described, two are classified as pseudoevaluations, thirteen as questions/methods-oriented approaches, three as improvement/accountability-oriented approaches, and four as social agenda/advocacy-directed approaches.

Each approach is analyzed in terms of ten descriptors: (1) advance organizers, that is, the main cues that evaluators use to set up a study; (2) main purpose(s) served; (3) sources of questions addressed; (4) questions that are characteristic of each study type; (5) methods typically employed; (6) persons who pioneered in conceptualizing each study type; (7) other persons who have extended development and use of each study type; (8) key considerations in determining when to use each approach; (9) strengths of the approach; and (10) weaknesses of the approach. Using these descriptors, comments on each of the 22 program evaluation approaches are presented. These assessments are then used to reach conclusions about which approaches should be avoided, which are most meritorious, and under what circumstances the worthy approaches are best applied.

Caveats

I acknowledge, without apology, that the assessments of approaches and the entries in the summary chart in this chapter are based on my best judgments. I have taken no poll, and no definitive research exists, to represent a consensus on the characteristics and strengths and weaknesses of the different approaches. My analyses reflect 35 years of experience in applying and studying different evaluation approaches. Hopefully, these will be useful to evaluators and evaluation students at least in the form of working hypotheses to be tested.

I have mainly looked at the approaches as relatively discrete ways to conduct evaluations. In reality, there are many occasions when it is functional to mix and match different approaches. A careful analysis of such combinatorial applications no doubt would produce several hybrid approaches that might merit examination. That analysis is beyond the scope of this chapter.

PSEUDOEVALUATIONS

Because this chapter is focused on describing and assessing the state of the art in evaluation, it is necessary to discuss bad and questionable practices, as well as the best efforts. Evaluators and their clients are sometimes tempted to shade, selectively release, or even falsify findings. While such efforts might look like sound evaluations, they are aptly termed pseudoevaluations if they fail to produce and report to all right-to-know audiences valid assessments of merit and worth.

Pseudoevaluations often are motivated by political objectives. For example, persons holding or seeking authority may present unwarranted claims about their achievements and/or the faults of their opponents, or hide potentially damaging information. These objectionable approaches are presented because they deceive through evaluation and can be used by those in power to mislead constituents or to gain and maintain an unfair advantage over others, especially persons with little power. If evaluators acquiesce to and support pseudoevaluations, they help promote and support injustice, mislead decision making, lower confidence in evaluation services, and discredit the evaluation profession.

Approach 1: Public Relations-Inspired Studies

The public relations approach begins with an intention to use data to convince constituents that a program is sound and effective. Other labels for the approach are “ideological marketing” (see Ferguson, June 1999), advertising, and infomercial. The public relations approach may meet the standard for addressing all right-to-know audiences but fails as a legitimate evaluation approach, because typically it presents a program’s strengths or an exaggerated view of them but not its weaknesses.

The advance organizer is the propagandist’s information needs. The study’s purpose is to help the program director/public relations official project a convincing, positive public image for a program. The guiding questions are derived from the public relations specialists’ and administrators’ conceptions of which questions constituents would find most interesting. In general, the public relations study seeks information that would most help an organization confirm its claims of excellence and secure public support. From the start, this type of study seeks not a valid assessment of merit and worth but information to help the program “put its best foot forward.” Such studies avoid gathering or releasing negative findings.

Typical methods used in public relations studies are biased surveys; inappropriate use of norms tables; biased selection of testimonials and anecdotes; “massaging” of obtained information; selective release of only the positive findings; reporting central tendency, but not variation; cover-up of embarrassing incidents; and the use of “expert” advocate consultants. In contrast to the “critical friends” employed in Australian evaluations, public relations studies use “friendly critics.” A pervasive characteristic of the public relations evaluator’s use of dubious methods is a biased attempt to nurture a good picture for a program. The fatal flaw of built-in bias to report only good things offsets any virtues of this approach. If an organization substitutes biased reporting of only positive findings for balanced evaluations of strengths and weaknesses, it soon will demoralize evaluators who are trying to conduct and report valid evaluations and may discredit its overall practice of evaluation.

By disseminating only positive information on a program’s performance while withholding information on shortcomings and problems, evaluators and clients may mislead taxpayers, constituents, and other stakeholders concerning the program’s true value. The possibility of such positive bias in advocacy evaluations underlies the long-standing policy of Consumers Union not to include advertising by the owners of the products and services being evaluated in its *Consumer Reports* magazine. To main-

tain credibility with consumers, Consumers Union has, for the most part, maintained an independent perspective and a commitment to identify and report both strengths and weaknesses in the items evaluated and not to supplement this information with biased ads. (An exception is that the magazine advertizes its own supplementary publications and services.)

A contact with an urban school district illustrates the public relations type of study. A superintendent requested a community survey for his district. The superintendent said, straightforwardly, that he wanted a survey that would yield a positive report on the district's performance and his leadership. He said such a positive report was desperately needed at the time to restore the confidence of the community in the school district and in him. The superintendent did not get the survey and positive report, and it soon became clear why he thought one was needed. Several weeks after making the request, he was summarily fired. Another example occurred when a large urban school district used one set of national norms to interpret pretest results and another norms table for the posttest. The result was a spurious portrayal and wrong conclusion that the students' test performance had substantially improved between the first and second test administrations.

Evaluators need to be cautious in how they relate to the public relations activities of their sponsors, clients, and supervisors. Certainly, public relations documents will reference information from sound evaluations. Evaluators should persuade their audiences to make honest use of the evaluation findings. Evaluators should not be party to misuses, especially in cases where erroneous reports are issued that predictably will mislead readers to believe that a seriously flawed program is effective. As one safeguard, evaluators can promote and help their clients arrange to have independent metaevaluators examine the organization's production and use of evaluation findings against professional standards for evaluations.

Approach 2: Politically Controlled Studies

The politically controlled study is an approach that can be either defensible or indefensible. A politically controlled study is illicit if the evaluator and/or client (a) withhold the full set of evaluation findings from audiences who have express, legitimate, and legal rights to see the findings; (b) abrogate their prior agreement to fully disclose the evaluation findings; or (c) bias the evaluation message by releasing only part of the findings. It is not legitimate for a client first to agree to make the findings of a commissioned evaluation publicly available and then, having previewed the results, to release none or only part of the findings. If and when a client or evaluator violates the formal written agreement on disseminating findings or applicable law, then the other party has a right to take appropriate actions and/or seek an administrative or legal remedy.

Clients sometimes can legitimately commission covert studies and keep the findings private, while meeting applicable laws and adhering to an appropriate advance agreement with the evaluator. This can be the case in the U.S. for private organizations not governed by public disclosure laws. Furthermore, an evaluator, under

legal contractual agreements, can plan, conduct, and report an evaluation for private purposes, while not disclosing the findings to any outside party. The key to keeping client-controlled studies in legitimate territory is to reach appropriate, legally defensible, advance, written agreements and to adhere to the contractual provisions concerning release of the study's findings. Such studies also have to conform to applicable laws on release of information.

The advance organizers for a politically controlled study include implicit or explicit threats faced by the client for a program evaluation and/or objectives for winning political contests. The client's purpose in commissioning such a study is to secure assistance in acquiring, maintaining, or increasing influence, power, and/or money. The questions addressed are those of interest to the client and special groups that share the client's interests and aims. The main questions of interest to the client are: What is the truth, as best can be determined, surrounding a particular dispute or political situation? What information would be advantageous in a potential conflict situation? What data might be used advantageously in a confrontation? Typical methods of conducting the politically controlled study include covert investigations, simulation studies, private polls, private information files, and selective release of findings. Generally, the client wants information that is as technically sound as possible. However, he or she may also want to withhold findings that do not support his or her position. The strength of the approach is that it stresses the need for accurate information. However, because the client might release information selectively to create or sustain an erroneous picture of a program's merit and worth, might distort or misrepresent the findings, might violate a prior agreement to fully release findings, or might violate a "public's right to know" law, this type of study can degenerate into a pseudoevaluation.

A superintendent of one of the nation's largest public school districts once confided that he possessed an extensive notebook of detailed information about each school building in his district. The information related to student achievement, teacher qualifications, racial mix of teachers and students, average per-pupil expenditure, socioeconomic characteristics of the student body, teachers' average length of tenure in the system, and so forth. The data revealed a highly segregated district with uneven distribution of resources and markedly different achievement levels across schools. When asked why all the notebook's entries were in pencil, the superintendent replied that it was absolutely essential that he be kept informed about the current situation in each school, but that it was also imperative that the community-at-large, the board, and special interest groups in the community, not have access to the information, for any of these groups might point to the district's inequities as a basis for protest and even removing the superintendent. Hence, one special assistant kept the document up-to-date, only one copy existed, and the superintendent kept that locked in his desk. The point of this example is not to negatively judge the superintendent's behavior. However, the superintendent's ongoing covert investigation and selective release of information was decidedly not a case of true evaluation, for what he disclosed to the right-to-know audiences did not fully and honestly inform them about the observed situation in the district. This example

may appropriately be termed a pseudoevaluation because it both underinformed and misinformed the school district's stakeholders.

Still another example was seen when an evaluator gave her superintendent a sound program evaluation report, showing both strengths and weaknesses of the targeted program. The evaluator was surprised and dismayed one week later when the superintendent released to the public a revised version showing only the program's strengths.

Cases like these undoubtedly led to the federal and state sunshine laws in the United States. Under current U.S. and state freedom of information provisions, most information obtained through the use of public funds must be made available to interested and potentially affected citizens. Thus, there exist legal deterrents to and remedies for illicit, politically controlled evaluations that use public funds.

While it would be unrealistic to recommend that administrators and other evaluation users not obtain and selectively employ information for political gain, evaluators should not lend their names and endorsements to evaluations presented by their clients that misrepresent the full set of relevant findings, that present falsified reports aimed at winning political contests, or that violate applicable laws and/or prior formal agreements on release of findings. Despite these warnings, it can be legitimate for evaluators to give private evaluative feedback to clients, provided they conform with applicable laws, statutes, and policies, and sound contractual agreements on release of findings are reached and honored.

QUESTIONS/METHODS-ORIENTED EVALUATION APPROACHES (QUASI-EVALUATION STUDIES)

Questions/methods-oriented program evaluation approaches (1) address specified questions, answers to which may or may not be sufficient to assess a program's merit and worth and/or (2) use some preferred method(s). Whether the questions and methodology are appropriate for developing and supporting value claims is a secondary consideration. These approaches may employ as their starting points operational objectives, standardized measurement devices, cost-analysis procedures, expert judgment, a theory or model of a program, case study procedures, management information systems, designs for controlled experiments, and/or an overriding commitment to employ a mixture of qualitative and quantitative methods. Most emphasize technical quality and posit that it is usually better to answer a few pointed questions well than to attempt a broad assessment of something's merit and worth.

These studies can be called quasi-evaluation studies, because sometimes they happen to provide evidence that fully assesses a program's merit and worth, while in other cases, their focus is too narrow or is only tangential to questions of merit and worth. While the approaches are typically labeled as evaluations, they may or may not meet the requirements of a sound evaluation. Quasi-evaluation studies have legitimate uses apart from their relationship to program evaluation, since they can investigate important though narrow questions. The main caution is that these types of studies not be uncritically equated to evaluation.

Approach 3: Objectives-Based Studies

The objectives-based study is the classic example of a questions/methods-oriented evaluation approach. Stufflebeam and Madaus (1988) provided a comprehensive look at this approach by publishing an edited volume of the classical writings of Ralph W. Tyler. In this approach, some statement of objectives provides the advance organizer. The objectives may be mandated by the client, formulated by the evaluator, or specified by the service providers. The usual purpose of an objectives-based study is to determine whether the program's objectives have been achieved. Typical audiences are program developers, sponsors, and managers who want to know the extent to which each stated objective was achieved.

The methods used in objectives-based studies essentially involve specifying operational objectives and collecting and analyzing pertinent information to determine how well each objective was achieved. A wide range of objective and performance assessments may be employed. Criterion-referenced tests are especially relevant to this evaluation approach.

Ralph Tyler is generally acknowledged to be the pioneer in the objectives-based type of study, although Percy Bridgman and E. L. Thorndike should also be credited (Travers, 1977). Many people have furthered Tylers seminal contribution by developing variations of his evaluation model. These include Bloom et al. (1956), Hammond (1972), Metfessel and Michael (1967), Popham (1969), Provus (1971), and Steinmetz (1983).

The objectives-based approach is especially applicable in assessing tightly focused projects that have clear, supportable objectives. Even then, such studies can be strengthened by judging project objectives against the intended beneficiaries' assessed needs, searching for side effects, and studying the process as well as the outcomes.

The objectives-based study has been the most prevalent approach in program evaluation. It has common-sense appeal; program administrators have had a great amount of experience with it; and it makes use of technologies of behavioral objectives and both norm-referenced and criterion-referenced testing. Common criticisms are that such studies lead to terminal information that is of little use in improving a program or other enterprise; that the information often is far too narrow to constitute a sufficient basis for judging the object's merit and worth; that the studies do not uncover positive and negative side effects; and that they may credit unworthy objectives.

Approach 4: Accountability, Particularly Payment By Results Studies

The accountability study became prominent in the early 1970s. It emerged because of widespread disenchantment with the persistent stream of evaluation reports indicating that almost none of the massive state and federal investments in educational and social programs were making any positive, statistically discernable differences. One proposed solution posited that accountability systems could be initiated to ensure both that service providers would fulfill their responsibilities to improve services and that evaluators would find the programs' effects and determine which persons and groups were succeeding and which were not.

The advance organizers for the accountability study are the persons and groups responsible for producing results, the service providers' work responsibilities, and the expected outcomes. The study's purposes are to provide constituents with an accurate accounting of results; ensure, though something akin to intimidation, that the results are primarily positive; and pinpoint responsibility for good and bad outcomes. Sometimes accountability programs administer sanctions to the responsible service providers, depending on the extent and quality of their services and achievement.

Accountability questions come from the program's constituents and controllers, such as taxpayers; parent groups; school boards; and local, state, and national funding organizations. Their main question concerns whether each involved service provider and organization charged with responsibility for delivering and improving services is carrying out its assignments and achieving all it should, given the resources invested to support the work.

A wide variety of methods have been used to ensure and assess accountability. These include performance contracting; Program Planning and Budgeting Systems (PPBS); Management by Objectives (MBO); Zero Based Budgeting; mandated "program drivers" and indicators; program input, process, output databases; independent goal achievement auditors; procedural compliance audits; peer reviews; merit pay for individuals and/or organizations; collective bargaining agreements; mandated testing programs; institutional report cards/profiles; self-studies; site visits by expert panels; and procedures for auditing the design, process, and results of self-studies. Also included are mandated goals and standards, decentralization and careful definition of responsibility and authority, payment by results, awards and recognition, sanctions, takeover/intervention authority by oversight bodies, and competitive bidding.

Lessinger (1970) is generally acknowledged as a pioneer in the area of accountability. Among those who have extended Lessinger's work are Stenner and Webster (1971), in their development of a handbook for conducting auditing activities, and Kearney, in providing leadership to the Michigan Department of Education in developing the first statewide educational accountability system. Kirst (1990) analyzed the history and diversity of attempts at accountability in education within the following six broad types of accountability: performance reporting, monitoring and compliance with standards or regulations, incentive systems, reliance on the market, changing locus of authority or control of schools, and changing professional roles. A recent major attempt at accountability, involving sanctions, was the Kentucky Instructional Results Information System (Koretz & Barron, 1998). This program's failure was clearly associated with fast pace implementation in advance of validation, reporting and later retraction of flawed results, results that were not comparable to those in other states, payment by results that fostered teaching to tests and other cheating in schools, and heavy expense associated with performance assessments that could not be sustained over time.

Accountability approaches are applicable to organizations and professionals funded and charged to carry out public mandates, deliver public services, and implement

specially funded programs. It behooves these program leaders to maintain a dynamic baseline of information needed to demonstrate fulfillment of responsibilities and achievement of positive results. They should focus accountability mechanisms especially on program elements that can be changed with the prospect of improving outcomes. They should also focus accountability to enhance staff cooperation toward achievement of collective goals rather than to intimidate or stimulate counterproductive competition. Moreover, accountability studies that compare programs should fairly consider the programs' contexts, especially beneficiaries' characteristics and needs, local support, available resources, and external forces.

The main advantages of accountability studies are that they are popular among constituent groups and politicians and are aimed at improving public services. They can also provide program personnel with clear expectations against which to plan, execute, and report on their services and contributions. They can be designed to give service providers freedom to innovate on procedures coupled with clear expectations and requirements for producing and reporting on accomplishments. Further, setting up healthy, fair competition between comparable programs can result in better services and products for consumers.

A main disadvantage is that accountability studies often result in invidious comparisons and thereby produce unhealthy competition and much political unrest and acrimony among service providers and between them and their constituents. Also, accountability studies often focus too narrowly on outcome indicators and can undesirably narrow the range of services. Another disadvantage is that politicians tend to force the implementation of accountability efforts before the needed instruments, scoring rubrics, assessor training, etc. can be planned, developed, field-tested, and validated. Furthermore, prospects for rewards and threats of punishment have often led service providers to cheat in order to assure positive evaluation reports. In schools, cheating to obtain rewards and avoid sanctions has frequently generated bad teaching, bad press, turnover in leadership, and abandonment of the accountability system.

Approach 5: Objective Testing Programs

Since the 1930s, American education has been inundated with standardized, multiple choice, norm-referenced testing programs. Probably every school district in the country has some such program. The tests are administered annually by local school districts and/or state education departments to inform students, parents, educators, and the public at large about the achievements of children and youth. The purposes of testing are to assess the achievements of individual students and groups of students compared to norms and/or standards. Typically, tests are administered to all students in selected grade levels. Because the test results focus on student outcomes and are conveniently available, many educators have tried to use the results to evaluate the quality of special projects and specific school programs by inferring that high scores reflect successful efforts and low scores reflect poor efforts. Such inferences can be wrong if the tests were not targeted on particular project or program

objectives or the needs of particular target groups of students and if students' background characteristics were not taken into account.

Advance organizers for standardized educational tests include areas of the school curriculum, curricular objectives, and specified norm groups. The testing programs' main purposes are to compare the test performance of individual students and groups of students to those of selected norm groups and/or to diagnose shortfalls related to particular objectives. Standardized test results are also often used to compare the performance of programs and schools and to examine achievement trends across years. Metrics used to make the comparisons typically are standardized individual and mean scores for the total test and subtests. The sources of test questions are usually test publishers and test development/selection committees.

The typical question addressed by testing is whether the test performance of individual students is at or above the average performance of local, state, and national norm groups. Other questions may concern the percentages of students who surpassed one or more cut-score standards, where the group of students ranks compared to other groups, and whether achievement is better than in prior years. The main process is to select, administer, score, interpret, and report the tests.

Lindquist (1951), a major pioneer in this area, was instrumental in developing the Iowa testing programs, the American College Testing Program, the National Merit Scholarship Testing Program, and the General Educational Development Testing Program, as well as the Measurement Research Center at the University of Iowa. Many individuals have contributed substantially to the development of educational testing in America, including Ebel (1965), Flanagan (1939), Lord and Novick (1968), and Thorndike (1971). Innovations to testing in the 1990s include the development of item response theory (Hambleton & Swaminathan, 1985) and value-added measurement (Sanders & Horn, 1994; Webster, 1995).

If a school's personnel carefully select tests and use them appropriately to assess and improve student learning and report to the public, the involved expense and effort is highly justified. Student outcome measures for judging specific projects and programs must be validated in terms of the particular objectives and the characteristics and needs of the students served by the program. However, tests should not be relied on exclusively for evaluating specially targeted projects and programs. Results should be interpreted in light of other information on student characteristics, program implementation, student participation, and other outcome measures.

The main advantages of standardized-testing programs are that they are efficient in producing valid and reliable information on student performance in many areas of the school curriculum and that they are a familiar strategy at every level of the school program in virtually all school districts in the United States. The main limitations are that they provide data only about student outcomes; they reinforce students' multiple-choice test-taking behavior rather than their writing and speaking behaviors; they tend to address only lower-order learning objectives; and, in many cases, they are perhaps a better indicator of the socioeconomic levels of the students in a given program, school, or school district than of the quality of teaching and

learning. Stake (1971) and others have argued effectively that standardized tests often are poor approximations of what teachers actually teach. Moreover, as has been patently clear in evaluations of programs for both disadvantaged and gifted students, norm-referenced tests often do not measure achievements well for low and high scoring students. Unfortunately, program evaluators often have made uncritical use of standardized test results to judge a program's outcomes, just because the results were conveniently available and had face validity for the public. Often, the contents of such tests do not match the program's objectives.

Approach 6: Outcomes Evaluation As Value-Added Assessment

Recurrent outcome/value-added assessment is a special case of the use of standardized testing to evaluate the effects of programs and policies. The emphasis is often on annual testing at all or a succession of grade levels to assess trends and partial out effects of the different components of an education system, including groups of schools, individuals and individual teachers. Characteristic of this approach is the annual collection of outcome measures based on standardized indicators, analysis to determine what value is being added to the achievements of students served by particular components of the evaluation system, and reporting of the results for policy, accountability, and improvement purposes. The main interest is in aggregates, not performance of individual students. A state education department may annually collect achievement data from all students (at a succession of grade levels), as is the case in the Tennessee Value-Added Assessment System. The evaluator may analyze the data to look at contrasting results for different schools. Results may be further broken out to make comparisons between curricular areas, teachers, elementary versus middle schools, size and resource classifications of schools, districts, and areas of a state. What differentiates the approach from the typical standardized achievement testing program is the emphasis on sophisticated analysis of data to partial out effects of system components and identify which ones should be improved and which ones should be commended and reinforced. Otherwise, the two approaches have much in common.

Advance organizers in outcome evaluation and employing value-added analysis are indicators of intended outcomes and a scheme for obtaining and classifying gain scores in order to examine policy issues and/or program effects. The purposes of outcome evaluation/value-added assessment systems are to provide direction for policymaking, accountability to constituents, and feedback for improving programs and services. The approach also ensures standardization of data for assessment and improvement throughout a system. The source of questions to be addressed by outcome evaluation originate from governing agencies, funding organizations, policymakers, the system's professionals, and constituents.

One form of outcome evaluation involves value-added assessment, which has been developed by Sanders and Horn (1994); Webster (1995); Webster, Mendro, and Almaguer (1994); and Tymms (1995). Illustrative questions addressed in this form of evaluation are: To what extent are particular programs adding value to students' achievements? What are the cross-year trends in outcomes? In what sectors of the

system is the program working best and poorest? What are key, pervasive shortfalls in particular program objectives that require further study and attention? To what extent are program successes and failures associated with the system's groupings of grade levels? Outcome monitoring involving value-added assessment is probably most appropriate in well-endowed state education departments and large school districts where there is strong support from policy groups, administrators, and service providers to make the approach work. It requires systemwide buy-in; politically effective leaders to continually explain and sell the program; annual testing at a succession of grade levels; a smoothly operating, dynamic, computerized baseline of relevant input and output information; highly skilled technicians to make it run efficiently and accurately; a powerful computer system; complicated statistical analysis; and high-level commitment to use the results for purposes of policy development, accountability, program evaluation, and improvement at all levels of the system.

The central advantage of outcomes monitoring involving value-added assessment is in the systematization and institutionalization of a database of outcomes that can be used over time and in a standardized way to study and find means to improve outcomes. This approach makes efficient use of standardized tests; is amenable to analysis of trends at state, district, school, and classroom levels; uses students as their own controls; and emphasizes service to every student. The approach also is conducive to using a standard of continuous progress across years for every student, as opposed to employing static cut scores. The latter, while prevalent in accountability programs, basically fail to take into account meaningful gains by low or high achieving students, since such gains usually are far removed from the static, cut score standards. Sanders and Horn (1994) have shown that use of static cut scores may produce a "shed pattern," in which students who began below the cut score make the greatest gains while those who started above the cut score standard make little progress. Like the downward slope, from left to right, of a tool shed, the gains are greatest for previously low scoring students and progressively lower for the higher achievers. This suggests that teachers are concentrating mainly on getting students to the cut score standard but not beyond it and thus "holding back the high achievers."

A major disadvantage of the value-added approach is that it is politically volatile, since it is used to identify responsibility for successes and failures down to the levels of schools and teachers. It is also heavily reliant on quantitative information such as that coming from standardized, multiple-choice achievement tests. Consequently, the complex and powerful analyses are based on a limited scope of outcome variables. Nevertheless, Sanders (1989) has argued that a strong body of evidence supports the use of well-constructed, standardized, multiple-choice achievement tests. Beyond the issue of outcome measures, the approach does not provide in-depth documentation of program inputs and processes and makes little if any use of qualitative methods. Despite advancements in objective measurement and the employment of hierarchical mixed models to defensibly partial out effects of a system's organizational components and individual staff members, critics of the approach argue that causal factors

are so complex that no measurement and analysis system can fairly fix responsibility to the level of teachers for the academic progress of individual and collections of students.

Approach 7: Performance Testing

In the 1990s, major efforts were made to offset the limitations of typical multiple-choice tests by employing performance or authentic measures. These devices require students to demonstrate their achievements by producing authentic responses to evaluation tasks, such as written or spoken answers, musical or psychomotor presentations, portfolios of work products, or group solutions to defined problems. Arguments for performance tests are that they have high face validity and model and reinforce students' needed life skills. After all, students are not being taught so that they will do well in choosing best answers from a list, but so that they will master underlying understandings and skills and effectively apply them in real life situations.

The advance organizers in performance assessments are life-skill objectives and content-related performance tasks, plus ways that their achievement can be demonstrated in practice. The main purpose of performance tasks is to compare the performance of individual students and groups of students to model performance on the tasks. Grades assigned to each respondent's performance, using set rubrics, enables assessment of the quality of achievements represented and comparisons across groups.

The sources of questions addressed by performance tests are analyses of selected life-skill tasks and content specifications in curricular materials. The typical assessment questions concern whether individual students can effectively write, speak, figure, analyze, lead, work cooperatively, and solve given problems up to the level of acceptable standards. The main testing process is to define areas of skills to be assessed; select the type of assessment device; construct the assessment tasks; determine scoring rubrics; define standards for assessing performance; train and calibrate scorers; validate the measures; and administer, score, interpret, and report the results.

In speaking of licensing tests, Flexner (1910) called for tests that ascertain students' practical ability to successfully confront and solve problems in concrete cases. Some of the pioneers in applying performance assessment to state education systems were the state education departments in Vermont and Kentucky (Kentucky Department of Education, 1993; Koretz, 1986, 1996; Koretz & Barron, 1998). Other sources of information about the general approach and issues in performance testing include Baker, O'Neil, and Linn (1993); Herman, Gearhart, and Baker (1993); Linn, Baker, and Dunbar (1991); Mehrens (1972); Messick (1994); Stillman, Haley, Regan, Philbin, Smith, O'Donnell, and Pohl (1991); Swanson, Norman, and Linn (1995); Torrance (1993); and Wiggins (1989).

It is often difficult to establish the necessary conditions to employ the performance testing approach. It requires a huge outlay of time and resources for development and application. Typically, state education departments and school districts should probably use this approach very selectively and only when they

can make the investment needed to produce valid results that are worth the large, required investment. On the other hand, students' writing ability is best assessed and nurtured through obtaining, assessing, and providing critical feedback on their writing.

One advantage of performance tests is minimization of guessing. Requiring students to construct responses to assessment tasks also reinforces writing, computation, scientific experimentation, and other life skills.

Major disadvantages of the approach are the heavy time requirements for administration; the high costs of scoring; the difficulty in achieving reliable scores; the narrow scope of skills that can feasibly be assessed; and lack of norms for comparisons, especially at the national level. In general, performance tests are inefficient, costly, and often of dubious reliability. Moreover, compared with multiple-choice tests, performance tests, in the same amount of testing time, cover a much narrower range of questions.

Approach 8: Experimental Studies

In using controlled experiments, program evaluators randomly assign students or groups of students to experimental and control groups and then contrast the outcomes when the experimental group receives a particular intervention and the control group receives no special treatment or some different treatment. This type of study was quite prominent in program evaluation during the late 1960s and early 1970s, when there was a federal requirement to assess the effectiveness of federally funded innovations. However, experimental program evaluations subsequently fell into disfavor and disuse. Apparent reasons for this decline are that educators rarely can meet the required experimental conditions and assumptions.

This approach is labeled a questions-oriented or quasi-evaluation strategy because it starts with questions and methodology that may address only a narrow set of questions needed to assess a program's merit and worth. In the 1960s, Campbell and Stanley (1963) and others hailed the true experiment as the only sound means of evaluating interventions. This piece of evaluation history reminds one of Kaplan's (1964) famous warning against the so-called "law of the instrument," whereby a given method is equated to a field of inquiry. In such a case, the field of inquiry is restricted to the questions that are answerable by the given method. Fisher (1951) specifically warned against equating his experimental methods with science. In general, experimental design is a method that can contribute importantly to program evaluation, as Nave, Miech, and Mosteller (Chapter 8, this volume) have demonstrated, but by itself it is often insufficient to address a client's full range of evaluation questions.

The advance organizers in experimental studies are problem statements, competing treatments, hypotheses, investigatory questions, and randomized treatment and comparison groups. The usual purpose of the controlled experiment is to determine causal relationships between specified independent and dependent variables, such as between a given instructional method and student standardized-test performance. It is particularly noteworthy that the sources of questions investigated in the experi-

mental study are researchers, program developers, and policy figures, and not usually a program's constituents and practitioners.

The frequent question in the experimental study is: What are the effects of a given intervention on specified outcome variables? Typical methods used are experimental and quasi-experimental designs. Pioneers in using experimentation to evaluate programs are Campbell and Stanley (1963), Cronbach and Snow (1969), and Lindquist (1953). Others who have developed the methodology of experimentation substantially for program evaluation are Boruch (1994), Glass and Maguire (1968), Suchman (1967), and Wiley and Bock (1967).

Evaluators should consider conducting a controlled experiment only when its required conditions and assumptions can be met. Often this requires substantial political influence, substantial funding, and widespread agreement—among the involved funders, service providers, and beneficiaries—to submit to the requirements of the experiment. Such requirements typically include, among others, a stable program that will not have to be studied and modified during the evaluation; the ability to establish and sustain comparable program and control groups; the ability to keep the program and control conditions separate and uncontaminated; and the ability to obtain the needed criterion measures from all or at least a representative group of the members of the program and comparison groups. Evaluability assessment was developed as a particular methodology for determining the feasibility of moving ahead with an experiment (Smith, 1989; Wholey, 1995).

Controlled experiments have a number of advantages. They focus on results and not just intentions or judgments. They provide strong methods for establishing relatively unequivocal causal relationships between treatment and outcome variables, something that can be especially significant when program effects are small but important. Moreover, because of the prevalent use and success of experiments in such fields as medicine and agriculture, the approach has widespread credibility.

These advantages, however, are offset by serious objections to experimenting on school students and other human subjects. It is often considered unethical or even illegal to deprive control group members of the benefits of special funds for improving services. Likewise, many parents do not want schools or other organizations to experiment on their children by applying unproven interventions. Typically, schools find it impractical and unreasonable to randomly assign students to treatments and to hold treatments constant throughout the study period. Furthermore, experimental studies provide a much narrower range of information than organizations often need to assess and strengthen their programs. On this point, experimental studies tend to provide terminal information that is not useful for guiding the development and improvement of programs and may in fact thwart ongoing modifications of programs.

Approach 9: Management Information Systems

Management Information Systems are like politically controlled approaches, except that they supply managers with information needed to conduct and report on

their programs, as opposed to supplying them with information needed to win a political advantage. The management information approach is also like the decision/accountability-oriented approach, which will be discussed later, except that the decision/accountability-oriented approach provides information needed to both develop and defend a program's merit and worth, which goes beyond providing information that managers need to implement and report on their management responsibilities.

The advance organizers in most management information systems include program objectives, specified activities, and projected program milestones or events. A management information system's purpose is to continuously supply managers with the information they need to plan, direct, control, and report on their programs or spheres of responsibility.

The sources of questions addressed are the management personnel and their superiors. The main questions they typically want answered are: Are program activities being implemented according to schedule, according to budget, and with the expected results? To provide ready access to information for addressing such questions, systems regularly store and make accessible up-to-date information on program goals, planned operations, actual operations, staff, program organization, operations, expenditures, threats, problems, publicity, and achievements.

Methods employed in management information systems include system analysis, Program Evaluation and Review Technique (PERT), Critical Path Method, Program Planning and Budgeting System (PPBS), Management by Objectives (MBO), computer-based information systems, periodic staff progress reports, and regular budgetary reporting.

Cook (1966) introduced the use of PERT to education, and Kaufman (1969) wrote about the use of management information systems in education. Business schools and programs in computer information systems regularly provide courses in management information systems. These focus mainly on how to set up and employ computerized information banks for use in organizational decision making.

W. Edwards Deming (1986) argued that managers should pay close attention to process rather than being preoccupied with outcomes. He advanced a systematic approach for monitoring and continuously improving an enterprise's process, arguing that close attention to the process will result in increasingly better outcomes. It is commonly said that, in paying attention to this and related advice from Deming, Japanese car makers and later Americans greatly increased the quality of automobiles (Aguaro, 1990). Bayless and Massaro (1992) applied Deming's approach to program evaluations in education. Based on this writer's observations, the approach was not well suited to assessing the complexities of educational processes—possibly because, unlike the manufacture of automobiles, educators have no definitive, standardized models for linking exact educational processes to specified outcomes. Nevertheless, given modern database technology, program managers often can and should employ management information systems in multiyear projects and programs. Program databases can provide information, not only for keeping programs on track,

but also for assisting in the broader study and improvement of program processes and outcomes.

A major advantage of the use of management information systems is in giving managers information they can use to plan, monitor, control, and report on complex operations. A difficulty with the application of this industry-oriented type of system to education and other social services, however, is that the products of many programs are not amenable to a narrow, precise definition as is the case with a corporation's profit and loss statement. Moreover, processes in educational and social programs often are complex and evolving rather than straightforward and standardized like those of manufacturing and business. The information gathered in management information systems typically lacks the scope of context, input, process, and outcome information required to assess a program's merit and worth.

Approach 10: Benefit-Cost Analysis Approach

Benefit-cost analysis as applied to program evaluation is a set of largely quantitative procedures used to understand the full costs of a program and to determine and judge what investments returned in objectives achieved and broader social benefits. The aim is to determine costs associated with program inputs, determine the monetary value of the program outcomes, compute benefit-cost ratios, compare the computed ratios to those of similar programs, and ultimately judge a program's productivity in economic terms.

The benefit-cost analysis approach to program evaluation may be broken down into three levels of procedure: (1) cost analysis of program inputs, (2) cost-effectiveness analysis, and (3) benefit-cost analysis. These may be looked at as a hierarchy. The first type, cost analysis of program inputs, may be done by itself. Such analyses entail an ongoing accumulation of a program's financial history, which is useful in controlling program delivery and expenditures. The program's financial history can be used to compare its actual and projected costs and how costs relate to the costs of similar programs. Cost analyses can also be extremely valuable to outsiders who might be interested in replicating a program.

Cost-effectiveness analysis necessarily includes cost analysis of program inputs to determine the cost associated with progress toward achieving each objective. For example, two or more programs' costs and successes in achieving the same objectives might be compared. A program could be judged superior on cost-effectiveness grounds if it had the same costs but superior outcomes as similar programs. Or a program could be judged superior on cost-effectiveness grounds if it achieved the same objectives as more expensive programs. Cost-effectiveness analyses do not require conversion of outcomes to monetary terms but must be keyed to clear, measurable program objectives.

Benefit-cost analyses typically build on a cost analysis of program inputs and a cost-effectiveness analysis. But the benefit-cost analysis goes further. It seeks to identify a broader range of outcomes than just those associated with program objectives. It examines the relationship between the investment in a program and the extent of positive and negative impacts on the program's environment. In doing so, it ascer-

tains and places a monetary value on program inputs and each identified outcome. It identifies a program's benefit-cost ratios and compares these to similar ratios for competing programs. Ultimately, benefit-cost studies seek conclusions about the comparative benefits and costs of the examined programs.

Advance organizers for the overall benefit-cost approach are associated with cost breakdowns for both program inputs and outputs. Program input costs may be delineated by line items (e.g., personnel, travel, materials, equipment, communications, facilities, contracted services, overhead), by program components, and by year. In cost-effectiveness analysis, a program's costs are examined in relation to each program objective, and these must be clearly defined and assessed. The more ambitious benefit-cost analyses look at costs associated with main effects and side effects, tangible and intangible outcomes, positive and negative outcomes, and short-term and long-term outcomes—both inside and outside a program. Frequently, they also may break down costs by individuals and groups of beneficiaries. One may also estimate the costs of foregone opportunities and, sometimes, political costs. Even then, the real value of benefits associated with human creativity or self-actualization are nearly impossible to estimate. Consequently, the benefit-cost equation rests on dubious assumptions and uncertain realities.

The purposes of these three levels of benefit-cost analysis are to gain clear knowledge of what resources were invested, how they were invested, and with what effect. In popular vernacular, cost-effectiveness and benefit-cost analyses seek to determine the program's "bang for the buck." There is great interest in answering this type of question. Policy boards, program planners, and taxpayers are especially interested to know whether program investments are paying off in positive results that exceed or are at least as good as those produced by similar programs. Authoritative information on the benefit-cost approach may be obtained by studying the writings of Kee (1995), Levin (1983), and Tsang (Chapter 9, this volume).

Benefit-cost analysis is potentially important in most program evaluations. Evaluators and their clients are advised to discuss this matter thoroughly with their clients, to reach appropriate advance agreements on what should and can be done to obtain the needed cost information, and to do as much cost-effectiveness and benefit-cost analysis as can be done well and within reasonable costs.

Benefit-cost analysis is an important but problematic consideration in program evaluations. Most evaluations are amenable to analyzing the costs of program inputs and maintaining a financial history of expenditures. The main impediment is that program authorities often do not want anyone other than the appropriate accountants and auditors looking into their financial books. If cost analysis, even at only the input levels, is to be done, this must be clearly provided for in the initial contractual agreements covering the evaluation work. Performing cost-effectiveness analysis can be feasible if cost analysis of inputs is agreed to; if there are clear, measurable program objectives; and if comparable cost information can be obtained from competing programs. Unfortunately, it is usually hard to meet all these conditions. Even more unfortunate is the fact that it is usually impractical to conduct a thorough benefit-cost analysis. Not only must it meet all the conditions of the analysis

of program inputs and cost-effectiveness analysis, it must also place monetary values on identified outcomes, both anticipated and not expected.

Approach 11: Clarification Hearing

The clarification hearing is one label for the judicial approach to program evaluation. The approach essentially puts a program on trial. Role-playing evaluators competitively implement both a damning prosecution of the program—arguing that it failed—and a defense of the program—arguing that it succeeded. A judge hears arguments within the framework of a jury trial and controls the proceedings according to advance agreements on rules of evidence and trial procedures. The actual proceedings are preceded by the collection of and sharing of evidence by both sides. The prosecuting and defending evaluators may call witnesses and place documents and other exhibits into evidence. A jury hears the proceedings and ultimately makes and issues a ruling on the program's success or failure. Ideally, the jury is composed of persons representative of the program's stakeholders. By videotaping the proceedings, the administering evaluator can, after the trial, compile a condensed videotape as well as printed reports to disseminate what was learned through the process.

The advance organizers for a clarification hearing are criteria of program effectiveness that both the prosecuting and defending sides agree to apply. The main purpose of the judicial approach is to ensure that the evaluation's audience will receive balanced evidence on a program's strengths and weaknesses. The key questions essentially are: Should the program be judged a success or failure? Is it as good or better than alternative programs that address the same objectives?

Robert Wolf (1975) pioneered the judicial approach to program evaluation. Others who applied, tested, and further developed the approach include Levine (1974), Owens (1973), and Popham and Carlson (1983).

Based on the past uses of this approach, it can be judged as only marginally relevant to program evaluation. Because of its adversarial nature, the approach encourages evaluators to present biased arguments in order to win their cases. Thus, truth seeking is subordinated to winning. The most effective debaters are likely to convince the jury of their position even when it is poorly founded. The approach is also politically problematic, since it generates considerable acrimony. Despite the attractiveness of using the law, with its attendant rules of evidence, as a metaphor for program evaluation, its promise has not been fulfilled. There are few occasions in which it makes practical sense for evaluators to apply this approach.

Approach 12: Case Study Evaluations

Program evaluation that is based on a case study is a focused, in-depth description, analysis, and synthesis of a particular program or other object. The investigators do not control the program in any way. Rather, they look at it as it is occurring or as it occurred in the past. The study looks at the program in its geographic, cultural, organizational, and historical contexts, closely examining its internal operations and how it uses inputs and processes to produce outcomes. It examines a wide range of

intended and unexpected outcomes. It looks at the program's multiple levels and also holistically at the overall program. It characterizes both central dominant themes and variations and aberrations. It defines and describes the program's intended and actual beneficiaries. It examines beneficiaries' needs and the extent to which the program effectively addressed the needs. It employs multiple methods to obtain and integrate multiple sources of information. While it breaks apart and analyzes a program along various dimensions, it also provides an overall characterization of the program.

The main thrust of the case study approach is to delineate and illuminate a program, not necessarily to guide its development or to assess and judge its merit and worth. Hence, this chapter characterizes the case study approach as a questions/methods-oriented approach rather than an improvement/accountability approach.

Advance organizers in case studies include the definition of the program, characterization of its geographic and organizational environment, the historical period in which it is to be examined, the program's beneficiaries and their assessed needs, the program's underlying logic of operation and productivity, and the key roles involved in the program. A case study program evaluation's main purpose is to provide stakeholders and their audiences with an authoritative, in-depth, well-documented explication of the program.

The case study should be keyed to the questions of most interest to the evaluation's main audiences. The evaluator must therefore identify and interact with the program's stakeholders. Along the way, stakeholders will be engaged to help plan the study and interpret findings. Ideally, the audiences include the program's oversight body, administrators, staff, financial sponsors, beneficiaries, and potential adopters of the program.

Typical questions posed by some or all of the above audiences are: What is the program in concept and practice? How has it evolved over time? How does it actually operate to produce outcomes? What has it produced? What are the shortfalls and negative side effects? What are the positive side effects? In what ways and to what degrees do various stakeholders value the program? To what extent did the program effectively meet beneficiaries' needs? What were the most important reasons for the program's successes and failures? What are the program's most important unresolved issues? How much has it cost? What are the costs per beneficiary, per year, etc.? What parts of the program have been successfully transported to other sites? How does this program compare with what might be called critical competitors? These questions only illustrate the range of questions that a case study might address, since each study will be tempered by the interests of the client and other audiences for the study and the evaluator's interests.

To conduct effective case studies, evaluators need to employ a wide range of qualitative and quantitative methods. These may include analysis of archives; collection of artifacts, such as work samples; content analysis of program documents; both independent and participant observations; interviews; logical analysis of operations; focus groups; tests; questionnaires; rating scales; hearings; forums; and maintenance of a

program database. Reports may incorporate in-depth descriptions and accounts of key historical trends; focus on critical incidents, photographs, maps, testimony, relevant news clippings, logic models, and cross-break tables; and summarize main conclusions. The case study report may include a description of key dimensions of the case, as determined with the audience, as well as an overall holistic presentation and assessment. Case study reports may involve audio and visual media as well as printed documents.

Case study methods have existed for many years and have been applied in such areas as anthropology, clinical psychology, law, the medical profession, and social work. Pioneers in applying the method to program evaluation include Campbell (1975), Lincoln and Guba (1985), Platt (1992), Smith and Pohland (1974), Stake (1995), and Yin (1992).

The case study approach is highly appropriate in program evaluation. It requires no controls of treatments and subjects and looks at programs as they naturally occur and evolve. It addresses accuracy issues by employing and triangulating multiple perspectives, methods, and information sources. It employs all relevant methods and information sources. It looks at programs within relevant contexts and describes contextual influences on the program. It looks at programs holistically and in depth. It examines the programs internal workings and how it produces outcomes. It includes clear procedures for analyzing qualitative information. It can be tailored to focus on the audience's most important questions. It can be done retrospectively or in real time. It can be reported to meet given deadlines and subsequently updated based on further developments.

The main limitation of the case study is that some evaluators may mistake its openness and lack of controls as an excuse for approaching it haphazardly and bypassing steps to ensure that findings and interpretations possess rigor as well as relevance. Furthermore, because of a preoccupation with descriptive information, the case study evaluator may not collect sufficient judgmental information to permit a broad-based assessment of a programs merit and worth. Users of the approach might slight quantitative analysis in favor of qualitative analysis. By trying to produce a comprehensive description of a program, the case study evaluator may not produce timely feedback needed to help in program development. To overcome these potential pitfalls, evaluators using the case study approach should fully address the principles of sound evaluation as related to accuracy, utility, feasibility, and propriety.

Approach 13: Criticism and Connoisseurship

The connoisseur-based approach grew out of methods used in art criticism and literary criticism. It assumes that certain experts in a given substantive area are capable of in-depth analysis and evaluation that could not be done in other ways. Just as a national survey of wine drinkers could produce information concerning their overall preferences for types of wines and particular vineyards, it would not provide the detailed, creditable judgments of the qualities of particular wines that might be derived from a single connoisseur who has devoted a professional lifetime

to the study and grading of wines and whose judgments are highly and widely respected.

The advance organizer for the connoisseur-based study is the evaluator's special expertise and sensitivities. The study's purpose is to describe, critically appraise, and illuminate a particular program's merits. The evaluation questions addressed by the connoisseur-based evaluation are determined by expert evaluators—the critics and authorities who have undertaken the evaluation. Among the major questions they can be expected to ask are: What are the program's essence and salient characteristics? What merits and demerits distinguish the particular program from others of the same general kind?

The methodology of connoisseurship includes critics' systematic use of their perceptual sensitivities, past experiences, refined insights, and abilities to communicate their assessments. The evaluator's judgments are conveyed in vivid terms to help the audience appreciate and understand all of the program's nuances. Eisner (1975, 1983) has pioneered this strategy in education.⁶ A dozen or more of Eisner's students have conducted research and development on the connoisseurship approach, e.g., Vallance (1973) and Flinders and Eisner (Chapter 12, this volume). This approach obviously depends on the chosen expert's qualifications. It also requires an audience that has confidence in, and is willing to accept and use, the connoisseur's report. I would willingly accept and use any evaluation that Dr. Elliott Eisner agreed to present, but there are not many Eisners out there.

The main advantage of the connoisseur-based study is that it exploits the particular expertise and finely developed insights of persons who have devoted much time and effort to the study of a precise area. Such individuals can provide an array of detailed information that an audience can then use to form a more insightful analysis than otherwise might be possible. The approach's disadvantage is that it is dependent on the expertise and qualifications of the particular expert doing the program evaluation, leaving room for much subjectivity.

Approach 14: Program Theory-Based Evaluation

Program evaluations based on program theory begin with either (1) a well-developed and validated theory of how programs of a certain type within similar settings operate to produce outcomes or (2) an initial stage to approximate such a theory within the context of a particular program evaluation. The former condition is much more reflective of the implicit promises in a theory-based program evaluation, since the existence of a sound theory means that a substantial body of theoretical development has produced and tested a coherent set of conceptual, hypothetical, and pragmatic principles, as well as associated instruments to guide inquiry. The theory can then aid a program evaluator to decide what questions, indicators, and assumed linkages between and among program elements should be used to evaluate a program covered by the theory.

Some theories have been used more or less successfully to evaluate programs, which gives this approach some measure of viability. For example, health education/behavior change programs are sometimes founded on theoretical frameworks,

such as the Health Belief Model (Becker, 1974; Janz & Becker, 1984; Mullen, Hersey, & Iverson, 1987). Other examples are the PRECEDE-PROCEED Model for health promotion planning and evaluation (Green & Kreuter, 1991), Bandura's (1977) Social Cognitive Theory, the Stages of Change Theory of Prochaska and DiClemente (1992), and Peters and Waterman's (1982) theory of successful organizations. When such frameworks exist, their use probably can enhance a program's effectiveness and provide a credible structure for evaluating its functioning. Unfortunately, few program areas are buttressed by well-articulated and tested theories.

Thus, most theory-based evaluations begin by setting out to develop a theory that appropriately could be used to guide the particular program evaluation. As will be discussed later in this characterization, such ad hoc theory development efforts and their linkage to program evaluations are problematic. In any case, let us look at what the theory-based evaluator attempts to achieve.

The point of the theory development or selection effort is to identify advance organizers to guide the evaluation. Essentially, these are the mechanisms by which program activities are understood to produce or contribute to program outcomes, along with the appropriate description of context, specification of independent and dependent variables, and portrayal of key linkages. The main purposes of the theory-based program evaluation are to determine the extent to which the program of interest is theoretically sound, to understand why it is succeeding or failing, and to provide direction for program improvement.

Questions for the program evaluation are derived from the guiding theory. Example questions include: Is the program grounded in an appropriate, well-articulated, and validated theory? Is the employed theory reflective of recent research? Are the program's targeted beneficiaries, design, operation, and intended outcomes consistent with the guiding theory? How well does the program address and serve the full range of pertinent needs of targeted beneficiaries? If the program is consistent with the guiding theory, are the expected results being achieved? Are program inputs and operations producing outcomes in the ways the theory predicts? What changes in the program's design or implementation might produce better outcomes? What elements of the program are essential for successful replication? Overall, was the program theoretically sound, did it operate in accordance with an appropriate theory, did it produce the expected outcomes, were the hypothesized causal linkages confirmed, what program modifications are needed, is the program worthy of continuation and/or dissemination, and what program features are essential for successful replication?

The nature of these questions suggests that the success of the theory-based approach is dependent on a foundation of sound theory development and validation. This, of course, entails sound conceptualization of at least a context-dependent theory, formulation and rigorous testing of hypotheses derived from the theory, development of guidelines for practical implementation of the theory based on extensive field trials, and independent assessment of the theory. Unfortunately, not many program areas in education and the social sciences are grounded in sound theories. Moreover, evaluators wanting to employ a theory-based evaluation do not

often find it feasible to conduct the full range of theory development and validation steps, and still get the evaluation done on time. Thus, in claiming to conduct a theory-based evaluation, evaluators often seem to promise much more than they can deliver.

The main procedure typically used in “theory-based program evaluations” is a model of the program’s logic. This may be a detailed flowchart of how inputs are thought to be processed to produce intended outcomes. It may also be a grounded theory, such as those advocated by Glaser and Strauss (1967). The network analysis of the former approach is typically an armchair theorizing process involving evaluators and persons who are supposed to know how the program is expected to operate and produce results. They discuss, scheme, discuss some more, network, discuss further, and finally produce networks in varying degrees of detail of what is involved in making the program work and how the various elements are linked to produce desired outcomes. The more demanding grounded theory requires a systematic, empirical process of observing events or analyzing materials drawn from operating programs, followed by an extensive modeling process.

Pioneers in applying theory development procedures to program evaluation include Glaser and Strauss (1967) and Weiss (1972, 1995). Other developers of the approach are Bickman (1990), Chen (1990), and Rogers (Chapter 13, this volume).

In any program evaluation assignment, it is reasonable for the evaluator to examine the extent to which program plans and operations are grounded in an appropriate theory or model. It can also be useful to engage in a modicum of effort to network the program and thereby seek out key variables and linkages. As noted previously, in the enviable but rare situation where a relevant, validated theory exists, the evaluator can beneficially apply it in structuring the evaluation and in analyzing findings.

However, if a relevant, defensible theory of the program’s logic does not exist, evaluators need not develop one. In fact, if they attempt to do so, they will incur many threats to their evaluation’s success. Rather than evaluating a program and its underlying logic, evaluators might usurp the program staff’s responsibility for program design. They might do a poor job of theory development, given limitations on time and resources to develop and test an appropriate theory. They might incur the conflict of interest associated with having to evaluate the theory they developed. They might pass off an unvalidated model of the program as a theory, when it meets almost none of the requirements of a sound theory. They might bog down the evaluation in too much effort to develop a theory. They might also focus attention on a theory developed early in a program and later discover that the program has evolved to be a quite different enterprise than what was theorized at the outset. In this case, the initial theory could become a “Procrustean bed” for the program evaluation.

Overall, there really is not much to recommend theory-based program evaluation, since doing it right is usually not feasible and since failed or misrepresented attempts can be highly counterproductive. Nevertheless, modest attempts to model programs—labeled as such—can be useful for identifying measurement variables, so

long as the evaluator does not spend too much time on this and so long as the model is not considered as fixed or as a validated theory. In the rare case where an appropriate theory already exists, the evaluator can make beneficial use of it to help structure and guide the evaluation and interpret the findings.

Approach 15: Mixed-Methods Studies

In an attempt to resolve the longstanding debate about whether program evaluations should employ quantitative or qualitative methods, some authors have proposed that evaluators should regularly combine these methods in given program evaluations (for example, see the National Science Foundation's 1997 *User-Friendly Handbook for Mixed Method Evaluations*). Such recommendations, along with practical guidelines and illustrations, are no doubt useful to many program staff members and to evaluators. But in the main, the recommendation for a mixed-method approach only highlights a large body of longstanding practice of mixed-methods program evaluation rather than proposing a new approach. All seven approaches discussed in the remainder of this section of the chapter employ both qualitative and quantitative methods. What sets them apart from the mixed-method approach is that their first considerations are not the methods to be employed but either the assessment of value or the social mission to be served. The mixed-methods approach is included in this section on questions/methods approaches, because it is preoccupied with using multiple methods rather than whatever methods are needed to comprehensively assess a program's merit and worth. As with the other approaches in this section, the mixed-methods approach may or may not fully assess a program's value; thus, it is classified as a quasi-evaluation approach.

The advance organizers of the mixed-methods approach are formative and summative evaluations, qualitative and quantitative methods, and intra-case or cross-case analysis. Formative evaluations are employed to examine a program's development and assist in improving its structure and implementation. Summative evaluations basically look at whether objectives were achieved, but may look for a broader array of outcomes. Qualitative and quantitative methods are employed in combination to assure depth, scope, and dependability of findings. This approach also applies to carefully selected single programs or to comparisons of alternative programs.

The basic purposes of the mixed method approach are to provide direction for improving programs as they evolve and to assess their effectiveness after they have had time to produce results. Use of both quantitative and qualitative methods is intended to ensure dependable feedback on a wide range of questions; depth of understanding of particular programs; a holistic perspective; and enhancement of the validity, reliability, and usefulness of the full set of findings. Investigators look to quantitative methods for standardized, replicable findings on large data sets. They look to qualitative methods for elucidation of the program's cultural context, dynamics, meaningful patterns and themes, deviant cases, and diverse impacts on individuals as well as groups. Qualitative reporting methods are applied to bring the findings to life, to make them clear, persuasive, and interesting. By using

both quantitative and qualitative methods, the evaluator secures cross-checks on different subsets of findings and thereby instills greater stakeholder confidence in the overall findings.

The sources of evaluation questions are the program's goals, plans, and stakeholders. The stakeholders often include skeptical as well as supportive audiences. Among the important stakeholders are program administrators and staff, policy boards, financial sponsors, beneficiaries, taxpayers, and program area experts.

The approach may pursue a wide range of questions. Examples of formative evaluation questions are: To what extent do program activities follow the program plan, time line, and budget? To what extent is the program achieving its goals? What problems in design or implementation need to be addressed? Examples of summative evaluation questions are: To what extent did the program achieve its goals? Was the program appropriately effective for all beneficiaries? What interesting stories emerged? What are program stakeholders' judgments of program operations, processes, and outcomes? What were the important side effects? Is the program sustainable and transportable?

The approach employs a wide range of methods. Among quantitative methods are surveys using representative samples, both cohort and cross-sectional samples, norm-referenced tests, rating scales, quasi-experiments, significance tests for main effects, and a posteriori statistical tests. The qualitative methods may include ethnography, document analysis, narrative analysis, purposive samples, single cases, participant observers, independent observers, key informants, advisory committees, structured and unstructured interviews, focus groups, case studies, study of outliers, diaries, logic models, grounded theory development, flow charts, decision trees, matrices, and performance assessments. Reports may include abstracts, executive summaries, full reports, oral briefings, conference presentations, and workshops. They should include a balance of narrative and numerical information.

Considering his book on service studies in higher education, Ralph Tyler (Tyler et al., 1932) was certainly a pioneer in the mixed-method approach to program evaluation. Other authors who have written cogently on the approach are Guba and Lincoln (1981), Kidder and Fine (1987), Lincoln and Guba (1985), Miron (1998), Patton (1990), and Schatzman and Strauss (1973).

It is almost always appropriate to consider using a mixed-methods approach. Certainly, the evaluator should take advantage of opportunities to obtain any and all potentially available information that is relevant to assessing a program's merit and worth. Sometimes a study can be mainly or only qualitative or quantitative, but usually such studies would be strengthened by including both types of information. The key point is to choose methods because they can effectively address the study's questions, not because they are either qualitative or quantitative.

Key advantages of using both qualitative and quantitative methods are that they complement each other in ways that are important to the evaluation's audiences. Information from quantitative methods tends to be standardized, efficient, amenable to standard tests of reliability, easily summarized and analyzed, and accepted as "hard"

data. Information from qualitative approaches adds depth; can be delivered in interesting, story-like presentations; and provides a means to explore and understand the more superficial quantitative findings. Using both types of method affords important cross-checks on findings.

The main pitfall in pursuing the mixed-methods approach is using multiple methods because this is the popular thing to do rather than because the selected methods best respond to the evaluation questions. Moreover, sometimes evaluators let the combination of methods compensate for a lack of rigor in applying them. Using a mixed methods approach can produce confusing findings if an investigator uncritically mixes positivistic and postmodern paradigms, since quantitative and qualitative methods are derived from different theoretical approaches to inquiry and reflect different conceptions of knowledge. Many evaluators do not possess the requisite foundational knowledge in both the sciences and humanities to effectively combine quantitative and qualitative methods. The approaches in the remainder of this chapter place proper emphasis on mixed methods, making choice of the methods subservient to the approach's dominant philosophy and to the particular evaluation questions to be addressed.

The mixed methods approach to evaluation concludes this chapter's discussion of the questions/methods approaches to evaluation. These 13 approaches tend to concentrate on selected questions and methods and thus may or may not fully address an evaluation's fundamental requirement to assess a program's merit and worth. The array of these approaches suggests that the field has advanced considerably since the 1950s when program evaluations were rare and mainly used approaches grounded in behavioral objectives, standardized tests, and/or accreditation visits.

IMPROVEMENT/ACCOUNTABILITY-ORIENTED EVALUATION APPROACHES

I now turn to three approaches that stress the need to fully assess a program's merit and worth. These approaches are expansive and seek comprehensiveness in considering the full range of questions and criteria needed to assess a program's value. Often they employ the assessed needs of a program's stakeholders as the foundational criteria for assessing the program's merit and worth. They also seek to examine the full range of pertinent technical and economic criteria for judging program plans and operations. They look for all relevant outcomes, not just those keyed to program objectives. Usually, they are objectivist and assume an underlying reality in seeking definitive, unequivocal answers to the evaluation questions. Typically, they must use multiple qualitative and quantitative assessment methods to provide cross-checks on findings. In general, the approaches conform closely to this chapter's definition of evaluation. The approaches are labeled Decisions/Accountability, Consumer-Oriented, and Accreditation. The three approaches emphasize respectively improvement through serving program decisions, providing consumers with assessments of optional programs and services, and helping consumers to gain assurances that given programs are professionally sound and effective.

Approach 16: Decision/Accountability-Oriented Studies

The decision/accountability-oriented approach emphasizes that program evaluation should be used proactively to help improve a program as well as retroactively to judge its merit and worth. The approach is distinguished from management information systems and from politically controlled studies because decision/accountability-oriented studies emphasize questions of merit and worth. The approach's philosophical underpinnings include an objectivist orientation to finding best answers to context-limited questions and subscription to the principles of a well-functioning democratic society, especially human rights, equity, excellence, conservation, and accountability. Practically, the approach engages stakeholders in focusing the evaluation, addressing their most important questions, providing timely, relevant information to assist decision making, and producing an accountability record.

The advance organizers for the approach include decision makers/stakeholders, decision situations, and program accountability requirements. Audiences include not just top managers but stakeholders at all organizational program levels. From the bottom up, such stakeholders may include beneficiaries, parents/guardians, service providers, administrators, program consultants, support personnel, policy makers, funding authorities, and taxpayers. The generic decision situations to be served may include defining goals and priorities, choosing from competing services, planning programs, budgeting, staffing, using services, guiding participation, judging progress, and recycling program operations. Key classes of needed evaluative information are assessments of needs, problems, and opportunities; identification and assessment of competing programs or program approaches; assessment of program plans; assessment of staff qualifications and performance; assessment of program facilities and materials; monitoring and assessment of process; assessment of intended and unintended and short-range and long-range outcomes; and assessment of cost-effectiveness.

The basic purpose of decision/accountability studies is to provide a knowledge and value base for making and being accountable for decisions that result in developing, delivering, and making informed use of cost-effective services. Thus, evaluators must interact with representative members of their audiences, discern their questions, and supply them with relevant, timely, efficient, and accurate information. The approach stresses that an evaluation's most important purpose is not to prove but to improve.

The sources of questions addressed by this approach are the concerned and involved stakeholders. These may include all persons and groups who must make choices related to initiating, planning, implementing, and using a program's services. Main questions addressed are: Has an appropriate beneficiary population been determined? What beneficiary needs should be addressed? What are the available alternative ways to address these needs, and what are their comparative merits and costs? Are plans of services and participation sound? Is there adequate provision for facilities, materials, and equipment? Is the program staff sufficiently qualified and cred-

ible? Have appropriate roles been assigned to the different participants? Are the participants effectively carrying out their assignments? Is the program working and should it be revised in any way? Is the program effectively reaching all the targeted beneficiaries? Is the program meeting participants' needs? Did beneficiaries play their part? Is the program better than competing alternatives? Is it affordable? Is it sustainable? Is it transportable? Is the program worth the required initial investment? Answers to these and related questions are to be based on the underlying standard of good programs, i.e., they must effectively reach and serve beneficiaries' targeted needs at a reasonable cost, and do so as well as or better than reasonably available alternatives.

Many methods may be used in decision/accountability-oriented program evaluations. These include, among others, surveys, needs assessments, case studies, advocate teams, observations, interviews, resident evaluators, and quasi-experimental and experimental designs. To make the approach work, the evaluator must regularly interact with a representative body of stakeholders. Typically, the evaluator should establish and engage a representative stakeholder advisory panel to help define evaluation questions, shape evaluation plans, review draft reports, and help disseminate findings. The evaluator's exchanges with this group involve conveyance of evaluation feedback that may be of use in program improvement and use, as well as determining what future evaluation reports would be most helpful to program personnel and other stakeholders. Interim reports may assist beneficiaries, program staff, and others to obtain feedback on the program's merits and worth and on the quality of their own participation. By maintaining a dynamic baseline of evaluation information and applications of the information, the evaluator can use this information to develop a comprehensive summative evaluation report, periodically update the broad group of stakeholders, and supply program personnel with findings for their own accountability reports.

The involvement of stakeholders is consistent with a key principle of the change process. An enterprise—read *evaluation* here—can best help bring about change in a target group's behavior if that group was involved in planning, monitoring, and judging the enterprise. By involving stakeholders throughout the evaluation process, decision-oriented evaluators lay the groundwork for bringing stakeholders to understand and value the evaluation process and apply the findings.

Cronbach (1963) advised educators to reorient their evaluations from an objectives orientation to a concern for making better program decisions. While he did not use the terms *formative* and *summative* evaluation, he essentially defined the underlying concepts. In discussing the distinctions between the constructive, proactive orientation on the one hand and the retrospective, judgmental orientation on the other, he argued for placing more emphasis on the former. He noted the limited functionality of the tradition of stressing retrospective outcomes evaluation. Later, I (Stufflebeam, 1966, 1967) argued that evaluations should help program personnel make and defend decisions keyed to meeting beneficiaries' needs. While I advocated an improvement orientation to evaluation, I also emphasized that evaluators must both inform decisions and provide information for accountability. I also emphasized

that the approach should interact with and serve the full range of stakeholders who need to make judgments and choices about a program. Others who have contributed to the development of a decision/accountability orientation to evaluation are Alkin (1969) and Webster (1975).

The decision/accountability-oriented approach is applicable in cases where program staff and other stakeholders want and need both formative and summative evaluation. It can provide the evaluation framework for both internal and external evaluation. When used for internal evaluation, it is usually important to commission an independent metaevaluation of the inside evaluator's work. Beyond program evaluations, this approach has proved useful in evaluating personnel, students, projects, facilities, and products.

A major advantage of the approach is that it encourages program personnel to use evaluation continuously and systematically to plan and implement programs that meet beneficiaries' targeted needs. It aids decision making at all program levels and stresses improvement. It also presents a rationale and framework of information for helping program personnel to be accountable for their program decisions and actions. It involves the full range of stakeholders in the evaluation process to ensure that their evaluation needs are well addressed and to encourage and support them to make effective use of evaluation findings. It is comprehensive in attending to context, inputs, process, and outcomes. It balances the use of quantitative and qualitative methods. It is keyed to professional standards for evaluations. Finally, the approach emphasizes that evaluations must be grounded in the democratic principles of a free society.

A major limitation is that the collaboration required between an evaluator and stakeholders introduces opportunities for impeding the evaluation and/or biasing its results, especially when the evaluative situation is politically charged. Further, when evaluators are actively influencing the course of a program, they may identify so closely with it that they lose some of the independent, detached perspective needed to provide objective, forthright reports. Moreover, the approach may overemphasize formative evaluation and give too little time and resources to summative evaluation. External metaevaluation has been employed to counteract opportunities for bias and to ensure a proper balance of the formative and summative aspects of evaluation. Though the charge is erroneous, this approach carries the connotation that only top decision makers are served.

Approach 17: Consumer-Oriented Studies

In the consumer-oriented approach, the evaluator is the "enlightened surrogate consumer." He or she must draw direct evaluative conclusions about the program being evaluated. Evaluation is viewed as the process of determining something's merit and worth, with evaluations being the products of that process. The approach regards a consumer's welfare as a program's primary justification and accords that welfare the same primacy in program evaluation. Grounded in a deeply reasoned view of ethics and the common good, together with skills in obtaining and synthesizing pertinent, valid, and reliable information, the evaluator should help developers produce and

deliver products and services that are of excellent quality and of great use to consumers (for example, students, their parents, teachers, and taxpayers). More importantly, the evaluator should help consumers identify and assess the merit and worth of competing programs, services, and products.

Advance organizers include societal values, consumers' needs, costs, and criteria of goodness in the particular evaluation domain. The purpose of a consumer-oriented program evaluation is to judge the relative merits and worth of the products and services of alternative programs and, thereby, to help taxpayers, practitioners, and potential beneficiaries make wise choices. The approach is objectivist in assuming an underlying reality and positing that it is possible, although often extremely difficult, to find best answers. It looks at a program comprehensively in terms of its quality and costs, functionally regarding the assessed needs of the intended beneficiaries, and comparatively considering reasonably available alternative programs. Evaluators are expected to subject their program evaluations to evaluations, what Scriven has termed *metaevaluation*.

The approach employs a wide range of assessment topics. These include program description, background and context, client, consumers, resources, function, delivery system, values, standards, process, outcomes, costs, critical competitors, generalizability, statistical significance, assessed needs, bottom-line assessment, practical significance, recommendations, reports, and metaevaluation. The evaluation process begins with consideration of a broad range of such topics, continuously compiles information on all of them, and ultimately culminates in a super-compressed judgment of the program's merit and worth.

Questions for the consumer-oriented study are derived from society, from program constituents, and especially from the evaluator's frame of reference. The general question addressed is: Which of several alternative programs is the best choice, given their differential costs, the needs of the consumer group, the values of society at large, and evidence of both positive and negative outcomes?

Methods include checklists, needs assessments, goal-free evaluation, experimental and quasi-experimental designs, modus operandi analysis, applying codes of ethical conduct, and cost analysis (Scriven, 1974a). A preferred method is for an external, independent consumer advocate to conduct and report findings of studies of publicly supported programs. The approach is keyed to employing a sound checklist of the program's key aspects. Scriven (1991) developed a generic "Key Evaluation Checklist" for this purpose. The main evaluative acts in this approach are scoring, grading, ranking, apportioning, and producing the final synthesis (Scriven, 1994a).

Scriven (1967) was a pioneer in applying the consumer-oriented approach to program evaluation, and his work parallels the concurrent work of Ralph Nader and the Consumers Union in the general field of consumerism. Glass (1969) has supported and developed Scriven's approach. Scriven coined the terms *formative* and *summative* evaluation. He allowed that evaluations can be divergent in early quests for critical competitors and explorations related to clarifying goals and making programs function well. However, he also maintained that ultimately evaluations must

converge on summative judgments about a program's merit and worth. While accepting the importance of formative evaluation, he also argued against Cronbach's (1963) position that formative evaluation should be given the major emphasis. According to Scriven, the fundamental aim of a sound evaluation is to judge a program's merit, comparative value, and overall worth. He sees evaluation as a trans-discipline encompassing all evaluations of various entities across all applied areas and disciplines and comprised of a common logic, methodology, and theory that transcends specific evaluation domains, which also have their unique characteristics (Scriven, 1991, 1994a).

The consumer-oriented study requires a highly credible and competent expert, together with either sufficient resources to allow the expert to conduct a thorough study or other means to obtain the needed information. Often, a consumer-oriented evaluator is engaged to evaluate a program after its formative stages are over. In these situations, the external consumer-oriented evaluator is often dependent on being able to access a substantial base of information that the program staff had accumulated. If no such base of information exists, the consumer-oriented evaluator will have great difficulty in obtaining enough information to produce a thorough, defensible summative program evaluation.

One of the main advantages of consumer-oriented evaluation is that it is a hard-hitting, independent assessment intended to protect consumers from shoddy programs, services, and products and to guide them to support and use those contributions that best and most cost-effectively address their needs. The approach's stress on independent/objective assessment and its attempt to achieve a comprehensive assessment of merit and worth yield high credibility with consumer groups. This is aided by Michael Scriven's (1991) Key Evaluation Checklist and his *Evaluation Thesaurus* (in which he presents and explains the checklist). The approach provides for a summative evaluation to yield a bottom-line judgment of merit and worth, preceded by a formative evaluation to assist developers to help ensure that their programs will succeed.

One disadvantage of the consumer-oriented evaluation is that it can be so independent from practitioners that it may not assist them to do a better job in serving consumers. If summative evaluation is applied too early, it can intimidate developers and stifle their creativity. However, if summative evaluation is applied only near a program's end, the evaluator may have great difficulty in obtaining sufficient evidence to confidently and credibly judge the program's basic value. This often iconoclastic approach is also heavily dependent on a highly competent, independent, and "bulletproof" evaluator.

Approach 18: Accreditation/Certification Approach

Many educational institutions, hospitals, and other service organizations have periodically been the subject of an accreditation study, and many professionals, at one time or another, have had to meet certification requirements for a given position. Such studies of institutions and personnel are in the realm of accountability-oriented evaluations, as well as having an improvement element. Institutions,

institutional programs, and personnel are studied to prove whether they meet requirements of given professions and service areas and whether they are fit to serve designated functions in society; typically, the feedback reports identify areas for improvement.

The advance organizers used in the accreditation/certification study usually are guidelines and criteria that some accrediting or certifying body has adopted. As previously suggested, the evaluation's purpose is to determine whether institutions, institutional programs, and/or personnel should be approved to perform specified functions.

The source of questions for accreditation or certification studies is the accrediting or certifying body. Basically, they address the question: Are institutions and their programs and personnel meeting minimum standards, and how can their performance be improved?

Typical methods used in the accreditation/certification approach are self-study and self-reporting by the individual or institution. In the case of institutions, panels of experts are assigned to visit the institution, verify a self-report, and gather additional information. The basis for the self-studies and the visits by expert panels are usually guidelines and criteria that have been specified by the accrediting agency.

Accreditation of education was pioneered by the College Entrance Examination Board around 1901. Since then, the accreditation function has been implemented and expanded, especially by the Cooperative Study of Secondary School Standards, dating from around 1933. Subsequently, the accreditation approach has been developed, further expanded, and administered by the North Central Association of Secondary Schools and Colleges, along with their associated regional accrediting agencies across the United States, and by many other accrediting and certifying bodies. Similar accreditation practices are found in medicine, law, architecture, and many other professions.

Any area of professional service that potentially could put the public at risk if services are not delivered by highly trained specialists in accordance with standards of good practice and safety should consider subjecting its programs to accreditation reviews and its personnel to certification processes. Such use of evaluation services is very much in the public interest and is also a means of getting feedback which can be of use in strengthening capabilities and practices.

The major advantage of the accreditation or certification study is that it aids lay persons in making informed judgments about the quality of organizations and programs and the qualifications of individual personnel. Major difficulties are that the guidelines of accrediting and certifying bodies often emphasize inputs and processes and not outcomes. Further, the self-study and visitation processes used in accreditation offer many opportunities for corruption and inept performance. As is the case for a number of the evaluation approaches described above, it is prudent to subject accreditation and certification processes themselves to independent metaevaluations.

The three improvement/accountability-oriented approaches emphasize the assessment of merit and worth, which is the thrust of the definition of evaluation used

to classify the 22 approaches considered in this chapter. The chapter turns next to the fourth and final set of program evaluation approaches—those concerned with using evaluation to further some social agenda.

SOCIAL AGENDA-DIRECTED/ADVOCACY APPROACHES

Social Agenda/Advocacy approaches are directed to making a difference in society through program evaluation. These approaches seek to ensure that all segments of society have equal access to educational and social opportunities and services. They have an affirmative action bent toward giving preferential treatment through program evaluation to the disadvantaged. If—as many persons have stated—information is power, then these approaches employ program evaluation to empower the disenfranchised.

The four approaches in this set are oriented to employing the perspectives of stakeholders as well as of experts in characterizing, investigating, and judging programs. They favor a constructivist orientation and the use of qualitative methods. For the most part, they eschew the possibility of finding right or best answers and reflect the philosophy of postmodernism, with its attendant stress on cultural pluralism, moral relativity, and multiple realities. They provide for democratic engagement of stakeholders in obtaining and interpreting findings.

There is a concern that these approaches might concentrate so heavily on serving a social mission that they fail to meet the standards of a sound evaluation. By giving stakeholders the authority for key evaluation decisions, related especially to interpretation and release of findings, evaluators empower these persons to use evaluation to their best advantage. Such delegation of authority over important evaluation matters makes the evaluation vulnerable to bias and other misuse. Further, if an evaluator is intent on serving the underprivileged, empowering the disenfranchised, and/or righting educational and/or social injustices, he or she might compromise the independent, impartial perspective needed to produce valid findings, especially if funds allocated to serve these groups would be withdrawn as a consequence of a negative report. In the extreme, an advocacy evaluation could compromise the integrity of the evaluation process to achieve social objectives and thus devolve into a pseudoevaluation.

Nevertheless, there is much to recommend these approaches, since they are strongly oriented to democratic principles of equity and fairness and employ practical procedures for involving the full range of stakeholders. The particular social agenda/advocacy-directed approaches presented in this chapter seem to have sufficient safeguards needed to walk the line between sound evaluation services and politically corrupted evaluations. Worries about bias control in these approaches increase the importance of subjecting advocacy evaluations to metaevaluations grounded in standards for sound evaluations.

Approach 19: Client-Centered Studies (or Responsive Evaluation)

The classic approach in this set is the client-centered study, or what Robert Stake (1983) has termed the responsive evaluation. The label *client-centered* evaluation is

used here, because one pervasive theme is that the evaluator must work with and for the support of a diverse client group including, for example, teachers, administrators, developers, taxpayers, legislators, and financial sponsors. They are the clients in the sense that they support, develop, administer, or directly operate the programs under study and seek or need evaluators' counsel and advice in understanding, judging, and improving programs. The approach charges evaluators to interact continuously with, and respond to, the evaluative needs of the various clients.

This approach contrasts sharply with Scriven's consumer-oriented approach. Stake's evaluators are not the independent, objective assessors of Scriven. The client-centered study embraces local autonomy and helps people who are involved in a program to evaluate it and use the evaluation for program improvement. The evaluator in a sense is the client's handmaiden as they strive to make the evaluation serve their needs. Moreover, the client-centered approach rejects objectivist evaluation, subscribing to the postmodernist view, wherein there are no best answers or clearly preferable values and subjective information is preferred. In this approach, the program evaluation may culminate in conflicting findings and conclusions, leaving interpretation to the eyes of the beholders. Client-centered evaluation is perhaps the leading entry in the "relativistic school of evaluation," which calls for a pluralistic, flexible, interactive, holistic, subjective, constructivist, and service-oriented approach. The approach is relativistic because it seeks no final authoritative conclusion, interpreting findings against stakeholders' different and often conflicting values. The approach seeks to examine a program's full countenance and prizes the collection and reporting of multiple, often conflicting perspectives on the value of a program's format, operations, and achievements. Side effects and incidental gains as well as intended outcomes are to be identified and examined.

The advance organizers in client-centered evaluations are stakeholders' concerns and issues in the program itself, as well as the program's rationale, background, transactions, outcomes, standards, and judgments. The client-centered program evaluation may serve many purposes. Some of these are helping people in a local setting gain a perspective on the program's full countenance; understanding the ways that various groups see the program's problems, strengths, and weaknesses; and learning the ways affected people value the program, as well as the ways program experts judge it. The evaluators process goal is to carry on a continuous search for key questions and to provide clients with useful information as it becomes available.

The client-centered/responsive approach has a strong philosophical base: evaluators should promote equity and fairness, help those with little power, thwart the misuse of power, expose the huckster, unnerve the assured, reassure the insecure, and always help people see things from alternative viewpoints. The approach subscribes to moral relativity and posits that, for any given set of findings, there are potentially multiple, conflicting interpretations that are equally plausible.

Community, practitioner, and beneficiary groups in the local environment, together with external program area experts, provide the questions addressed by the client-centered study. In general, the groups usually want to know what the program achieved, how it operated, and how it was judged by involved persons and experts

in the program area. The more specific evaluation questions emerge as the study unfolds based on the evaluator's continuing interactions with stakeholders and their collaborative assessment of the developing evaluative information.

This approach reflects a formalization of the longstanding practice of informal, intuitive evaluation. It requires a relaxed and continuous exchange between evaluator and clients. It is more divergent than convergent. Basically, the approach calls for continuing communication between evaluator and audience for the purposes of discovering, investigating, and addressing a program's issues. Designs for client-centered program evaluations are relatively open-ended and emergent, building to narrative description, rather than aggregating measurements across cases. The evaluator attempts to issue timely responses to clients' concerns and questions by collecting and reporting useful information, even if the needed information was not anticipated at the study's beginning. Concomitant with the ongoing conversation with clients, the evaluator attempts to obtain and present a rich set of information on the program. This includes its philosophical foundation and purposes, history, transactions, and outcomes. Special attention is given to side effects, the standards that various persons hold for the program, and their judgments of the program.

Depending on the evaluation's purpose, the evaluator may legitimately employ a range of different methods. Preferred methods are the case study, expressive objectives, purposive sampling, observation, adversary reports, story telling to convey complexity, sociodrama, and narrative reports. Client-centered evaluators are charged to check for the existence of stable and consistent findings by employing redundancy in their data-collecting activities and replicating their case studies. They are not expected to act as a program's sole or final judges, but should collect, process, and report the opinions and judgments of the full range of the program's stakeholders as well as those of pertinent experts. In the end, the evaluator makes a comprehensive statement of what the program is observed to be and references the satisfaction and dissatisfaction that appropriately selected people feel toward the program. Overall, the client-centered/responsive evaluator uses whatever information sources and techniques seem relevant to portraying the program's complexities and multiple realities, and communicates the complexity even if the result instills doubt and makes decisions more difficult.

Stake (1967) is the pioneer of the client-centered/responsive type of study, and his approach has been supported and developed by Denny (1978), MacDonald (1975), Parlett and Hamilton (1972), Rippey (1973), and Smith and Pohland (1974). Guba's (1978) early development of constructivist evaluation was heavily influenced by Stake's writings on responsive evaluation. Stake has expressed skepticism about scientific inquiry as a dependable guide to developing generalizations about human services, and pessimism about the potential benefits of formal program evaluations.

The main condition for applying the client-centered approach is a receptive client group and a confident, competent, responsive evaluator. The client must be willing to endorse a quite open, flexible evaluation plan as opposed to a well-developed,

detailed, preordinate plan and must be receptive to equitable participation by a representative group of stakeholders. The client must find qualitative methods acceptable and usually be willing to forego anything like a tightly controlled experimental study, although in exceptional cases a controlled field experiment might be employed. Clients and other involved stakeholders need tolerance, even appreciation for ambiguity, and should hold out only modest hopes of obtaining definitive answers to evaluation questions. Clients must also be receptive to ambiguous findings, multiple interpretations, the employment of competing value perspectives, and the heavy involvement of stakeholders in interpreting and using findings. Finally, clients must be sufficiently patient to allow the program evaluation to unfold and find its direction based on ongoing interactions between the evaluator and the stakeholders.

A major strength of the responsive/client-centered approach is that it involves action-research, in which people funding, implementing, and using programs are helped to conduct their own evaluations and use the findings to improve their understanding, decisions, and actions. The evaluations look deeply into the stakeholders' main interests and search broadly for relevant information. They also examine the program's rationale, background, process, and outcomes. They make effective use of qualitative methods and triangulate findings from different sources. The approach stresses the importance of searching widely for unintended as well as intended outcomes. It also gives credence to the meaningful participation in the evaluation by the full range of interested stakeholders. Judgments and other inputs from all such persons are respected and incorporated in the evaluations. The approach also provides for effective communication of findings.

A major weakness is the approach's vulnerability regarding external credibility, since people in the local setting, in effect, have considerable control over the evaluation of their work. Similarly, evaluators working so closely with stakeholders may lose their independent perspectives. The approach is not very amenable to reporting clear findings in time to meet decision or accountability deadlines. Moreover, rather than bringing closure, the approach's adversarial aspects and divergent qualities may generate confusion and contentious relations among stakeholders. Sometimes, this cascading, evolving approach may bog down in an unproductive quest for multiple inputs and interpretations.

Approach 20: Constructivist Evaluation

The constructivist approach to program evaluation is heavily philosophical, service oriented, and paradigm-driven. Constructivism rejects the existence of any ultimate reality and employs a subjectivist epistemology. It sees knowledge gained as one or more human constructions, uncertifiable, and constantly problematic and changing. It places the evaluators and program stakeholders at the center of the inquiry process, employing all of them as the evaluation's "human instruments." The approach insists that evaluators be totally ethical in respecting and advocating for all the participants, especially the disenfranchised. Evaluators are authorized, even expected, to maneuver the evaluation to emancipate and empower involved or affected disenfranchised

people. Evaluators do this by raising stakeholders' consciousness so that they are energized, informed, and assisted to transform their world. The evaluator must respect participants' free will in all aspects of the inquiry and should empower them to help shape and control the evaluation activities in their preferred ways. The inquiry process must be consistent with effective ways of changing and improving society. Thus, stakeholders must play a key role in determining the evaluation questions and variables. Throughout the study, the evaluator regularly and continuously informs and consults stakeholders in all aspects of the study. The approach rescinds any special privilege of scientific evaluators to work in secret and control/manipulate human subjects. In guiding the program evaluation, the evaluator balances verification with a quest for discovery, balances rigor with relevance and the use of quantitative and qualitative methods. The evaluator also provides rich and deep description in preference to precise measurements and statistics. He or she employs a relativist perspective to obtain and analyze findings, stressing locality and specificity over generalizability. The evaluator posits that there can be no ultimately correct conclusions. He or she exalts openness and the continuing search for more informed and illuminating constructions.

This approach is as much recognizable for what it rejects as for what it proposes. In general, it strongly opposes positivism as a basis for evaluation, with its realist ontology, objectivist epistemology, and experimental method. It rejects any absolutist search for correct answers. It directly opposes the notion of value-free evaluation and attendant efforts to expunge human bias. It rejects positivism's deterministic and reductionist structure and its belief in the possibility of fully explaining studied programs.

Advance organizers of the constructivist approach are basically the philosophical constraints placed on the study, as noted above, including the requirement of collaborative, unfolding inquiry. The main purpose of the approach is to determine and make sense of the variety of constructions that exist or emerge among stakeholders. Inquiry is kept open to ongoing communication and to the gathering, analysis, and synthesis of further constructions. One construction is not considered more "true" than others, but some may be judged as more informed and sophisticated than others. All evaluation conclusions are viewed as indeterminate, with the continuing possibility of finding better answers. All constructions are also context dependent. In this respect, the evaluator does define boundaries on what is being investigated.

The questions addressed in constructivist studies cannot be determined independently of participants' interactions. Evaluator and stakeholders together identify the questions to be addressed. Questions emerge in the process of formulating and discussing the study's rationale, planning the schedule of discussions, and obtaining various initial persons' views of the program to be evaluated. The questions develop further over the course of the approach's hermeneutic and dialectic processes. Questions may or may not cover the full range of issues involved in assessing something's merit and worth. The set of questions to be studied is never considered fixed.

The constructivist methodology is first divergent, then convergent. Through the use of hermeneutics, the evaluator collects and describes alternative individual constructions on an evaluation question or issue, ensuring that each depiction meets with the respondent's approval. Communication channels are kept open throughout the inquiry, and all respondents are encouraged and facilitated to make their inputs and are kept apprised of all aspects of the study. The evaluator then moves to a dialectical process aimed at achieving as much consensus as possible among different constructions. Respondents are provided with opportunities to review the full range of constructions along with other relevant information. The evaluator engages the respondents in a process of studying and contrasting existing constructions, considering relevant contextual and other information, reasoning out the differences among the constructions, and moving as far as they can toward a consensus. The constructivist evaluation is, in a sense, never-ending. There is always more to learn, and finding ultimately correct answers is considered impossible.

Guba and Lincoln (1985, 1989) are pioneers in applying the constructivist approach to program evaluation. Bhola (1998), a disciple of Guba, has extensive experience in applying the constructivist approach to evaluating programs in Africa. In agreement with Guba, he stresses that evaluations are always a function not only of the evaluator's approach and interactions with stakeholders, but also of his or her personal history and outlook. Thomas Schwandt (1984), another disciple of Guba, has written extensively about the philosophical underpinnings of constructivist evaluation. Fetterman's (1994) empowerment evaluation approach is closely aligned with constructivist evaluation, since it seeks to engage and serve all stakeholders, especially those with little influence. However, there is a key difference between the constructivist and empowerment evaluation approaches. While the constructivist evaluator retains control of the evaluation and works with stakeholders to develop a consensus, the empowerment evaluator "gives away" authority for the evaluation to stakeholders, serving in a technical assistance role.

The constructivist approach can be applied usefully when evaluator, client, and stakeholders in a program fully agree that the approach is appropriate and that they will cooperate. They should reach agreement based on an understanding of what the approach can and cannot deliver. They need to accept that questions and issues to be studied will unfold throughout the process. They also should be willing to receive ambiguous, possibly contradictory findings, reflecting stakeholders' diverse perspectives. They should know that the shelf life of the findings is likely to be short (not unlike any other evaluation approach, but clearly acknowledged in the constructivist approach). They also need to value qualitative information that largely reflects the variety of stakeholders' perspectives and judgments. However, they should not expect to receive definitive pre-post measures of outcomes or statistical conclusions about causes and effects. While these persons can hope for achieving a consensus in the findings, they should agree that such a consensus might not emerge and that in any case such a consensus would not generalize to other settings or time periods.

This approach has a number of advantages. It is exemplary in fully disclosing the whole evaluation process and its findings. It is consistent with the principle of effective change that people are more likely to value and use something (read *evaluation* here) if they are consulted and involved in its development. It also seeks to directly involve the full range of stakeholders who might be harmed or helped by the evaluation as important, empowered partners in the evaluation enterprise. It is said to be educative for all the participants, whether or not a consensus is reached. It also lowers expectations for what clients can learn about causes and effects. While it does not promise final answers, it moves from a divergent stage, in which it searches widely for insights and judgments, to a convergent stage in which some unified answers are sought. In addition, it uses participants as instruments in the evaluation, thus taking advantage of their relevant experiences, knowledge, and value perspectives; this greatly reduces the burden of developing, field-testing, and validating information collection instruments before using them. The approach makes effective use of qualitative methods and triangulates findings from different sources.

The approach, however, is limited in its applicability and has some disadvantages. Because of the need for full involvement and ongoing interaction through both the divergent and convergent stages, it is often difficult to produce the timely reports that funding agencies and decision makers demand. Further, if the approach is to work well, it requires the attention and responsible participation of a wide range of stakeholders. The approach seems to be unrealistically Utopian in this regard: widespread, grass-roots interest and participation are often hard to obtain and sustain throughout a program evaluation. The situation can be exacerbated by a continuing turnover of stakeholders. While the process emphasizes and promises openness and full disclosure, some participants do not want to tell their private thoughts and judgments to the world. Moreover, stakeholders sometimes are poorly informed about the issues being addressed in an evaluation and thus are poor data sources. It can be unrealistic to expect that the evaluator can and will take the needed time to inform, and then meaningfully involve, those who begin as basically ignorant of the program being assessed. Further, constructivist evaluations can be greatly burdened by itinerant evaluation stakeholders who come and go, and reopen questions previously addressed and question consensus previously reached. There is the further issue that some evaluation clients do not take kindly to evaluators who are prone to report competing, perspectivist answers, and not take a stand regarding a program's merit and worth. Many clients are not attuned to the constructivist philosophy and they may value reports that mainly include hard data on outcomes and assessments of statistical significance. They may expect reports to be based on relatively independent perspectives that are free of program participants' conflicts of interest. Since the constructivist approach is a countermeasure to assigning responsibility for successes and failures in a program to certain individuals or groups, many policy boards, administrators, and financial sponsors might see this rejection of accountability as unworkable and unacceptable. It is easy to say that all persons in a program should share the glory or the disgrace; but try to tell this to an excep-

tionally hardworking and effective teacher in a school program where virtually no one else tries or succeeds.

Approach 21: Deliberative Democratic Evaluation

Perhaps the newest entry in the program evaluation models enterprise is the deliberative democratic approach advanced by House and Howe (Chapter 22, this volume). The approach functions within an explicit democratic framework and charges evaluators to uphold democratic principles in reaching defensible conclusions. It envisions program evaluation as a principled, influential societal institution, contributing to democratization through the issuing of reliable and valid claims.

The advance organizers of deliberative democratic evaluation are seen in its three main dimensions: democratic participation, dialogue to examine and authenticate stakeholders' inputs, and deliberation to arrive at a defensible assessment of a program's merit and worth. All three dimensions are considered essential in all aspects of a sound program evaluation.

In the democratic dimension, the approach proactively identifies and arranges for the equitable participation of all interested stakeholders throughout the course of the evaluation. Equity is stressed, and power imbalances in which the message of powerful parties would dominate the evaluation message are not tolerated. In the dialogic dimension, the evaluator engages stakeholders and other audiences to assist in compiling preliminary findings. Subsequently, the collaborators seriously discuss and debate the draft findings to ensure that no participant's views are misrepresented. In the culminating deliberative stage, the evaluator(s) honestly considers and discusses with others all inputs obtained but then renders what he or she considers a fully defensible assessment of the program's merit and worth. All interested stakeholders are given voice in the evaluation, and the evaluator acknowledges their views in the final report, but may express disagreement with some of them. The deliberative dimension sees the evaluator(s) reaching a reasoned conclusion by reviewing all inputs; debating them with stakeholders and others; reflecting deeply on all the inputs; then reaching a defensible, well-justified conclusion.

The purpose of the approach is to employ democratic participation in the process of arriving at a defensible assessment of a program. The evaluator(s) determines the evaluation questions to be addressed, but does so through dialogue and deliberation with engaged stakeholders. Presumably, the bottom-line questions concern judgments about the program's merit and its worth to stakeholders.

Methods employed may include discussions with stakeholders, surveys, and debates. Inclusion, dialogue, and deliberation are considered relevant at all stages of an evaluation—inception, design, implementation, analysis, synthesis, write-up, presentation, and discussion. House and Howe present the following ten questions for assessing the adequacy of a democratic deliberative evaluation: Whose interests are represented? Are major stakeholders represented? Are any excluded? Are there serious power imbalances? Are there procedures to control imbalances? How do people participate in the evaluation? How authentic is their participation? How involved is

their interaction? Is there reflective deliberation? How considered and extended is the deliberation?

Ernest House originated this approach. He and Kenneth Howe say that many evaluators already implement their proposed principles, and point to an article by Karlsson (1998) to illustrate their approach. They also refer to a number of authors who have proposed practices that at least in part are compatible with the democratic dialogic approach.

The approach is applicable when a client agrees to fund an evaluation that requires democratic participation of at least a representative group of stakeholders. Thus, the funding agent must be willing to give up sufficient power to allow inputs from a wide range of stakeholders, early disclosure of preliminary findings to all interested parties, and opportunities for the stakeholders to play an influential role in reaching the final conclusions. Obviously, a representative group of stakeholders must be willing to engage in open and meaningful dialogue and deliberation at all stages of the study.

The approach has many advantages. It is a direct attempt to make evaluations just. It strives for democratic participation of stakeholders at all stages of the evaluation. It seeks to incorporate the views of all interested parties, including insiders and outsiders, disenfranchised persons and groups, as well as those who control the purse strings. Meaningful democratic involvement should direct the evaluation to the issues that people care about and incline them to respect and use the evaluation findings. The approach employs dialogue to examine and authenticate stakeholders' inputs. A key advantage over some other advocacy approaches is that the democratic deliberative evaluator expressly reserves the right to rule out inputs that are considered incorrect or unethical. The evaluator is open to all stakeholders' views, carefully considers them, but then renders as defensible a judgment of the program as possible. He or she does not leave the responsibility for reaching a defensible final assessment to a majority vote of stakeholders—some of whom are sure to have conflicts of interest and be uninformed. In rendering a final judgment, the evaluator ensures closure.

As House and Howe have acknowledged, the democratic dialogic approach is, at this time, unrealistic and often cannot be fully applied. The approach—in offering and expecting full democratic participation in order to make an evaluation work—reminds me of a colleague who used to despair of ever changing or improving higher education. He would say that changing any aspect of our university would require getting every professor to withhold her or his veto. In view of the very ambitious demands of the democratic dialogic approach, House and Howe have proposed it as an ideal to be kept in mind even though evaluators will seldom, if ever, be able to achieve it.

Approach 22: Utilization-Focused Evaluation

The utilization-focused approach is explicitly geared to ensure that program evaluations make an impact (Patton, Chapter 23, this volume). It is a process for making choices about an evaluation study in collaboration with a targeted group of

priority users, selected from a broader set of stakeholders, in order to focus effectively on their intended uses of the evaluation. All aspects of a utilization-focused program evaluation are chosen and applied to help the targeted users obtain and apply evaluation findings to their intended uses, and to maximize the likelihood that they will. Such studies are judged more for the difference they make in improving programs and influencing decisions and actions than for their elegance or technical excellence. No matter how good an evaluation report is, if it only sits on the shelf gathering dust, then it will not contribute positively to the evaluation and possibly should not have been written.

Placement of Patton's evaluation approach within the category system used in this chapter was problematic. His chapter was placed in the Social Agenda section because it requires democratic participation of a representative group of stakeholders, whom it empowers to determine the evaluation questions and information needs. Patton gives away such authority over the evaluation to increase the likelihood that the findings will be used. However, utilization-focused evaluations do not necessarily advocate any social agenda, such as affirmative action to right injustices and better serve the poor. While the approach is in agreement with the improvement/accountability-oriented approaches in guiding decisions, promoting impacts, and invoking the Joint Committee (1994) *Program Evaluation Standards*, it does not quite fit there. It does not, for example, require assessments of merit and worth. In fact, Patton essentially has said that his approach is pragmatic and ubiquitous. In the interest of getting findings used, he will draw upon any legitimate approach to evaluation, leaving out any parts that might impede use. As for the dilemma of categorizing the Utilization-Based Evaluation Model, the reader will note that we placed Patton's chapter in this volume, not in the Social Agenda Evaluation Models section, but in the section on overarching matters (Section V).

The advance organizers of utilization-focused program evaluations are, in the abstract, the possible users and uses to be served. Working from this initial conception, the evaluator moves as directly as possible to identify in concrete terms the actual users to be served. Through careful and thorough analysis of stakeholders, the evaluator identifies the multiple and varied perspectives and interests that should be represented in the study. He or she then selects a group that is willing to pay the price of substantial involvement and that represents the program's stakeholders. The evaluator then engages this client group to clarify why they need the evaluation, how they intend to apply its findings, how they think it should be conducted, and what types of reports (e.g., oral and/or printed) should be provided. He or she facilitates users' choices by supplying a menu of possible uses, information, and reports for the evaluation. This is done not to supply the choices but to help the client group thoughtfully focus and shape the study. The main possible uses of evaluation findings contemplated in this approach are assessment of merit and worth, improvement, and generation of knowledge. The approach also values the evaluation process itself, seeing it as helpful in enhancing shared understandings among stakeholders, bringing support to a program, promoting participation in it, and developing and strengthening organizational capacity. According to Patton, when

the evaluation process is sound and functional, a printed final report may not be needed.

In deliberating with intended users, the evaluator emphasizes that the program evaluation's purpose must be to give them the information they need to fulfill their objectives. Such objectives may include socially valuable aims such as combating problems of illiteracy, crime, hunger, homelessness, unemployment, child abuse, spouse abuse, substance abuse, illness, alienation, discrimination, malnourishment, pollution, and bureaucratic waste. However, it is the targeted users who determine the program to be evaluated, what information is required, how and when it must be reported, and how it will be used.

In this approach, the evaluator is no iconoclast, but rather the intended users' servant. Among other roles, he or she is a facilitator. The evaluation should meet the full range of professional standards for program evaluations, not just utility. The evaluator must therefore be an effective negotiator, standing on principles of sound evaluation, but working hard to gear a defensible program evaluation to the targeted users' evolving needs. The utilization-focused evaluation is considered situational and dynamic. Depending on the circumstances, the evaluator may play any of a variety of roles—trainer, measurement expert, internal colleague, external expert, analyst, spokesperson, or mediator.

The evaluator works with the targeted users to determine the evaluation questions. Such questions are to be determined locally, may address any of a wide range of concerns, and probably will change over time. Example foci are processes, outcomes, impacts, costs, and cost benefits. The chosen questions are kept front and center and provide the basis for information collection and reporting plans and activities, so long as users continue to value and pay attention to the questions. Often, however, the evaluator and client group will adapt, change, or refine the questions as the evaluation unfolds.

All evaluation methods are fair game in a utilization-focused program evaluation. The evaluator will creatively employ whatever methods are relevant (e.g., quantitative and qualitative, formative and summative, naturalistic and experimental). As far as possible, the utilization-focused evaluator puts the client group in "the driver's seat" in determining evaluation methods to ensure that the evaluator focuses on their most important questions; collects the right information; applies the relevant values; answers the key action-oriented questions; uses techniques they respect; reports the information in a form and at a time to maximize use; convinces stakeholders of the evaluation's integrity and accuracy; and facilitates the users' study, application, and—as appropriate—dissemination of findings. The bases for interpreting evaluation findings are the users' values, and the evaluator will engage in values clarification to ensure that evaluative information and interpretations serve users' purposes. Users are actively involved in interpreting findings. Throughout the evaluation process, the evaluator balances the concern for utility with provisions for validity and cost-effectiveness.

In general, the method of utilization-focused program evaluation is labeled "active-reactive-adaptive and situationally responsive," emphasizing that the method-

ology evolves in response to ongoing deliberations between evaluator and client group, and in consideration of contextual dynamics. Patton (1997) says that “Evaluators are active in presenting to intended users their own best judgments about appropriate evaluation focus and methods; they are reactive in listening attentively and respectfully to others’ concerns; and they are adaptive in finding ways to design evaluations that incorporate diverse interests . . . while meeting high standards of professional practice (p. 383).”

Patton (1980, 1982, 1994, 1997, Chapter 23, this volume) is the leading proponent of utilization-based evaluation. Other advocates of the approach are Alkin (1995), Cronbach and Associates (1980), Davis and Salasin (1975), and the Joint Committee on Standards for Educational Evaluation (1981b, 1994).

As defined by Patton, the approach has virtually universal applicability. It is situational and can be tailored to meet any program evaluation assignment. It carries with it the integrity of sound evaluation principles. Within these general constraints, the evaluator negotiates all aspects of the evaluation to serve specific individuals who need to have a program evaluation performed and who intend to make concrete use of the findings. The evaluator selects from the entire range of evaluation techniques those that best suit the particular evaluation. And the evaluator plays any of a wide range of evaluation and improvement-related roles that fit the local needs. The approach requires a substantial outlay of time and resources by all participants, both for conducting the evaluation and the needed follow-through.

The approach is geared to maximizing evaluation impacts. It fits well with a key principle of change: Individuals are more likely to understand, value, and use the findings of an evaluation if they were meaningfully involved in the enterprise. As Patton (1997) says, “by actively involving primary intended users, the evaluator is training users in use, preparing the groundwork for use, and reinforcing the intended utility of the evaluation” (p. 22). The approach engages stakeholders to determine the evaluation’s purposes and procedures and uses their involvement to promote the use of findings. It takes a more realistic approach to stakeholder involvement than some other advocacy approaches. Rather than trying to reach and work with all stakeholders, Patton’s approach works concretely with a select, representative group of users. The approach emphasizes values clarification and attends closely to contextual dynamics. It may selectively use any and all relevant evaluation procedures and triangulates findings from different sources. Finally, the approach stresses the need to meet all relevant standards for evaluations.

Patton sees the main limitation of the approach to be turnover of involved users. Replacement users may require that the program evaluation be renegotiated, which may be necessary to sustain or renew the prospects for evaluation impacts. But it can also derail or greatly delay the process. Further, the approach seems to be vulnerable to corruption by user groups, since they are given so much control over what will be looked at, the questions addressed, the methods employed, and the information obtained. Stakeholders with conflicts of interest may inappropriately influence the evaluation. For example, they may inappropriately limit the evaluation to a subset of the important questions. It may also be almost impossible to get a

representative users group to agree on a sufficient commitment of time and safeguards to ensure an ethical, valid process of data collection, reporting, and use. Moreover, effective implementation of this approach requires a highly competent, confident evaluator who can approach any situation flexibly without compromising basic professional standards. Strong skills of negotiation are essential, and the evaluator must possess expertise in the full range of quantitative and qualitative evaluation methods, strong communication and political skills, and working knowledge of all applicable standards for evaluations. Unfortunately, not many evaluators are sufficiently trained and experienced to meet these requirements.

The utilization-based approach to evaluation concludes this chapter's discussion of social agenda/advocacy approaches. The four approaches concentrate on making evaluation an instrument of social justice and modesty and candor in presenting findings that often are ambiguous and contradictory. All four approaches promote utilization of findings through involvement of stakeholders.

BEST APPROACHES FOR 21ST CENTURY EVALUATIONS

Of the variety of evaluation approaches that emerged during the 20th century, nine can be identified as the strongest and most promising for continued use and development beyond the year 2000. The other 13 approaches also have varying degrees of merit, but I chose in this section to focus on what I judged to be the most promising approaches. The ratings of these approaches appear in Table 1. They are listed in order of merit, within the categories of Improvement/Accountability, Social Mission/Advocacy, and Questions/Methods evaluation approaches. The ratings are based on the Joint Committee *Program Evaluation Standards* and were derived by the author using a special checklist keyed to the *Standards*.

All nine of the approaches earned overall ratings of Very Good, except Accreditation, which was judged Good overall.¹ The Utilization-Based and Client-Centered approaches received Excellent ratings in the standards area of Utility, while the Decision/Accountability approach was judged Excellent in provisions for Accuracy. The rating of Good in the Accuracy area for the Outcomes Monitoring/Value-Added approach was due not to this approach's low merit in what it does, but to the narrowness of questions addressed and information used; in its narrow sphere of application, the approach provides technically sound information. The comparatively lower ratings given to the Accreditation approach result from its being labor intensive and expensive; its susceptibility to conflict of interest; its overreliance on self-reports and brief site visits; and its insular resistance to independent metaevaluations. Nevertheless, the distinctly American and pervasive accreditation approach is entrenched. All who use it are advised to strengthen it in the areas of weakness identified in this chapter. The Consumer-Oriented approach also deserves its special place, with its emphasis on independent assessment of developed products and services. While the approach is not especially applicable to internal evaluations for improvement, it complements such approaches with an outsider, expert view that becomes important when products and services are put up for dissemination.

The Case Study approach scored surprisingly well, considering that it is focused on use of a particular technique. An added bonus is that it can be employed on its own or as a component of any of the other approaches. As mentioned previously, the Democratic Deliberative approach is new and appears to be promising for testing and further development. Finally, the Constructivist approach is a well-founded, mainly qualitative approach to evaluation that systematically engages interested parties to help conduct both the divergent and convergent stages of evaluation. All in all, the nine approaches summarized in Table 1 bode well for the future application and further development of alternative program evaluation approaches.










SUMMARY AND CONCLUSIONS

The last half of the 20th century saw considerable development of program evaluation approaches. In this chapter, 22 identified approaches were grouped as pseudoevaluations, questions/methods-oriented evaluations, improvement/accountability-oriented evaluations, and social mission/advocacy evaluations. Apart from pseudoevaluations, there is among the approaches an increasingly balanced quest for rigor, relevance, and justice. Clearly, the approaches are showing a strong orientation to stakeholder involvement and the use of multiple methods.

When compared to professional standards for program evaluations, the best approaches are decision/accountability, utilization-based, client-centered, consumer-oriented, case study, democratic deliberative, Constructivist, accreditation, and outcomes monitoring. While House and Howe's (Chapter 22, this volume) democratic deliberative approach is new and in their view utopian, it has many elements of a sound, effective evaluation approach and merits study, further development, and trial. The worst bets were found to be the politically controlled, public relations, accountability (especially payment by results), clarification hearings, and program theory-based approaches. The rest fell in the middle. A critical analysis of the approaches has important implications for evaluators, those who train evaluators, theoreticians concerned with devising better concepts and methods, and those engaged in professionalizing program evaluation.

A major consideration for the practitioner is that evaluators may encounter considerable difficulties if their perceptions of the study being undertaken differ from those of their clients and audiences. Frequently, clients want a politically advantageous study performed, while the evaluator wants to conduct questions/methods-oriented studies that allow him or her to exploit the methodologies in which he or she was trained. Moreover, audiences usually want values-oriented studies that will help them determine the relative merits and worth of competing programs, or advocacy evaluations that will give them voice in the issues that affect them. If evaluators ignore the likely conflicts in purposes, the program evaluation is probably doomed to fail. At an evaluation's outset, evaluators must be keenly sensitive to their own agenda for the study, as well as those that are held by the client and the other right-to-know audiences. Further, the evaluator should advise involved parties of possible conflicts in the evaluation's purposes and should, at the beginning, negotiate a common understanding of the evaluation's purpose and the appropriate

Table 1. RATINGS Strongest Program Evaluation Approaches Within types, listed in order of compliance with *The Program Evaluation Standards*

Evaluation Approach	Graph of overall merit	Overall Score & Rating	UTILITY Rating	FEASIBILITY Rating	PROPRIETY Rating	ACCURACY Rating
	0 P F G VG E 100					
IMPROVEMENT/ACCOUNTABILITY						
Decision/Accountability		92 (V G)	90 (V G)	92 (V G)	88 (V G)	98 (E)
Consumer Orientation		81 (V G)	81 (V G)	75 (V G)	91 (V G)	81 (V G)
Accreditation		60 (G)	71 (V G)	58 (G)	59 (G)	50 (G)
SOCIAL MISSION/ADVOCACY						
Utilization-Focused		87 (V G)	96 (E)	92 (V G)	81 (V G)	79 (V G)
Client-Centered		87 (V G)	93 (E)	92 (V G)	75 (V G)	88 (V G)
Democratic Deliberative		83 (V G)	96 (E)	92 (V G)	75 (V G)	69 (V G)
Constructivist		80 (V G)	82 (V G)	67 (G)	88 (V G)	83 (V G)
QUESTIONS/METHODS						
Case Study		80 (V G)	68 (V G)	83 (V G)	78 (V G)	92 (V G)
Outcomes Monitoring/Value-Added		72 (V G)	71 (V G)	92 (V G)	69 (V G)	56 (G)

The procedures behind the ratings: The author rated each evaluation approach on each of the 30 Joint Committee program evaluation standards by judging whether the approach endorses each of 10 key features of the standard. He judged the approach's adequacy on each standard as follows: 9-10 Excellent, 7-8 Very Good, 5-6 Good, 3-4 Fair, 0-2 Poor. The score for the approach on each of the 4 categories of standards (Utility, Feasibility, Propriety, Accuracy) was then determined by summing the following products: 4 x number of Excellent ratings, 3 x number of Very Good ratings, 2 x number of Good ratings, 1 x number of Fair ratings. Judgments of the approach's strength in satisfying each category of standards were then determined according to percentages of the possible quality points for the category of standards as follows: 93%-100% Excellent, 68%-92% Very Good, 50%-67% Good, 25%-49% Fair, 0%-24% Poor. This was done by converting each category score to the percent of the maximum score for the category and multiplying by 100. The 4 equalized scores were then summed, divided by 4, and compared to the total maximum value, 100. The approach's overall merit was then judged as follows: 93-100 Excellent, 68-92 Very Good, 50-67 Good, 25-49 Fair, 0-24 Poor. Regardless of the approach's total score and overall rating, a notation of unacceptable would have been attached to any approach receiving a poor rating on the vital standards of P1 Service Orientation, A5 Valid Information, A10 Justified Conclusions, A11 Impartial Reporting. The author's ratings were based on his knowledge of the Joint Committee Program Evaluation Standards, his many years of studying the various evaluation models and approaches, and his experience in seeing and assessing how some of these models and approaches worked in practice. He chaired the Joint Committee on Standards for Educational Evaluation during its first 13 years and led the development of the first editions of both the program and personnel evaluation standards. Nevertheless, his ratings should be viewed as only his personal set of judgments of these models and approaches. Also, his conflict of interest is acknowledged, since he was one of the developers of the Decision/Accountability approach. The scale ranges in the above graphs are P=Poor, F=Fair, G=Good, VG=Very Good, E=Excellent.

approach. Evaluators should also regularly inform participants in their evaluations of the selected approach's logic, rationale, process, and pitfalls. This will enhance stakeholders' cooperation and constructive use of findings.

Evaluation training programs should effectively address the ferment over and development of new program evaluation approaches. Trainers should directly teach their students about expanding and increasingly sophisticated program evaluation approaches. When students clearly understand the approaches, and provided they know when and how to apply them they will be in a position to discern which approaches are worth using and which are not.

For the theoretician, a main point is that the approaches all have inherent strengths and weaknesses. In general, the weaknesses of the politically oriented studies are that they are vulnerable to conflicts of interest and may mislead an audience into developing an unfounded, perhaps erroneous, judgment of a program's merit and worth. The main problem with the questions/methods-oriented studies is that they often address questions that are too narrow to support a full assessment of merit and worth. However, it is also noteworthy that these types of studies compete favorably with improvement/accountability-oriented evaluation studies and social agenda/advocacy studies in the efficiency of methodology employed. Improvement/accountability-oriented studies, with their concentration on merit and worth, undertake a very ambitious task, for it is virtually impossible to fully and unequivocally assess any program's ultimate worth. Such an achievement would require omniscience, infallibility, an unchanging environment, and an unquestioned, singular value base. Nevertheless, the continuing attempt to address questions of merit and worth is essential for the advancement of societal programs. Finally, the social mission/advocacy studies are to be applauded for their quest for equity as well as excellence in programs being studied. They model their mission by attempting to make evaluation a participatory, democratic enterprise. Unfortunately, many pitfalls attend such utopian approaches. These approaches are especially susceptible to bias, and they face practical constraints in involving, informing, and empowering targeted stakeholders.

For the evaluation profession itself, the review of program evaluation approaches underscores the importance of standards and metaevaluations. Professional standards are needed to maintain a consistently high level of integrity in uses of the various approaches. All legitimate approaches are enhanced when evaluators key their studies to professional standards for evaluation and obtain independent reviews of their evaluations.

NOTE

1. A test to determine differences between overall ratings of models based on one approach that sums across 30 equally weighted standards and the approach used in Table 1 that provides the average of scores for four equally weighted categories of standards (having different numbers of standards in each category) yielded a Pearson correlation of .968.