CHAPTER 1

OVERVIEW OF THE EVALUATION FIELD

Evaluation is perhaps society's most fundamental discipline; it is an essential characteristic of the human condition; and it is the single most important and sophisticated cognitive process in the repertoire of human reasoning and logic (Osgood, Suci, & Tannenbaum, 1957). It permeates all areas of human activity and has important implications for maintaining and improving services and protecting citizens in all areas of interest to society. Evaluation is a process for giving attestations to such matters as reliability, effectiveness, cost-effectiveness, efficiency, safety, ease of use, and probity. Society and individual clients are at risk to the extent that services, products, and other objects of interest are of poor quality. Evaluation serves society by providing affirmations of worth, value, progress, accreditation, and accountability—and, when necessary, a credible, defensible, nonarbitrary basis for terminating bad programs or, conversely, expanding good programs.

What Are Appropriate Objects of Evaluations and Related Subdisciplines of Evaluation?

In general, we refer to objects of evaluations as evaluands. When the evaluand is a person, however, we follow Scriven's recommendation to label the person whose qualifications or performance is being evaluated as the evaluee (Scriven, 1991). Objects of evaluations may be programs, projects, policies, proposals, products, equipment, services, concepts and theories, data and other types of information, individuals, or organizations, among others. Although the practice of evaluation largely concentrates on

LEARNING OBJECTIVES

In this chapter you will learn about the following:

- **The distinction between formal and informal evaluation**
- **The potential contributions and limitations of formal evaluation**
- **Evaluation as a profession and its relationship to other professions**
- **Conceptual and operational definitions of evaluation**
- **Key criteria for evaluating programs, including merit and worth**
- **The roles of values clarification and setting standards in reaching evaluative conclusions**
- **Four main uses of evaluation**
- **Distinctions between formative evaluation and summative evaluation**
- **Distinctions between research and evaluation**
- **Historical milestones in the development of professional evaluation**

program evaluation, one can refer to a range of other areas of evaluative inquiry, such as personnel evaluation, product evaluation, portfolio evaluation, performance evaluation, proposal evaluation, and policy evaluation. The scope of evaluation applications broadens greatly when one considers the wide range of disciplines, activities, and endeavors to which evaluation applies. One can speak, for example, of educational evaluation, social and human services evaluation, arts evaluation, consumer product evaluation, human resources development and evaluation, city planning and evaluation, real estate appraising, engineering testing and evaluation, hospital evaluation, drug testing, manufacturing evaluation, science policy evaluation, evaluation of international development and international aid, agricultural experimentation, and environmental evaluation.

Are Evaluations Enough to Control Quality, Guide Improvement, and Protect Consumers?

The presence of sound evaluation does not necessarily guarantee high quality in services or that those in authority will heed the lessons of evaluation and take needed corrective actions. Evaluations provide only one of the ingredients needed for quality assurance and improvement. There are many examples of defective products that have harmed consumers not because of a lack of pertinent evaluative information, but because of a failure on the part of decision makers to heed and act on rather than ignore or cover up alarming evaluative information. The continued sales of the Corvair automobile after its developers and marketers knew of its rear-end collision fire hazard provides one clear example (see also Nader, 1965). Here we see that society has a critical need not only for competent evaluators but for evaluation-oriented decision makers as well. For evaluations to make a positive difference, policymakers, regulatory bodies, service providers, and others must obtain and act responsibly on evaluation findings. The production and appropriate use of sound evaluation constitute one of the most vital contributors to strong services and societal progress.

Evaluation as a Profession and Its Relationship to Other Professions

As a profession with important roles in society, evaluation has technical aspects requiring thorough and ongoing training. It possesses an extensive and rapidly developing professional literature containing information on evaluation models and methods and findings from research on evaluation (Christie, 2011; Coryn & Westine, 2013). Its research material evolves from, and is closely connected to, the wide range of evaluations conducted in all fields. Evaluation has many professional organizations, including the American Evaluation Association (AEA) and other state and national evaluation associations. Among the earliest known professional societies were the May 12th Group, Division H of the American Educational Research Association (AERA), the Evaluation Network (E-Net), and the Evaluation Research Society (ERS), all of which originated in the late 1960s and early 1970s.

In 1995 there were only five evaluation organizations worldwide, including AEA (ensuing from the merger of E-Net and ERS in 1986), the Canadian Evaluation Society (CES), the Australasian Evaluation Society (AES), the European Evaluation Society, and the Central American Evaluation Society. By 2006 there were more than fifty national and regional evaluation organizations throughout the world, most in developing countries (Segone & Ocampo, 2006). There are also university training programs in evaluation, among them the Interdisciplinary PhD in Evaluation (IDPE) program and the Evaluation, Measurement, and Research (EMR) program at Western Michigan University (Coryn, Stufflebeam, Davidson, & Scriven, 2010), as well as other evaluation graduate programs at Claremont Graduate University, the University of Illinois, The Ohio State University, the University of Minnesota, the University of North Carolina, the University of Virginia, and the University of California at Los Angeles (for historical trends in graduate training in evaluation, see LaVelle and Donaldson [2010]). In addition, the field has developed recognized standards for evaluation services, including the Joint Committee on Standards for Educational Evaluation's standards for evaluating programs, personnel, and students (1981, 1988, 1994, 2003, 2009, 2011) and the U.S. Government Accountability Office's *Government Auditing Standards* (U.S. General Accounting Office, 2002; U.S. Government Accountability Office, 2003, 2007), plus AEA's *Guiding Principles for Evaluators* (2004).

To communicate and disseminate developments in, thinking about, and critiques of evaluation theory, methods, and practice, professional journals and other types of publications dedicated exclusively to evaluation scholarship and practice began to appear in the 1970s (Coryn, 2007a). One of the field's earliest publications, which first appeared in 1974, was the journal *Evaluation and Program Planning*. This was followed in 1975 by the journal *Studies in Evaluation;* in 1976 by *Evaluation Review: A Journal of Applied Social Research;* some years later by the *American Journal of Evaluation* (formerly published under the titles *Evaluation News,* prior to 1986, and *Evaluation Practice,* between 1986 and 1997); *New Directions for Evaluation* (formerly *New Directions for Program Evaluation*) and *Evaluation & the Health Professions,* both of which appeared in 1978; and *Educational Evaluation and Policy Analysis,* which first appeared in 1979.

The 1980s were marked by the appearance of the *Canadian Journal of Program Evaluation,* which emerged in 1986; the *Journal of Personnel Evaluation in Education* (now published under the title *Educational Assessment, Evaluation, and Accountability*), which was first published in 1987; and *Practical Assessment, Research and Evaluation,* which was launched in 1988.

In the 1990s several additional journals appeared, including *Research Evaluation* in 1991, which is published in the Netherlands; *Evaluation: The International Journal of Theory, Research and Practice,* which is published in the United Kingdom; and the *Journal of Evaluation in Clinical Practice,* the last two having first been published in 1995. In the next decade several more scholarly journals devoted to evaluation emerged, including the *Evaluation Journal of Australasia,* which was first published in 2000, and the *Journal of MultiDisciplinary Evaluation,* which first appeared in 2004.

Despite the burgeoning number of scholarly evaluation journals, many evaluation scholars and practitioners disseminate their work in discipline-specific journals, including those found

in education, health and medicine, philosophy, psychology, and sociology, to name but a few. In addition to publishing in evaluation and discipline-specific journals, other evaluation scholars publish their work in subject-specific areas, such as measurement, research, and statistics.

As a distinct profession, evaluation is supportive of all other professions and in turn is supported by many of them; no profession could excel without evaluation. Services and research can lead to progress and stand up to public and professional scrutiny only if they are regularly subjected to rigorous evaluation and shown to be sound. Also, improvement-oriented self-evaluation is a hallmark of professionalism. Program leaders and all members of any profession are obligated to serve their clients well. This requires that they regularly evaluate, improve, and be accountable for their contributions. In the sense of assessing and improving quality and meeting accountability requirements, all professions (including evaluation) are dependent on evaluation. Moreover, evaluation draws concepts, criteria, and methods from such other fields as philosophy, political science, psychology, sociology, anthropology, education, economics, communication, public administration, information technology, statistics, and measurement. Clearly it is important for evaluators to recognize and build on the symbiotic relationships between evaluation and other fields of study and practice.

Improvements in programs and other evaluands can be enhanced and made more enduring to the extent that supporting evaluations are relevant, systematic, rigorous, and timely, and to the extent that clients make responsible use of findings. Evaluations that lack these aspects of discipline typically are fruitless, wasteful, and misleading. It bears mention, however, that evaluators can only do their best, and despite strenuous efforts to involve clients in evaluations, there is no certainty that clients will heed and act on sound evaluation findings. If rigorous evaluations are to make a positive difference, clients must play their part by helping focus evaluations, supporting their conduct, and making sound use of findings. Accordingly, evaluation training programs should prepare evaluation specialists *and* evaluation clients to collaborate effectively in conducting evaluations that are both rigorous and useful.

What Is Evaluation?

Mainly because there have been different approaches to evaluation over the years, definitions of the term *evaluation* have themselves varied. In earlier times, for example, evaluation was commonly associated with assessing achievement against clearly defined objectives, or (in schools and universities) conducting norm-referenced testing, or (in such fields as agriculture and experimental psychology) conducting controlled experiments. Also, particularly during the 1970s, many evaluations were keyed only or mainly to professional judgment. Subsequently, there was a growing belief that useful evaluations are ones that provide quality information for making and assessing decisions. These and other concepts of evaluation have elements of credibility, depending often on the type of evaluation study being undertaken and especially the needs of the evaluation users.

One of the earliest and still most prominent definitions of evaluation states that it means determining whether objectives have been achieved. Although following this definition can guide one to assess accomplishments in achieving one's valued goals, in a broader

sense the practice of objectives-based evaluation has serious limitations and can even be counterproductive. Especially from the perspective of an independent evaluation of consumer products or services, employing the objectives-based evaluation approach can cause an evaluation to fail. One of this approach's problems is that some objectives are unworthy of achievement. Surely evaluators must avoid judging a program as successful solely because it achieved its own objectives. Objectives might well be corrupt, dysfunctional, unimportant, not oriented to the needs of intended beneficiaries, or mainly reflective of a developer's profit motive or other conflicts of interest. Another problem is that this approach steers evaluations in the direction of looking only at outcomes. Many evaluations also should examine a program's objectives, structure, and processes, especially if the evaluation is to contribute to program improvement or adoption and adaptation by other service providers. Moreover, a focus on objectives might cause evaluators not to search for important, unintended consequences (often called side effects). These can be beneficial or harmful, as is often seen in prescription drugs that may do as much harm as good for particular users. In addition to the deficiencies already noted, evaluators employing an objectives-based evaluation approach provide feedback only at the completion of a program. Depending on the needs of the client group, evaluators often should also deliver timely findings for use in planning and in guiding programs toward successful outcomes.

Definitions that equate evaluation with any one methodology should be rejected. Sometimes evaluations based on randomized experiments can provide consumers with useful information on the comparative outcomes of competing programs, products, or services. However, in many evaluations, a controlled experimental approach would not be feasible, or it would be counterproductive; it might be unethical; or it might fail to address key questions about needs, objectives, plans, processes, side effects, and other important aspects of a program. Similarly, other useful methods—such as sample surveys, standardized testing, site visits, or self-studies—are far too narrow in the information they yield to provide a sufficient basis for most program evaluations. Evaluation, therefore, rather than being equated with any one methodology, should encompass all methods that are necessary and useful to reach defensible judgments of programs or other entities, and evaluators should selectively apply appropriate methods.

In this book we advocate a basic definition of evaluation put forth by the Joint Committee in 1994.[1] We present three variations of the definition. First, we present the definition as the Joint Committee stated it.[2] The committee's definition is general, calling for evaluations to be systematic and focused on determining an object's value. We then extend the general definition to highlight a range of important, generic criteria for consideration when assessing programs. Finally, we expand the definition further to outline the key steps involved in carrying out a sound evaluation and to stress the importance of obtaining both descriptive and judgmental information. We see the Joint Committee definition as especially appropriate and useful when conversing with uninitiated audiences and focusing their attention on the essence of evaluation. The second rendition can be helpful when discussing with clients or other stakeholder groups the values that should be referenced when evaluating a particular program or other object. The third version is especially appropriate when planning the required evaluation work.

Joint Committee Definition of Evaluation

The Joint Committee's 1994 definition states that "evaluation is the systematic assessment of the worth or merit of an object" (p. 3). Advantages of this definition are that it is concise and consistent with common dictionary meanings of evaluation. We see this as the definition to use when discussing evaluation at a general level. Notably, some alternative definitions of evaluation often also include significance, resulting in a formal definition of evaluation as the act or process of determining the merit, worth, or significance of something or the product of that process (Davidson, 2005; Scriven, 1991).

Evaluation's root term, *value,* denotes that evaluations essentially involve making value judgments. Accordingly, evaluations are not value-free (Scriven, 1993). They need to reference pertinent values. Depending on the particular program or other evaluand, such values may include effectiveness, efficiency, usability, cost, safety, legality, and so on. Also, the evaluation itself should be grounded in some defensible set of values for judging evaluations. Here we see that an evaluation is an evaluand that should adhere to relevant values for judging evaluations. These may include professionally defined principles (as in AEA's *Guiding Principles for Evaluators* or the Joint Committee program evaluation standards). Essentially, an evaluation—be it an assessment of a program or of an evaluation—should assess the evaluand's standing against the referenced values. This truism presents evaluators with the impetus to choose appropriate values for judging an evaluand. For example, in evaluating public services in the United States, evaluators should be true to, and sometimes specifically invoke, such democratic precepts as freedom, equity, due process of law, and the need for an enlightened society. Moreover, as will be explained in Chapter 3, evaluators should hold their evaluations to meeting such values as the Joint Committee–defined standards of utility, feasibility, propriety, accuracy, and evaluation accountability.

The Joint Committee's 1994 definition partially addresses the need to determine values by denoting that evaluations should assess merit or worth. Scriven (1991) pointed to the nontrivial differences between these two concepts and their important role in determining an evaluand's value. According to both Scriven (1991) and the Joint Committee (1994), merit essentially involves excellence or quality (that is, intrinsic value), whereas worth includes merit within the context of a particular culture and its associated needs, costs, and related circumstances (that is, extrinsic value). In Table 1.1, the essential characteristics and nature of these concepts are summarized, with further discussion following.

Merit

In general, one needs to look at the merit or quality of an evaluand. For example, does a state's special program for preparing middle school history teachers succeed in producing teachers who confidently and effectively teach middle school students about pertinent areas and periods of history? In general, does an evaluand do well what it is supposed to do? If so, it rates high on merit. The criteria of merit reside in the standards of the evaluand's particular discipline or area of service. In the example here, an evaluator might base her or

Table 1.1 Characteristics of Merit and Worth

Merit	Worth
May be assessed on any object of interest	Is assessed only on objects that have demonstrated an acceptable level of quality
Pertains to the intrinsic value of the object	Pertains to the extrinsic value of the object
Pertains to quality, that is, an object's level of excellence	Pertains to an object's quality and value or importance within a given context
Is assessed using the question, Does the object do well what it is intended to do?	Is assessed using the question, Is the object of high quality and also something a target group needs?
Is tied to accepted standards of quality for the type of object being evaluated	Is tied to accepted standards of quality and to data from a pertinent needs assessment
Concerns the object's rating on standards of quality and against competitive objects of the same type	Entails judgments of the object's quality and importance and value to a particular consumer group
May be assessed through comparison of an object with standards or competitive objects	Assessments of worth may be comparative or noncomparative

his assessment of merit on published standards of effective teaching and the state's required content for middle school history programs. Graduates of the program would thus be assessed on knowledge of the required history content and effectiveness in teaching the content. The subject program would be judged high on merit to the extent that graduates scored high on pertinent measures of content knowledge and teaching competence. Merit (or quality) then can be broadly understood as intrinsic excellence in the absence of costs.

Worth

An evaluand that rates high on merit might not be worthy. By worth, we refer to an evaluand's combination of excellence and service in an area of clear need within a specified context and considering the costs involved (both monetary and nonmonetary). Suppose the middle school program is a special emergency program developed and funded at a previous time when the state's colleges and universities were graduating too few history teachers to meet the needs of schools in the state. Suppose further that more recently, the state's universities have increased their production of competent middle school history teachers, and many of these new teachers cannot find jobs. Arguably, the state no longer needs the special emergency program, because the state's universities are now supplying more qualified middle school history teachers than the schools can employ. In this situation, although the state's special program has good merit, it now has low worth to the state and does not warrant continued investment of the state's scarce resources. We see in this example that this high-quality program's worth could be gauged only after an assessment of the need for the program's graduates. Here, we see that assessments of worth have to be keyed to assessments of need within the context of a particular setting and time period. Broadly, then, worth (or value) is quality under consideration of context and costs.

Needs

By a need, we refer to something that is necessary or useful for fulfilling a defensible purpose, without which satisfactory functioning cannot occur. We define a defensible purpose as a legitimately defined, desired end that is consistent with a guiding philosophy, set of professional standards, institutional mission, mandated curriculum, national constitution, or public referendum, for example. Other terms to describe defensible purposes are *legitimatized mandates, goals,* and *priorities*. In the middle school illustration, presumably the state curriculum requires that all students in the state be well educated in designated areas of history. This "defensible purpose" requires further that school districts employ competent history teachers. In this case, a competent history teacher fits our definition of an entity that is necessary or useful for fulfilling the defensible purpose of sound history instruction—that is, a need. Because of the state's finding that this need is now being fulfilled by state colleges and universities, this excellent special program would now meet the criterion of merit but not the criterion of worth. In reaching judgments of something's worth, evaluators should identify needs, then determine whether they are being met, partially met, or unmet in the context of interest (Stufflebeam, McCormick, Brinkerhoff, & Nelson, 1985; also see Coryn, Gugiu, Davidson, & Schröter, 2008).

Needs may be of either the outcome or the treatment variety (also see Davidson, 2005). An outcome need is a level of achievement or outcome in a particular area required to fulfill a defensible purpose, such as preparing students for higher education. For example, high school students need to develop competencies in mathematics, science, social studies, and language arts to enter top-notch colleges and universities. A treatment need is a certain service, service provider, or other helping agent required to meet an outcome need. To continue the example, a school district needs an appropriate curriculum and competent teachers (the treatment needs) to help students attain areas and levels of competence (the outcome needs) required for admission to high-level colleges and universities. One assesses both treatment and outcome needs to determine whether they are being met or unmet and whether they are consonant.

Typically, the meeting of outcome needs depends on meeting the treatment needs. For example, a dentist would be likely to check patients with tooth decay for use of fluoridated water or toothpaste. Here the outcome need (for cavity-free teeth) is not being met, and it is prudent to check on the treatment need related to fluoridation. In contrast, if patients evidence no tooth decay, the dentist would be unlikely to check them for use of fluoridated water or toothpaste.

Needs Assessments

In general, a needs assessment is a systematic investigation of the extent to which treatment and/or outcome needs are being met (Stufflebeam, McCormick, et al., 1985). One might posit that comprehensive high schools should serve the defensible purpose of developing students in all areas of human growth and development: intellectual, psychological, social, physical, moral, vocational, and aesthetic. In an appropriate range of curricular areas, a comparison of students' scores on standardized achievement tests to criterion-referenced standards or norms would give an indication of whether students' intellectual outcome needs were being met. However, considering the school's intention to develop students also in physical, aesthetic, psychological,

Table 1.2 Concepts Related to Needs and Needs Assessment

Concept	Definition	Example
Defensible purpose	A desired end that has been legitimated	Students' development of basic academic skills
Need	Something that is necessary or useful for fulfilling a defensible purpose	Competent, effective instruction in the basic skill areas
Outcome need	An achievement or outcome required to meet a defensible purpose	Students' demonstration of proficiency in specified areas, such as twelfth-grade math, science, and language arts
Treatment need	A certain service, competent service provider, or other helping agent	Competent instructors in twelfth-grade courses in math, science, and language arts
Needs assessment	A systematic investigation of the extent to which treatment and/or outcome needs are being met	Examination of students' scores on national tests and evaluation of the involved teachers

social, moral, and vocational areas, the achievement test scores would be insufficient to assess the full range of questions concerning students' outcome needs. To be valid, needs assessments have to be keyed to the full range of intended outcomes.

Some needs assessments will have a narrow scope and appropriately address a quite restricted construction of outcome needs. Even in a narrowly focused program, however, it can be important to consider a broad range of outcome and associated treatment needs. For example, school-based instrumental music programs contribute to students' development in such areas as social relations, psychological well-being, discipline, and employment. In general, an assessment of a program's worth should assess and gauge its quality and outcomes against the assessed outcome and treatment needs of beneficiaries. Table 1.2 offers a summary of key concepts related to needs and needs assessment.

Evaluations Should Be Systematic

Beyond its focus on merit and worth, the Joint Committee's 1994 definition of evaluation requires evaluations to be systematic. We acknowledge that the broad meaning of evaluation encompasses haphazard or unsystematic evaluations as well as carefully conducted evaluations. In this book, we are advocating for and discussing the latter. Indeed, this book is intended as a countermeasure to careless or corrupt inquiry processes that masquerade as evaluations and often lead to biased or otherwise erroneous interpretations of something's value. Instead, we seek the kind of evaluation that is conducted with great care—not only in collecting information of high quality but also in clarifying and providing a defensible rationale for the value perspectives used to interpret the findings and reach judgments and in communicating evaluation findings to the client and other audiences.

An Extended, Values-Oriented Definition of Evaluation

Although the Joint Committee's 1994 definition of evaluation has the positive features just noted, it omits mention of other key generic values. We thus extend the definition of evaluation as follows: evaluation is the systematic assessment of an object's merit, worth, probity, feasibility,

safety, significance, and/or equity. We see the values referenced in this definition as particularly important in a free and democratic society, but also acknowledge that we might have included additional values. Of course, evaluators have to engage in a good deal of values clarification as they plan their studies. Those included in our extended definition of evaluation are a good set to consider, but evaluators and their clients often should invoke additional values that pertain to the contexts of particular studies and the unique cultures and interests of stakeholders. Nonetheless, many sound and defensible evaluations will be strongly influenced by some or all of the five values we have added to merit and worth. In the following paragraphs we discuss and elucidate each of the values noted in the extended definition of evaluation.

Probity

During the writing of the first edition of this book (Stufflebeam & Shinkfield, 2007), there was a rash of public scandals in which major corporations based in the United States defrauded shareholders and others out of billions of dollars. Moreover, at least one major audit firm that contracted to evaluate a corporation's financial conditions and lawful operations was found to have complicity in that corporation's fraud. This audit firm compromised its independence and credibility. Not only did it fail to report on the probity of the corporation's accounting practices, but also it was alleged to have distorted and covered up information to hide the company's unethical, unlawful practices. Here we see that the corporation cheated its shareholders, workers, and ultimately the public, and that the audit firm was charged with aiding and abetting the fraud. On another front, there have been despicable scandals across the globe in which clergy and teachers have been found to be pedophiles.

Clearly, the public interest (broadly defined) requires that evaluations address considerations of probity: assessments of honesty, integrity, and ethical behavior. Unless there is no prospect for fraud or other illicit behavior, evaluators should check on a program's uncompromising adherence to moral standards. However, when probity breaches are expected, there is cause to err on the side of too much consideration of probity in evaluations of programs and institutions. To the extent required to form a defense against unethical behavior, probity considerations should be addressed in many evaluations of programs and in evaluations of evaluations.

Feasibility

Although a program (or service, or other type of service-oriented evaluand) might be of high quality, directed to an area of high need, and unimpeachable on ethical grounds, it still could fail on the criterion of feasibility. For example, it might consume more resources than required or cause no end of political turmoil. If either is the case, the program should at least be modified in these areas to make it more feasible. Obviously a good evaluation of the program should speak to this issue and, where appropriate, provide direction for making the program easy to apply, efficient in the use of time and resources, and politically and culturally viable. Evaluation of a program's feasibility sometimes justifies a cancellation decision. This argument in favor of assessing feasibility seems applicable to all programs (and to all service-oriented evaluands).

Safety

Many evaluations focus squarely on the issue of safety. Obvious cases are evaluations of new pharmaceutical products, medical treatments, laboratory equipment, meat and other food products, automobiles, railroad transportation services, air traffic control, oil and gas production and distribution, stepladders, electrical equipment, children's toys, and insecticides. Consumers are at risk to the extent that such commodities and services are manufactured, sold, and dispensed or delivered without rigorous safety checks and appropriate cautions. Moreover, many programs also require evaluations that examine the safety of facilities, equipment, activity regimens, crowd control practices, and others. To see the importance of safety evaluations in programs, one need only recall head injuries in football, lost teeth in ice hockey, heat strokes in a variety of outdoor sports, fires and explosions in school laboratories, fires resulting in many deaths due to improper fire escapes or fire drills, and fatalities due to faulty school buses or incompetent bus drivers. The criterion of safety applies to evaluations in all fields and to evaluations of programs as well as of products and services.

Significance

Another criterion that sometimes comes into play is a program's significance: its potential influence, importance, and visibility. Many programs are of only local or short-term interest. Other programs that have far-reaching implications should be examined and judged on the significance of their mission and outcomes. Such an assessment can be especially important in deciding whether and how far to disseminate lessons learned and in helping interested parties make sound decisions concerning adopting, adapting, and/or disseminating all or particular aspects of a program. Evaluators should consider the possibility that the program under study has far-reaching implications outside the local arena and possibly should be evaluated for its significance over time and in other settings.

Equity

The last generic evaluative criterion to be mentioned here is equity, which is predominantly tied to democratic societies. It argues for equal opportunities for all people and emphasizes freedom for all (also see House & Howe, 2000a). In the United States, an educational evaluation of a public educational service would be incomplete if it did not assess whether the service is provided for, and made available to, public school students from all sectors of society. This concept of equity is complex. It is not enough to say that public educational services may be sought and used by all people. As Kellaghan (1982) has argued, for example, when there is true equity in education, there will be seven indications of its existence:

1. A society's public educational services will be provided for all people.
2. People from all segments of the society will have equal access to the services.
3. There will be close to equal participation by all groups in the use of the services.
4. Levels of attainment—for example, years in the education system—will be substantially the same for different groups.

5. Levels of proficiency in achieving all of the education system's objectives will be equivalent for different groups.
6. Levels of aspiration for life pursuits will be similar across societal groups.
7. The education system will make similar impacts on improving the life accomplishments of all segments of the population (especially ethnic, gender, and socioeconomic groups) that the educational system serves.

We assert that equity, in the broadest sense, is an important criterion for all evaluations that involve delivering programs to groups of people.

Operationalizing Our Definition of Evaluation

The extended definition of evaluation has provided an expanded look at key generic criteria for evaluating programs. From the discussion, it is evident that the Joint Committee's 1994 definition of evaluation and our adaptation focused on generic evaluative criteria are deceptive in their apparent simplicity. When one takes seriously the root term *value,* then inevitably one must consider value perspectives of individuals, groups, and organizations, as well as information. The combining of these in efforts to reach determinations of the value of something cannot be ignored. To serve the needs of clients and other interested persons, the information supplied to support evaluative judgments should reflect the full range of appropriate values.

We now expand the definition to outline the main tasks in any program evaluation and denote the types of information to be collected. Our operational definition of evaluation states that evaluation is the systematic process of delineating, obtaining, reporting, and applying descriptive and judgmental information about some object's merit, worth, probity, feasibility, safety, significance, and/or equity. One added element in this definition concerns the generic steps in conducting an evaluation. The other new element is that evaluations should produce both descriptive and judgmental information.

It is important to note that the work of evaluation includes both interface/communication and technical tasks. In regard to the interface aspects, evaluators communicate with clients and other stakeholders in the interest of planning relevant evaluations; conveying clear, timely findings; and assisting with use of the findings. To ensure an evaluation's relevance and impact, the evaluator needs to effectively engage stakeholders in the evaluation's planning and use. The technical tasks are concerned with the research aspects of an evaluation: the collection, organization, analysis, and synthesis of information. Evaluators need to be competent in both the communication and technical aspects of evaluation (also see Stevahn, King, Ghere, & Minnema, 2005). This competence is best acquired through formal courses and experiences in planning, conducting, and reporting on a wide range of evaluations. We have characterized the work of evaluation in four tasks: delineating, obtaining, reporting, and applying. Part Four of this book addresses these process tasks in detail.

Delineating

The delineating task entails the evaluator's interacting with the client and other program stakeholders. The aim here is to focus the evaluation on key questions, identify key audiences,

clarify pertinent values and criteria, determine information requirements, project needed analyses, construct an evaluation budget, and effect contractual agreements to both govern and facilitate the evaluation work. Basically, the delineating task encompasses effective, interactive communication involving evaluator, client, and other interested parties and culminates in negotiated terms for the evaluation. Particular areas of needed expertise include audience analysis, listening, developing rapport, interviewing, situational and cultural analysis, values clarification, conceptualization, proposal development, negotiation, contracting, and budgeting. The results of these actions should set the stage for the ensuing data collection work. In fact, delineating activities extend throughout the evaluation in response to the program's changing circumstances, identification of new audiences, continuing interaction with stakeholders, and emerging information needs. Moreover, a delineation process that is carried out thoroughly and professionally establishes a basis for essential trust and rapport between an evaluator and a client group.

Obtaining

The obtaining task encompasses all of the work involved in collecting, correcting, organizing, analyzing, and synthesizing information. Key areas of required expertise are research design, sampling, measurement, interviewing, observation, site visits, archival studies, case studies, focus groups, photography, database development and management, statistics, content analysis, cost analysis, policy analysis, synthesis, and computer technology. Program evaluators need expertise in these and related technical areas to provide clients with sound, meaningful, and creditable information. Results of the obtaining work are grist for preparing and presenting oral and printed evaluation reports.

Reporting

In the reporting task, the evaluator provides the client and other audiences with feedback. Typically such work includes preparing and delivering interim oral and printed reports, multimedia presentations, press releases, printed final reports, and executive summaries, as well as ongoing informal exchanges with the evaluation's client and, often, stakeholders. The point of all such reporting activities is to communicate effectively and accurately the evaluation's findings in a timely manner to interested and right-to-know audiences and to foster effective uses of evaluation findings. Reporting activities, in various forms, occur throughout and after completion of an evaluation (Coryn, 2006). Particular areas of needed expertise are writing, formatting reports, editing, information technology, oral communication, leading of group discussions, and dissemination. Effective reporting sets the stage for applying the evaluation findings.

Applying

The applying task is under the control of the client and other users of the evaluation. Nevertheless, the evaluator should at least offer to assist in the application of findings. Such assistance might be follow-up workshops, a critique of the client group's plans to apply findings, coordination of focus group deliberations, or responses to questions from the client

or other users. We have found that clients appreciate this kind of assistance from evaluators. It is seen as a continuation of the evaluation itself, provided that the initiative comes from the client after the evaluator offers this "rounding-off" service. Assisting in the sound use of evaluation findings requires forethought and funding. In starting an evaluation, therefore, the evaluator and client should consider the possibility of the evaluator's involvement in the application stage and should plan, budget, and contract for such follow-up assistance as appropriate. To be effective in supporting the application of evaluation findings, evaluators need to be knowledgeable about principles and procedures of effective change and research on evaluation use (see also Alkin, Daillak, & White, 1979; Patton, 1997, 2008). Also, they need skills in the areas of communication, consulting, group process, and counseling (see also Dewey, Montrosse, Schröter, Sullins, & Mattox, 2008).

Descriptive and Judgmental Information

The final major feature of our operational definition of evaluation concerns the nature of information included in evaluations. From experience, we know that sound, useful evaluations are grounded in descriptive and judgmental information. In general, audiences for evaluation reports want to know what program was evaluated, how well it was carried out, and how good it was, requiring the evaluator to collect and report both descriptive and judgmental information.

Descriptive Information A final evaluation report should describe a program's goals, plans, funding, staffing, operations, and outcomes objectively (that is, as factual statements). As much as possible, the descriptive information should be kept separate from judgments of the program. Relatively pure, dispassionate descriptions of a program are needed to help evaluation audiences know, for example, what the evaluated program was like, how it was staffed and financed, how it operated, how much time was required for implementation, how much it cost, and what would be required to replicate it. The evaluator also has a vested interest in getting a clear view of the program apart from how other observers judged it. This is especially important when interpreting a program's outcomes and judging its success. For example, in judging the effects of a community's immunization program on childhood diseases, an evaluator needs to determine and report the extent to which the pertinent inoculations were administered to all the targeted children as planned. If they were not, the deficient outcome more likely is due to poor program implementation than defects in the program plan.

Judgmental Information Beyond the collection of descriptive information, it is equally important to gather, assess, and synthesize judgments of a program. According to the values-oriented definition of evaluation given earlier, sound evaluations involve judging an evaluand against a set of values. Values-oriented feedback can be a vital, positive force when it is integral to development, directed toward identifying strengths as well as weaknesses, focused on improving the evaluand, and grounded in evidence or at least experience with the program. Appropriate sources of judgments include program beneficiaries, program staff, pertinent experts, and (of course) the evaluator, among others. Such judgments are typically reached

through the integration or synthesis of facts (that is, descriptive information) and values, or the synthesis of multiple statements of value (Coryn, 2007; Davidson, 2005; Scriven, 1991, 1993).

How Good Is Good Enough? How Bad Is Intolerable? How Are These Questions Addressed?

Many evaluations carry a need to draw a definitive conclusion or make a definite decision on quality, safety, or some other variable. For example, funding organizations regularly have to decide which proposed projects to fund, basing their decisions on these projects' relative quality, costs, and importance compared with other possible uses of available funds (also see Coryn, Hattie, Scriven, & Hartmann, 2007; Coryn & Scriven, 2008; Scriven & Coryn, 2008). For a project already funded, the funding organization often needs to determine after a funding cycle whether the project is sufficiently good and important to continue or increase its funds. In trials, a court has to decide whether the accused is guilty or not guilty. In determinations of how to adjudicate drunk-driving charges, state or other government agencies set decision rules concerning the level of alcohol in a driver's blood that is legally acceptable. These examples are not just abstractions. They reflect true, frequent circumstances in society in which evaluations have to be definitive and decisive.

The problem of how to reach a just, defensible, clear-cut decision never has an easy solution. In a sense, most protocols for such precise evaluative determinations are arbitrary, but they are not necessarily capricious. Although many decision rules are set carefully in light of relevant research and experience or legislative processes, the rules are human constructions, and their precise requirements arguably could vary, especially over time. The arbitrariness of a cut score (for example, a score that classifies scores above it [the cut line] as good and those below it as unsatisfactory) is also apparent in different α (alpha) and β (beta) levels that investigators may invoke for determining statistical significance. Typically, α is set, by convention, at 0.05 or 0.01, but it might as easily be set at 0.06 or 0.02. In spite of the difficulties in setting and defending criterion levels, societal groups have devised workable procedures that more or less are reasonable and defensible for drawing definitive evaluative conclusions and making associated decisions. These procedures include applying courts' rules of evidence and engaging juries of peers to reach consensus on a defendant's guilt or innocence; setting levels for determining statistical significance and statistical power; using fingerprints and DNA testing to determine identity; rating institutions or consumer products; ranking job applicants or project proposals for funding; applying cut scores to students' achievement test results; polling constituents; grading school homework assignments; contrasting students' tested performance with national norms; appropriating and allocating available funds across competing services; and charging an authority figure with deciding, or engaging an expert panel to determine, a project's future. Although none of these procedures is beyond challenge, as a group they have addressed society's need for workable, defensible, nonarbitrary decision-making tools (also see Cizek & Bunch, 2007).

Some of these procedures have in common the advance setting of cut scores, standards, or decision rules. In the United States, for example, it is known in advance that all twelve

(or sometimes six) members of a jury must vote "guilty" for a defendant in a criminal trial to be found guilty beyond a reasonable doubt. Advance determinations of criteria and acceptable levels also apply to evaluations of new drugs; drunk-driving convictions; and certification of safe levels in water, air quality, food products, and bicycle helmets, for example.

When it is feasible and appropriate to set standards, criterion levels, or decision rules in advance, a general process can be followed to reach precise evaluative conclusions. The steps suggested here would be approximately as follows: (1) define the evaluand and its boundaries; (2) determine the key evaluation questions; (3) identify and define crucial criteria of goodness or acceptability; (4) determine as much as possible the rules for answering the key evaluation questions, such as cut scores and decision rubrics; (5) describe the evaluand's context, cultural circumstances, structure, operations, and outcomes; (6) take appropriate measurements related to the evaluative criteria; (7) thoughtfully examine and analyze the obtained measures and descriptive information; (8) follow a systematic, transparent, documented process to reach the needed evaluative conclusions; (9) subject the total evaluation to an independent assessment; and (10) confirm or modify the evaluative conclusions.

Although this process is intended to provide rationality, rigor, fairness, balance, and transparency in reaching evaluative conclusions, it rarely is applicable to most of the program evaluations treated in this book. This is so because often one cannot precisely define beforehand the appropriate standards and evaluative criteria, plus defensible levels of soundness for each one and for all as a group. So how do evaluators function when they have to make plans, identify criteria, and interpret outcomes without the benefit of advance decisions on these matters? There is no single answer to this question. More often than not, criteria and decision rules have to be determined along the way. We suggest that it is often best to address the issues in defining criteria through an ongoing, interactive approach to evaluation design, analysis, and interpretation and, especially, by including the systematic engagement of a representative range of stakeholders in the deliberative process.

What Are Performance Standards? How Should They Be Applied?

Often evaluation is characterized as comparing a performance to a standard (see also Fournier, 1995; Scriven, 1980). Constructing or setting performance standards is the process of setting one or more cut scores against which the performance of something is judged, with the cuts score(s) representing two or more states, conditions, or degrees of performance. Cut scores divide a distribution of performances into two or more discrete categories. So, for example, in the case where only a single cut score is set, its application results in the creation of only two possible performance categories, such as pass or fail (for example, for issuing a driver's license or a license to practice medicine or law). In some contexts, however, multiple cut scores may be required, the application of which results in the creation of more than two performance categories. Here, such cut scores are exemplified by the method used in a typical grading system (that is, A, B, C, D, or F) or that used for the National Assessment of Educational Progress (that is, advanced, proficient, or basic). Such classification methods can most easily be described as approaches that are either norm referenced or criterion referenced.

Norm-referenced methods include those whereby a distribution of a norm group's scores on a variable of interest is established and used to determine where the score of a separate evaluand places in the normative distribution, such as how far the obtained score is above or below the mean of the normative distribution's scores. Such norms-based conclusions typically are expressed in the evaluand's percentile rank against the normative distribution's table of scores. In contrast to norm-referenced methods of standard setting, and more commonly used, are criterion-referenced methods. With criterion-referenced methods of standard setting, performance does not depend on how well other objects perform. Criterion-referenced methods are also sometimes referred to as absolute methods, in contrast to the relativistic nature of norm-referenced methods.

The concept of criterion-referenced assessment is perhaps clearest in the judging of livestock, cats, and dogs, where associations of breeders publish the standards for particular breeds. Similarly, the sports of diving, gymnastics, and figure skating have published standards against which to judge performances by athletes. However, observers often view with disdain the lack of transparency, reliability, and validity of rendered judgments. The problems are even more acute in most standards-based program evaluations in which there are no juried, published standards for particular classes of programs. In such cases, evaluators and clients often have to concoct and agree on standards by which to judge particular programs.

Sometimes clients and their evaluators define behavioral objectives that, among other things, specify cut scores for distinguishing good performance from poor performance on each variable of interest. Many problems follow from this practice. The objectives are arbitrary and often unrealistic. They may not reflect the assessed needs of the intended beneficiaries. They may be more appropriate for average performers than for very high or low performers. For example, beneficiaries who already far exceed the cut score standard may find a disincentive for improvement in the program's low expectations. At the other end of the distribution, beneficiaries who are far below the standard may believe it is futile to attempt to reach the cut score standard, and may consequently give up. Also, cut score standards have a tendency to narrow a program's focus; lock it into predetermined objectives; and inhibit it from responding over time to emergent needs, developments, insights, and opportunities to exceed past performance.

An alternative to this narrow, preordinate approach to standards-based evaluation (*preordinate* being a coined term common in evaluation circles signifying rigid, advance stipulation of an evaluation's questions, standards for interpreting findings, and measurement and analysis procedures) is to view the evaluation process as a flexible, creative, evolving, responsive approach to assessing and supporting the client group's continuing quest for program improvement. W. Edwards Deming (see Walton, 1986) sold a similar notion to Japanese automobile manufacturers in the 1970s and helped spawn an amazing trend of continuing improvement in the quality of automobiles that eventually spread throughout the world. Deming's notion was not to attain and continue to achieve at any given level of quality, but continually to strive for better and better quality. Moreover, the recommended focus was on continuously improving the quality of manufacturing processes under the assumption (which proved true) that this

would result in both improved manufacturing processes and better and better outcomes. In the education field, W. L. Sanders and Horn (1994) argued similarly that the standard for educational programs should be continued growth and improvement for every student, whatever her or his prior level of achievement.

We believe it makes no sense to close the gap between high and low achievers, because sound education that helps all students reach their fullest potential will inevitably widen the achievement gap. This claim can be rejected only if one also rejects the claim that humans vary in abilities and capacities. To do the latter would require discarding society's huge store of evidence from research on individual differences.

Why Is It Appropriate to Consider Multiple Values?

Many evaluations face the challenge of multiple value perspectives. This is part and parcel of the world's increasingly pluralistic societies. Addressing competing and often conflicting values and cultures of different members of an evaluation audience is a necessary and difficult task in evaluations (also see Shadish, Cook, & Leviton, 1991). We would argue that it is the shared and differential needs of the consumers of a given service that should be ascertained as a basis for determining what information to collect and what standards to invoke in determining the worth of that service.

Sometimes an evaluator should address the value conflict issue by separately interpreting process and outcome information against the distinct sets of values or priorities held by different segments of the stakeholder population. Moreover, the evaluator might beneficially seek out and assess alternative programs or services to determine which ones best meet the needs of different stakeholder groups.

In planning evaluations, evaluators should deal directly with the important matter of choosing and applying pertinent values (also see Scriven, 1994b, 2007). They should determine what sets of values will be referenced in interpreting findings and sometimes in searching for and analyzing program options. Such determinations require evaluators to work within their basic philosophical convictions—that is, to act with integrity. Evaluators also should take into account a program's mission and the pertinent cultures, values, needs, and priorities of the program's leaders as well as impactees and other stakeholder groups. In issuing evaluative conclusions or putting forward assessments of alternative programs, evaluators should report the employed values and explain why they were chosen.

Addressing conflicting values is not an easy task for evaluators, if for no other reason than that they are not the sole arbiters of one set of values over another. Our advice is, first, never to take the side of one group rather than another's and, second, to take a dispassionate view of the needs of differing value groups and work toward the formulation of a sound set of guiding values that reflects integrity and the interests of the different parties to the evaluation. That being said, evaluators should not set aside their basic values, such as those concerning human rights. They should not proceed with an evaluation if doing so would aid and abet unethical or immoral decisions and actions. Clearly, an evaluator should decline an evaluation assignment if it is alien to his or her beliefs about what is sound, moral behavior.

Should Evaluations Be Comparative, Noncomparative, or Both?

Evaluators may focus on a single product or service or compare it with alternatives. Depending on the circumstances, an evaluation legitimately may be comparative or noncomparative. A main consideration is the nature of the audience and what evaluative information it needs. If the audience is composed of consumers who need to choose a product or service, the evaluation should be comparative and help consumers learn what alternatives are available and how they compare on critical criteria. If the audience includes developers or consumers who are already committed to the development or use of a given program, the evaluation might focus intensively on the workings of the program and help provide direction for improving it. Periodically, however, even if a group is firmly devoted to a certain service or product, it might get a better version from the provider of this service or product or find a better alternative by opening consideration to other providers, or by engaging in a systematic process of invention and innovation.

In general, we think that evaluations should be comparative before the purchase of a product or service or the beginning of a program, noncomparative during program development or use of a service, and periodically comparative after development or sustained use to open the way for improvements or better alternatives. Whether an evaluation should be comparative depends on the intended uses of the evaluation. If, for example, a selection is to be made from among alternative programs or uses of resources, then the evaluation should clearly be comparative.

How Should Evaluations Be Used?

We see four main uses of evaluations: improvement, accountability, dissemination, and enlightenment.

Formative Evaluations for Improvement

The first use is to provide information for developing a service, ensuring its quality, or improving it. Evaluations to serve this use typically are labeled formative evaluations (Scriven, 1967). Basically, they provide feedback for improvement. They are prospective and proactive. They are typically conducted during development of a program or its ongoing operation. Formative evaluations offer guidance to those who are responsible for ensuring and improving the program's quality and who should, in doing so, pay close attention to the nature and needs of the program's consumers. In formative evaluations, evaluators assess and assist with the formulation of goals and priorities, provide direction for planning by assessing alternative courses of action and draft plans, and guide program management by assessing implementation of plans and interim results.

Information from a formative evaluation is directed toward improving operations, especially those that are in the process of development. In the main, formative evaluations serve quality assurance purposes. In formative evaluations, the evaluator should interact closely with program staff and provide guidance for decision making. The evaluation plan needs to be

flexible and responsive. When the main aim is to improve an existing program, the evaluation should resemble a case study more than a comparative experiment. In fact, locked-in, controlled experiments that require random assignment of program participants to alternative program treatments and keeping treatments stable and unchanging typically prevent the evaluator from giving to program personnel the ongoing feedback for improvement that is the essence of formative evaluations.

Summative Evaluations for Accountability

The second main use of evaluations is to produce summative reports (Scriven, 1967). These are retrospective assessments of such evaluands as completed projects, established programs, finished products, or services rendered. Summative evaluations typically occur following development of a product, completion of a program, or end of a service cycle. They draw together and supplement previously collected information and provide an overall judgment of the evaluand's value. Summative evaluations are useful in ascertaining accountability for successes and failures, informing consumers about the quality and safety of products and services, and helping interested parties increase their understanding of the assessed phenomena. Summative evaluation reports are not aimed primarily at the development staff, but at the sponsor and consumers. The reports should convey a cumulative record of what was done and accomplished and an assessment of the evaluand's cost-effectiveness. Information derived from in-depth case studies and field tests is of interest to audience members in such situations. Field tests can involve productive use of comparative experiments. In the medical field, for example, results from double-blind studies comparing a newly developed treatment or other evaluand to a placebo or another competitive treatment can help potential users decide whether to use the new contribution. Whereas in general we argue against the use of experimental design in formative evaluations, it can be useful in some summative evaluations. This is especially the case in evaluations designed to undergird dissemination of a final product, service, program, project, or other evaluand. But even then, a randomized experiment is only part of a sound summative evaluation.

Relationship Between Formative and Summative Evaluations

Table 1.3 summarizes main features of formative evaluation and summative evaluation.

Both formative and summative evaluations are needed in developing and certifying evaluands, including programs, projects, products, or services; or, in the case of personnel, they help in developing potential and gauging the extent to which required criteria for certification, tenure, promotion, and the like are met. Too often, only summative evaluation is carried out—for judging an evaluand's past performance. This restricts development processes and may lead to inadequate or even incorrect conclusions. Subjecting a trainee nurse to an accountability assessment, for example, while ignoring the obvious advantages of fostering improvement through formative methodologies is foolish. Similarly, when a new model of an agricultural combine is being designed and developed, a lack of formative information covering cost, efficiency, reliability, safety, ease of use, durability, effectiveness, and potential marketing

Table 1.3 Formative Evaluation and Summative Evaluation

Descriptor	Formative Evaluation	Summative Evaluation
Purpose	Quality assurance; improvement	Providing an overall judgment of the evaluand
Use	Guiding decision making	Ascertaining accountability for successes and failures; promoting understanding of assessed phenomena
Functions	Provides feedback for improvement	Informs consumers about an evaluand's value (for example, its quality, cost, utility, competitive advantage, and safety)
Orientation	Prospective and proactive	Retrospective and retroactive
When conducted	During development or ongoing operations	After completion of development
Particular types of services	Assists with goal setting, planning, and management	Assists consumers in making wise decisions
Foci	Goals, alternative courses of action, plans, implementation of plans, interim results	Completed projects, established programs, or finished products; ultimate outcomes; costs; side effects
Variables	All aspects of an evolving, developing program	A comprehensive range of dimensions having to do with merit, worth, probity, safety, equity, and significance
Audience	Managers, staff; connected closely to insiders	Sponsors, consumers, and other interested stakeholders; projected especially to outsiders
Nature of evaluation plans	Flexible, emergent, responsive, interactive	Relatively fixed, not emergent or evolving
Typical methods	Case studies, observation, interviews (controlled experiments typically are inappropriate here)	A wide range of methods, including case studies, controlled experiments, and checklists
Nature of reports	Periodic, often relatively informal, responsive to client and staff requests	Containing a cumulative record and assessment of what was done and accomplished, a comparison between the evaluand and critical competitors, and a cost-effectiveness analysis
Relationship between formative evaluation and summative evaluation	Often forms the basis for and supplements summative evaluations	Involves compiling, assessing, and building on previously collected formative evaluative information

would be disastrous for the manufacturers. Evaluations delayed until the near completion of an employee's training and probationary period, or a product's development, or a project or program's implementation, or a service's full period of delivery may be too late to foster needed improvements and produce successful outcomes.

The relative emphases of formative and summative evaluations will change according to the nature of and circumstances surrounding the evaluand. In general, as portrayed in Figure 1.1, formative evaluation will be dominant in a program's early stages and less so as the program matures. Summative evaluation will take over as the program concludes and certainly will be used after it is completed. All concerned in these evaluations should have a clear understanding of when and in what circumstances formative evaluation may give way to summative evaluation. The conclusion should not be drawn, however, that all evaluations fall

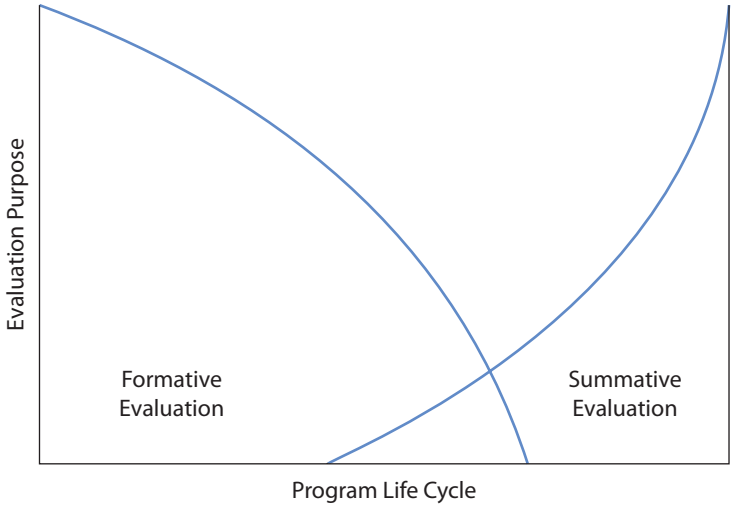

Figure 1.1 Relationship Between Program Life Cycle and Evaluation Purpose

into one or both categories. Many of the evaluation approaches depicted later in this book can be used for formative purposes, summative purposes, or both. Moreover, additional purposes (such as program monitoring) have been put forth, and have been widely debated (for example, Chen, 1996; Patton, 1996; Scriven, 1993, 1996; Wholey, 1996), yet we believe that formative and summative evaluations adequately reflect the large majority of work that evaluators engage in.

Stake (1969) made an interesting observation, apropos the relationship between formative and summative evaluations, that formative evaluations are closely connected to "insiders"—that is, program developers—whereas summative evaluations are of more interest to "outsiders"—that is, the potential users of the developing (or developed) programs. This does not assume that formative evaluations are necessarily undertaken by internal personnel or that summative evaluations are always conducted externally. A wide array of factors, such as timelines, finances, and the competency of personnel to undertake evaluations, will often determine whether evaluations, either formative or summative, are internal or external. The dominant question to be answered is whether the process and findings are credible. Ideally, internal summative evaluations are subjected to external audits (see Chapter 25 on metaevaluations).

Finally, formative evaluations often form the basis for summative evaluations. If this is to occur, both the evaluators and those who commission the studies must agree and also make clear to all involved that a formative evaluation will be conducted and used to form the basis for a subsequent summative evaluation. It should also be recognized that on occasion, the soundness and utility of a formative evaluation may be strengthened by the intervention of interim summative evaluations (usually carried out by external personnel) at critical points of a program's development. Such interplay between separately conducted formative and summative evaluations requires sound professional collaboration, which is a hallmark of good evaluation practice. (For other dimensions of this topic, see "Why Are Internal Evaluation Mechanisms Needed?" later in this chapter.)

Evaluations to Assist Dissemination Efforts

The third use of evaluations is to help developers disseminate proven practices or products and help consumers make wise adoption or purchasing decisions. Here the evaluator must critically compare the service or product with competitors. Perhaps the best example of evaluations aimed at serving dissemination and informing adoption decisions are those found in *Consumer Reports*. Each issue of this well-known monthly magazine provides independent evaluations of alternatives for consumer products and services: automobiles, insurance policies, mortgages, breakfast cereals, chain saws, refrigerators, computers, cameras, cell phones, restaurant chains, supermarket chains, hotel chains, and house paints, to name just a few. The unique feature of evaluations for dissemination is their focus on questions of practical interest to consumers. In Parts Two and Three, we describe Michael Scriven's consumer-oriented evaluation approach, which is predominately premised on a product model of evaluation.

A more recent example is the numerous evidence-based repositories, including the Campbell Collaboration, Cochrane Collaboration, and What Works Clearinghouse, among many others, that have become increasingly commonplace in the last few decades as mechanisms for using evaluation results to disseminate effective practices or products (Coryn, Tarsilla, & Hobson, 2010). These repositories largely provide information about the results of meta-analyses and randomized controlled trials in health and medicine, human and social services, and education. In part, these repositories emerged due to the climate of increasingly scarce resources and greater demands for accountability, in which policymakers and those in practice-based disciplines and professions have been seeking high-quality, nonarbitrary, and defensible evidence for formulating, endorsing, and, occasionally, enforcing best policies and practices (Flay et al., 2005).

Evaluations to Foster Enlightenment

The fourth use of evaluations is to foster enlightenment, or new understandings arising from evaluations (also see Chelimsky, 1997; Patton 1997, 2008). Basically, evaluation and research are different enterprises. Evaluators attempt to consider all criteria that apply in determining value, whereas researchers may be restricted to the study of selected variables that are of interest in testing theory, diagnosing problems, or answering particular questions. Evaluations typically involve subjective approaches and are not as tightly controlled and subject to manipulation as is the typical research investigation. However, efforts over a period of time to evaluate a program or set of similar programs may produce information of use in evolving and testing theory. Certainly evaluation results often should and do lead to focused, applied research efforts and sometimes to development of institutional or social policies. Hence, we believe that in planning studies, evaluators should consider how their findings might contribute to new insights in matters of interest to theorists, policymakers, and scientists, especially through formulation of testable hypotheses. With some forethought, careful planning, and appropriate budgeting, evaluations may serve not only to guide operating programs, sum up and assess their contributions, and lead to the dissemination of effective products and services but also to address particular research, theory, or policy questions.

Why Is It Important to Distinguish Between Informal Evaluation and Formal Evaluation?

By this point, it should be clear that program evaluation is a demanding field of practice. At the same time, everybody essentially evaluates constantly, whether making choices about the trivial or the critical (see also Posavac & Carey, 2003). We believe it is important to distinguish formal evaluation from informal evaluation. In fact, the distinction is at the root of the need for and emergence of the evaluation profession. Just as most individuals employ home remedies and over-the-counter medications in addressing their minor ailments, almost everybody recognizes that some health issues require diagnosis and treatment by competent physicians in accordance with the standards of the medical profession. Similarly, many evaluations can and must be conducted on an informal basis, whereas others require a rigorous, systematic approach, often including an independent perspective.

Informal Evaluations

Everybody performs informal evaluations whenever judging and making decisions about the things observed, thought about, interacted with, or being considered for purchase. For example, we do this when purchasing food, cars, tools, refrigerators, computers, computer programs, stocks, correspondence courses, insurance policies, or termite protection services. Depending on the nature of the evaluand, one might look for options, read labels, consult friends who have pertinent experience, form a committee or task group to deliberate on the evaluative questions of interest, call the Better Business Bureau, consult other consumer information sources, search the Internet for consumers' assessments of items purchased, or try out something before deciding to keep it. These are all good and appropriate evaluative moves and fit within our general conception of informal evaluation. The conduct of informal evaluations, however, is prone to haphazard data collection, crediting and using propaganda and other forms of misinformation, errors of judgment, strong influence by salespersons, acting on old preferences or prejudices, relying on out-of-date information, depending on an inadequate or biased sample of customer feedback, or making expedient choices. In many cases, the steps in an informal evaluation are unsystematic, lacking in rigor, and based on biased perspectives. Thus, informal evaluations typically offer a weak basis for convincing decision makers and others of the validity of evaluation findings and the appropriateness of ensuing conclusions and recommendations. We can get by with weak, informal evaluations when only we have to pay the price and abide by the consequences. Better, more formal evaluations are called for when there is a need to inform critically important decisions, especially ones that will affect many people, require substantial expenditures, or pose substantial risk.

Formal Evaluations

In accordance with the definition of evaluation given earlier, formal evaluations should be systematic and rigorous. By the term *systematic,* we refer to evaluations that are relevant, designed and executed to control bias, kept consistent with appropriate professional standards,

documented, reported to right-to-know audiences, and otherwise made useful and defensible. Especially, we define formal evaluations as ones that are held up to scrutiny against appropriate standards of the evaluation profession. The kind of formal evaluation we are promoting requires systematic effort by one or more persons who have the requisite evaluation competencies. We do not disparage the informal evaluations that are part and parcel of everybody's daily life, any more than we would advise people not to make prudent use of home remedies and over-the-counter medications. Moreover, not all formal evaluations need to be conducted by outside evaluation experts. What is required is that those conducting the evaluation meet the standards of the evaluation field. In Chapter 3 we summarize professionally developed guiding principles for evaluators (AEA, 2004); professional standards for program evaluations (Joint Committee, 1981, 1994, 2011); and the U.S. government auditing standards (U.S. General Accounting Office, 2002; U.S. Government Accountability Office, 2003, 2007). Building on these, this book is designed to help evaluation and research students, practicing evaluators, evaluation clients, research methodologists, and other interested parties attain the perspectives and basic level of proficiency required to undertake defensible formal evaluations grounded in professional standards and principles for practice.

How Do Service Organizations Meet Requirements for Public Accountability?

We cannot stress too strongly that society is dependent on sound evaluations to obtain safe, high-quality goods and services from a wide range of professionals and the organizations in which they work. Such organizations include school districts, universities, research centers, hospitals, government departments, charitable foundations, churches, community service organizations, and others. Any operatives should deliver services that are of high quality, up to date, safe, efficient, fairly priced, honest, and generally in the public interest. To meet accountability requirements, each profession, public service area, and society should regularly subject services to formal evaluations. Some evaluation work is appropriately directed at regulation and protection of the public interest. This work should be conducted by independent bodies, including government agencies, accrediting boards, and external evaluators. Equally important are the formative and summative evaluations of services that professionals and other service providers and their organizations themselves conduct. These internal or self-evaluations are an important aid to continually scrutinizing and improving services and also supplying data needed by independent or external evaluators.

Accreditation

A wide range of accrediting organizations periodically assess the performance of member organizations against formally established standards. Typical accreditation evaluations are grounded in clear accreditation criteria and guidelines for self-assessment. In the accreditation process, the institution or program to be evaluated proceeds by conducting a lengthy process of self-assessment, typically lasting at least a year. A team of external evaluators, appointed

by the accrediting organization, then reviews the self-report, conducts a site visit, and writes an independent evaluation report. The accrediting organization subsequently uses the report to make decisions on whether, to what extent, and for what period the subject institution or program is to be accredited and submits its report to the institution or program. Typically accreditation is awarded for a finite period, such as five years. The accrediting body then updates its publicly available list of accredited institutions or programs. In some cases, provisional accreditation is provided pending corrective actions by the assessed institution or program. A prime accreditation criterion often is that the subject institution or program will operate an internal evaluation mechanism and make use of its findings.

Why Are Internal Evaluation Mechanisms Needed?

Some large school districts, medical schools, foundations, and government agencies maintain well-funded and adequately staffed evaluation offices, and their evaluators have succeeded in helping their institutions be accountable to constituents, obtain guidance for planning and administering their services, win grants and contracts, and meet requirements of accrediting organizations or other oversight bodies. To keep their services up to date and ensure that they are effectively and safely meeting their clients' needs, service institutions and programs should continually obtain pertinent evaluative feedback. This process includes studying the outcome and treatment needs of their clients; evaluating relevant approaches that are being proposed or used elsewhere; evaluating the performance of personnel; closely monitoring and assessing the delivery of services; assessing immediate and long-term outcomes; and searching for ways to make services more efficient, effective, and safe. Conducting such internal evaluations is a challenging task. The credibility of internal evaluation is enhanced when it is subjected periodically to metaevaluation (Scriven, 1969b; Stufflebeam, 1978, 2001c; also see Chapter 25), in which an independent evaluator evaluates and reports publicly on the quality of internal evaluation work. Such independent metaevaluation also provides direction for strengthening the internal evaluation services. Optimally, metaevaluations are both formative and summative.

Chapter 26 provides in-depth information on how organizations may institutionalize and mainstream a systematic process of internal evaluation.

Why Is Evaluation a Personal as Well as an Institutional Responsibility?

Even if an organization has a strong internal evaluation unit, every professional in the organization needs to engage in systematic evaluation. There is no escaping the fact that evaluation is a personal as well as an organizational responsibility. Offices of evaluation and accrediting firms can help organizations meet their major responsibilities in regard to continuous evaluation and accountability. An office of evaluation can also provide an organization's staff with in-service training and technical support in evaluation. However, all professionals bear responsibility for formally evaluating their own performance. It is in their interest to do so, because evaluation is an essential means of finding out and acting on what is going right and wrong. Moreover, conducting and acting on sound evaluation constitute a fundamental part of what it means to be a professional—a member of an established profession

who continually works to deliver better services. We hope this book will both inspire and assist individual professionals and other service providers as well as evaluation students and specialists, enabling them to develop evaluation competencies and effectively carry out systematic evaluations.

What Are the Methods of Formal Evaluation?

One aspect that distinguishes formal evaluation from informal evaluation is the area of methodology. When we move our consideration away from evaluations that involve quick, intuitive judgments toward those that entail rigorously gathered findings and effective communication, we must necessarily deal with the complex areas of epistemology, rules of evidence, information sciences, research design, measurement, statistics, communication, and some others. Many principles, tools, and strategies within these areas are pertinent to systematic evaluation. The well-prepared evaluator will have a good command of concepts and techniques in all these areas and will stay informed about potentially useful technological developments. Evaluators who would exert leadership and help advance their profession should contribute to the critiquing of existing methods and the development of new ones.

Over the years, many evaluators have exclusively chosen and used—even championed—a narrow set of techniques. Some evaluators have equated evaluation with their favorite methods—for example, experimental design, standardized testing, surveying, case studies, site visits by teams of experts, or participant observation. Other leaders have sharply attacked narrow views of which methods are appropriate and argued for a broader, more eclectic approach, which is where we find ourselves (also see Mark, Henry, & Julnes, 2000). A key point in the latter position is that the use of multiple methods and perspectives enhances the dependability of inferences and conclusions and yields appropriate levels of circumspection.

We believe that evaluators should know about a wide range of pertinent techniques and how well they apply in different evaluation contexts. Then, in each evaluation, they can assess which techniques are potentially applicable and which most likely would work best and in combination to serve the given study's particular purposes. Among the technical areas in which we think the professional evaluator should be proficient are proposal writing, research design, budgeting, contracting, scheduling, system analysis, logic models, interviewing, focus groups, survey research, case studies, content analysis, observation, checklists, goal-free evaluation, advocate teams, test construction, rating scales, database development and management, statistical analysis, cost analysis, technical writing, and project administration.

What Is the Evaluation Profession, and How Strong Is It?

The formal profession of evaluation emerged only during the last third of the twentieth century. In so short a time period, this young profession has made remarkable progress, but it still has far to go. The evaluation field now has national and state professional societies of evaluators; annual conventions; a substantial literature and knowledge base, including numerous professional journals (also see Coryn, 2007) and a wide range of theoretical and

technical books; specialized Web sites; discussion groups, blogs, and listservs (also see Christie & Azzam, 2004); master's and doctoral programs (also see LaVelle & Donaldson, 2010); institutes and workshops on specialized evaluation topics (for example, the Evaluators' Institute, which presents annual training sessions in specific evaluation procedures); client organizations that fund a wide range of evaluations; evaluation companies; guiding principles for evaluators; and standards for program, personnel, and student evaluations. These are substantial gains considering the field's status in 1964, when it had none of these elements. The evaluation field is still immature, however, when compared with established professions, such as medicine, law, engineering, and accounting, and other service areas, such as those of master plumbers, licensed electricians, and dental hygienists. In particular, the evaluation field lacks some of the hallmarks of a mature profession. For example, membership in AEA is open to anyone regardless of training and expertise in evaluation. Furthermore, the field has no mechanisms for certifying or licensing competent evaluators (also see S. C. Jones & Worthen, 1999; Worthen, 1999), although the Canadian Evaluation Society began a credentialing effort for evaluators in 2010. Despite the field's substantial progress, clients of evaluation have no formal means of determining which self-proclaimed evaluators have been certified as competent. And even though evaluations are widely recognized as essential to the health of any organization, acceptance of tertiary training to gain qualifications as an evaluator is lagging worldwide. The evaluation field's stature and credibility are threatened by its lack of professional certification and quality control. This is especially so because there is "much gold in the evaluation hills," and because, in our experience, all too often ill-prepared evaluators obtain high-cost contracts to conduct evaluation assignments for which they lack the needed expertise.

What Are the Main Historical Milestones in the Evaluation Field's Development?

The evaluation field has evidenced only modest efforts to systematically record and analyze its history (for example, Shadish & Luellen, 2005). Any profession, to effectively serve its clients, must evolve in response to changing societal needs and in consideration of theoretical and technical advancements. Unless the members of a profession develop and maintain a historical perspective on their work, they are likely to persevere in using a stagnant conception of their role, not to remember valuable lessons of the past, not to stimulate and contribute to innovation in their field, and all too frequently to return to deficient methods of the past. It has been said often that those who do not learn from their history are doomed to repeat it.

In this section we focus on the history of the program evaluation field, especially as evaluation theory and practice have evolved in the area of education (Stufflebeam, Madaus, & Kellaghan, 2000).[3] We believe this is appropriate and will be instructive, because the profession of evaluation developed earliest and most heavily within the field of education. We provide only a brief historical sketch, noting the most significant developments in educational program evaluation.

Our historical analysis is grounded in the seminal work of Ralph W. Tyler (described later in this book), who is often spoken of as the father of educational evaluation. Using his

initial contributions as the main reference point, we have identified six major periods: (1) the Pre-Tylerian Period, which includes developments before 1930; (2) the Tylerian Age, which spans 1930 to 1945; (3) the Age of Innocence, which runs from 1946 to 1957; (4) the Age of Realism, which covers years 1958 through 1972; (5) the Age of Professionalism, which includes developments from 1973 to 2004; and (6) the Age of Global and Multidisciplinary Expansion, from 2005 to the present.

The Pre-Tylerian Period: Developments Before 1930

Systematic evaluation was not unknown before 1930, but it was not a recognizable movement. In the mid-1840s in the United States, the common method of assessing student learning and the quality of instruction was an annual oral examination conducted by school committees. Because of a desire for more dependable inspections of schools, in 1845 Boston replaced the oral exams with the first systematic school survey using printed tests. Horace Mann championed this approach and advised Boston to base school policies on factual results from testing the eldest class in each of the city's nineteen schools. The committee running the survey faced problems similar to those seen in today's large testing programs. In particular, teachers felt threatened because they knew that their students' test scores would be viewed as an indicator of their teaching competence.

The initial tests reflected the curriculum of the day, mainly requiring abstract renderings consistent with the prevalent Puritan philosophy. They were chalk-and-slate or quill-and-paper tests, requiring students mainly to recall facts but, in a minor way, also to demonstrate application of what they had learned. Members of the school committees administered the tests during six hours over two days. Test results overall were discouraging. Reports contained a brief, often negative evaluative statement about each school. Mann saw these new methods of inspecting schools as impartial, thorough, and accurate in assessing what pupils had been taught and lauded their use in arriving at independent judgments of schools. In today's language, we could say he judged the new evaluation approach as meeting conditions of objectivity, validity, and reliability. Although the Boston survey spawned similar examination projects elsewhere in the United States, it was not until the end of the nineteenth century that end-of-semester printed tests became a common feature in schools nationwide.

It is generally recognized that Joseph Rice conducted the first formal educational program evaluation in the United States. An education reformer who provided educational administrators in New York City with leadership, Rice in 1895 launched the most ambitious plan ever undertaken to collect data on education. His goal was to confirm that student learning was deficient. Over the next decade, he obtained test scores in spelling and mathematics from about sixteen thousand students. A key finding was that the amount of time spent on spelling each day related little to spelling achievement. The Boston and Rice surveys gave publicity to the survey technique as a means of collecting and analyzing data to help identify and correct deficiencies in the schools and form sound educational policies. Its use in the twentieth century was evident in the 1915 publication of the Cleveland Education Survey. Sponsored by the Survey Committee of the Cleveland Foundation, the twenty-five-volume report assessed every

aspect of the school system and was heralded as the most comprehensive study of an entire school system ever completed.

The dawning of the twentieth century saw the emergence of yet another approach to evaluation. In applying the concepts of efficiency and standardization to manufacturing, Frederick Taylor had found standardization to contribute to efficiency and assurance of consistent quality in manufactured products. Taylor's success in manufacturing influenced leaders in education to seek standardization and efficiency in schools. Consequently, under the leadership of Edward Thorndike and others, educators launched the now massive enterprise of standardized testing. They believed that standardized tests could check the effectiveness of education and thereby show the way to more efficient student learning. Technology for measuring student achievement and other human characteristics developed strongly in the United States, Great Britain, and some other countries throughout the twentieth century, and, in this century, continues to be developed and widely applied. Educators and the public have often looked to scores from standardized tests as a basis for judging schools, programs, teachers, and students. Nevertheless, perhaps no other educational practice has generated so much criticism and controversy as has standardized testing, especially when high stakes have been attached to the results (American Evaluation Association Task Force on High Stakes Testing, 2002).

As a countermovement to rigid testing practices, a progressive education movement developed during the 1920s that espoused the ideas of John Dewey and even earlier writers. Travers (1983) stated the matter extremely well:

> Those engaged in the progressive education movement viewed the new emphasis on standardized achievement testing as a menace to everything they hoped to accomplish. They wanted to make radical changes in the curriculum, but the standardized tests tended to encourage the retention of the established curriculum content. They wanted to emphasize the development of thinking skills, but the tests placed emphasis on the memorization of facts. They wanted to emphasize self-evaluation, with the child's own evaluation of himself as the point from which progress should be measured, but the achievement testers encouraged a competitive system in which a child was judged in terms of his position in a group. The use of criterion-referenced tests was minimal in the 1920s and 1930s, and although such tests would have answered this last criticism of the progressive educators, it would not have resolved even a small fraction of the misgivings that the progressives had about the new achievement testing. (p. 144)

Despite a continuing flow of criticism, the use of objective achievement tests has continued to expand. The limitations of tests in measuring important educational outcomes, such as abilities to understand, apply, and critique, often are discounted in favor of obtaining quick and easy measures. In the service of educational evaluation, large-scale testing programs have been extremely expensive. We also judge them as grossly inadequate for assessing programs and institutions on merit, worth, probity, feasibility, significance, safety, and equity.

Objective testing can play a useful role in educational program evaluations, but it can provide only a small part of the needed information.

Although program evaluation has only recently been identified as a field of professional practice, this account illustrates that systematic program evaluation is not a completely recent phenomenon. Some of the modern evaluation work (testing commissions, surveys, accreditation, and experimental comparison of competitors) continues to draw from ideas and techniques that were applied long ago.

The Tylerian Age: 1930 to 1945

In the early 1930s Tyler coined the term *educational evaluation* and published a broad and innovative view of both curriculum and evaluation. Over about fifteen years, he developed his ideas until they constituted an approach that provided a clear-cut alternative to other views (Madaus, 2004; Madaus & Stufflebeam, 1988).

What mainly distinguished his approach was its concentration on clearly stated objectives. In fact, he defined evaluation as determining whether objectives have been achieved. In light of this definition, evaluators were supposed to help curriculum developers clarify the student behaviors that were to be produced through the implementation of a curriculum. The resulting behavioral objectives were then to provide the basis for both curriculum and test development. Curriculum design was thus influenced away from the content to be taught and toward the student behaviors to be developed. The technology of test development was to be expanded to provide for tests and other assessment exercises referenced to objectives as well as those referenced to individual differences and national or state norms.

During the 1930s the United States, as well as the rest of the world, was in the depths of the Great Depression. Schools and other public institutions had stagnated from a lack of resources and optimism. Just as Franklin Roosevelt tried to lead the American economy out of this abyss through his New Deal program, Dewey and others tried to help education become a dynamic, innovative, and self-renewing system. Called progressive education, this movement reflected the philosophy of pragmatism and employed the tools of behavioral psychology.

Tyler was drawn into this movement when he was commissioned to direct the research component of the now famous Eight-Year Study (E. R. Smith & Tyler, 1942), which was designed to examine the effectiveness of certain innovative curricula and teaching strategies being employed in thirty schools throughout the United States. The study is noteworthy because it helped Tyler at once expand, test, and demonstrate his conception of educational evaluation.

Through this nationally visible study, Tyler was able to publicize what he saw as clear-cut advantages of his approach over others. Because Tylerian evaluation involves internal comparisons of outcomes with objectives, it does not require costly and disruptive comparisons between experimental and control groups. The approach concentrates on direct measures of achievement, as opposed to indirect approaches that measure such inputs as quality of teaching, number of books in the library, extent of materials, and community involvement. Tylerian evaluations need not be heavily concerned with reliability of differences between the scores

of individual students, and they typically cover a wider range of outcome variables than those covered by norm-referenced tests. Tyler's arguments were well received throughout American education, and by the mid-1940s Tyler had set the stage for exerting a heavy influence on how educators and other program evaluators viewed evaluation for the next twenty-five years.

The Age of Innocence: 1946 to 1957

In the ensuing years, Tyler's recommendations were more discussed than applied. Throughout American society, the late 1940s and 1950s were a time to forget the war, leave the depression behind, build and expand capabilities, acquire resources, and engineer and enjoy a good life. We might have called this era the Period of Expansion, except that there was also widespread complacence in regard to serious societal problems. We therefore think this time is better referred to as the Age of Innocence, or even as the Age of Social Apathy.

More to the point of educational evaluation, there was expansion of educational offerings, personnel, and facilities. New buildings were erected. New kinds of educational institutions, such as community colleges, emerged. Small school districts consolidated with others to provide the wide range of educational services that were common in larger school systems: mental and physical health services, guidance, food services, music instruction, expanded sports programs, business and technical education, and community education. Enrollment in teacher education programs ballooned, and college enrollment generally increased dramatically.

This general scene in society and education was reflected in educational evaluation. Although there was great expansion of education, society had no particular interest in holding educators accountable, identifying and addressing the needs of the underprivileged, or identifying and solving problems in the U.S. education system. Although educators wrote about evaluation and collected considerable data, they seem not to have related these efforts to attempts to improve educational services. This lack of a mission carried over into the development of the technical aspects of evaluation as well. There was considerable expansion of tools and strategies for applying the various approaches to evaluation: testing, comparative experimentation, operationalizing objectives, and comparing outcomes and objectives. As a consequence, educators were provided with new tests and test scoring services, algorithms for writing behavioral objectives, taxonomies of objectives, new experimental designs, and new statistical procedures for analyzing educational data. But these contributions were not derived from any analysis of what information was needed to assess and improve education, and they were not an outgrowth of school-based experience.

During this period, educational evaluations were, as they had been previously, primarily the purview of local school districts. Schools could do evaluation or not, depending on local interest and expertise. Federal and state agencies had not yet become deeply involved in the evaluation of programs. Funds for evaluations came from local coffers, foundations, or professional organizations. This lack of external pressure and dearth of support for evaluations at all levels of education would end with the arrival of the next period in the history of evaluation.

The Age of Realism: 1958 to 1972

The Age of Innocence in evaluation came to an abrupt end in the late 1950s and early 1960s with the call for evaluations of large-scale curriculum development projects funded by federal monies. Educators would find during this period that they no longer could do or not do evaluations as they pleased, and that further development of evaluation methodologies would have to be grounded in concern for accountability, usability, and relevance. Their rude awakening during this period would mark the end of an era of complacency and help launch profound changes, guided by the public interest and dependent on taxpayer monies for support, which would help evaluation expand as an industry and into a profession.

The federal government responded to the Russian launch of Sputnik I in 1957 by enacting the National Defense Education Act of 1958. Among other things, this act provided for new educational programs in mathematics, science, and foreign languages and expanded counseling and guidance services and testing programs in school districts. A number of new national curriculum development projects, especially in science and mathematics, were established. Eventually funds were allocated to evaluate these programs.

Four approaches to evaluation were represented in the evaluations done during this period. First, the Tylerian approach was used to help define objectives for the new curricula and to assess the degree to which the objectives were later realized. Second, new nationally standardized tests were developed to better reflect the objectives and content of the new curricula and to begin monitoring the educational progress of the nation's youth (L. V. Jones, 2003). Third, the professional judgment approach typically engaged experts to rate proposals and make periodic site visits to check on the efforts of contractors. Finally, many evaluators studied curriculum development efforts through the use of controlled field experiments.

In the early 1960s some leaders in educational evaluation realized that their work and their results were not particularly helpful to curriculum developers or responsive to the questions about the programs being raised by those who wanted to assess their effectiveness. The "best and the brightest" of the educational evaluation community were involved in these efforts to evaluate the new curricula; they were adequately financed, and they carefully applied the technology that had been developed during the past decade or more. Despite all this, they began to recognize that their efforts were not succeeding.

This negative assessment was well reflected in a landmark article by the educational psychologist Lee Cronbach (1963). In looking at the evaluation efforts of the recent past, he sharply criticized the guiding conceptualizations of evaluation for their lack of relevance and utility and advised evaluators to turn away from their penchant for evaluations based on comparisons of the norm-referenced test scores of experimental and control groups. Cronbach counseled evaluators to reconceptualize evaluation not in terms of a horse race between competing programs, but instead as a process of gathering and reporting information that could help guide curriculum development. Cronbach argued that analysis and reporting of test item scores would be likely to prove more useful to teachers than the reporting of average total scores. Initially, Cronbach's counsel and recommendations went largely unnoticed except by a small circle of evaluation specialists. Nonetheless, his article was seminal, containing

hypotheses about new approaches to conceptualizing and conducting evaluations that were to be developed and tested within a few years.

The War on Poverty was launched in 1965. It was grounded in the previous pioneering work of Senator Hubert Humphrey and the charismatic leadership of President John F. Kennedy before his untimely death in 1963. President Lyndon Johnson subsequently picked up the reins and used his great political skill to get this landmark legislation passed. Its programs poured billions of dollars into reforms aimed at equalizing and upgrading opportunities for all U.S. citizens across a broad array of health, social, and educational services. The expanding economy enabled the federal government to finance these programs, and there was widespread support throughout the nation for developing what President Johnson termed the Great Society. Accompanying this massive effort to help those in need was a concern in some quarters that the investments might be wasted if appropriate accountability requirements were not imposed.

In response to this concern, Senator Robert Kennedy and some of his colleagues in Congress amended the Elementary and Secondary Education Act of 1965 to include specific evaluation requirements. As a result, Title I of that act (aimed at providing compensatory education to disadvantaged children) specifically required each school district receiving funds under this title to evaluate Title I projects annually using appropriate standardized test data and thereby to assess the extent to which the projects had achieved their objectives.

This requirement, with its specific reference to standardized test data and an assessment of congruence between outcomes and objectives, reflects the state of the art in educational evaluation at that time, which was based largely on the use of standardized educational achievement tests and superficially on Tyler's objectives-based approach. More important, the requirement forced educators to move their concern for educational evaluation from the realm of theory and supposition into the realm of practice and implementation. When school districts began to respond to the evaluation requirements of Title I, they quickly found that the existing concepts, tools, and strategies employed by their evaluators were largely inappropriate for the task.

Available standardized tests had been designed to rank-order students of average ability; they were of little use in diagnosing needs and assessing the gains of disadvantaged children whose educational development lagged far behind that of their middle-class peers. Furthermore, these tests were found to be relatively insensitive to differences between schools and programs, mainly because of their psychometric properties and content coverage. Instead of being measures of outcomes directly relating to a school or a particular program, these tests were at best indirect indicators of learning, measuring much the same traits as general ability tests (Kellaghan, Madaus, & Airasian, 1982).

The use of standardized tests entailed another problem, because it conflicted with the precepts of the Tylerian approach. Because Tyler recognized and encouraged differences in objectives from locale to locale, this model became difficult to adapt to nationwide standardized testing programs. To be commercially viable, these standardized testing programs had to overlook, to some extent, objectives stressed by particular locales in favor of objectives stressed in the majority of districts.

Also, the Tylerian rationale itself proved inadequate to the evaluation task. There was insufficient information about the needs and achievement levels of disadvantaged children to

guide teachers in developing meaningful behavioral objectives for this population of learners. In retrospect, the enormous investment school districts across the United States made in training and leading educators to write behavioral objectives was largely unsuccessful and a waste of much time and money. Typically educators learned how to meet the technical requirements of good behavioral objectives. However, these technically sound statements of objectives often proved to be of little practical use in that they did not reflect empirical assessments of the needs and problems of the students to be served. When the teachers actually met their students, they often found it prudent to set aside as irrelevant the objectives that had been so carefully prepared in advance of the project.

Attempts to isolate the effects of Title I projects through the use of experimental and control group designs also failed. Typically such studies showed "no significant differences" in achievement between treated Title I students and comparison groups. This approach was widely tried but was doomed not to succeed. Title I evaluators could not begin to meet the assumptions required by experimental designs. For example, they usually could not, in a timely manner, obtain valid measures; could not hold treatments constant during the study period; and legally could not randomly assign Title I (disadvantaged) students to control and experimental groups. When the finding of no results was reported, as was generally the case, there was little information on what the treatment was supposed to be and often no data on the degree to which it had in fact been implemented. Also, the emphasis on pre- and posttest scores diverted attention from consideration of the treatment or of treatment implementation. This hugely expensive experiment in testing the utility and feasibility of experimental design evaluations in the Title I program demonstrated rather decisively that this technique is not amenable to evaluating highly dynamic, field-based, generalized assistance programs, especially in the course of such programs' development.

As a result of growing disquiet concerning evaluation efforts and consistently negative findings, Phi Delta Kappa set up the National Study Committee on Evaluation (Stufflebeam et al., 1971). After surveying the scene, this committee concluded that educational evaluation was seized with a great illness and called for the development of new theories and methods of evaluation as well as for new training programs for evaluators. This committee's indictment of educational evaluation practice was consistent with a study of government-sponsored evaluations by Guba (1966) and an analysis of the Title I evaluation efforts by Stufflebeam (1966b).

At the same time, many new conceptualizations of evaluation began to emerge. Provus (1969), Hammond (1967), Eisner (1975), and Metfessel and Michael (1967) proposed reformulations of the Tylerian model. R. Glaser (1963), R. W. Tyler (1967), and Popham (1971) pointed to criterion-referenced testing as an alternative to norm-referenced testing. D. L. Cook (1966) called for the use of system analysis techniques to evaluate programs. Scriven (1967, 1974); Stufflebeam (1967); Stufflebeam et al. (1971); and Stake (1967) introduced new models for evaluation that departed radically from prior approaches. These conceptualizations stemmed from recognition of the need to evaluate goals, look at inputs, examine implementation and delivery of services, and measure intended as well as unintended program outcomes. Developers of these new approaches also emphasized the need to make (or collect) judgments about the merit and/or worth of the object being evaluated.

The late 1960s and early 1970s were vibrant with descriptions, discussions, and debates concerning how evaluation should be conceived. The chapters in Part Three deal in depth with the alternative approaches that began to take shape during this period. Lessons had been learned, often by uneasy experience.

The Age of Professionalism: 1973 to 2004

Beginning in about 1973, the field of evaluation began to crystallize and emerge as a distinct profession in its own right—related to, but quite distinct from, its forerunners of research and testing. The field of evaluation has advanced considerably as a profession, yet it is instructive to consider the development in the Age of Professionalism in the context of the field in the previous period.

In the late 1960s and early 1970s, evaluators faced an identity crisis. They were uncertain of their role—whether they should be researchers, testers, reformers, administrators, teachers, consultants, or philosophers. What special qualifications, if any, they should possess was unclear. There were no professional organizations dedicated to evaluation as a field, nor were there specialized journals through which evaluators could exchange information about their work. Essentially no literature about evaluation existed, except for unpublished papers that circulated through a small underground network of scholars. There was a paucity of pre-service and in-service training opportunities in evaluation. Articulated standards of good practice were confined to educational and psychological tests. The field of evaluation was amorphous and fragmented. Many evaluations had been conducted by untrained personnel or research methodologists who had tried unsuccessfully to fit their experimental methods to evaluations (Guba, 1966). Evaluation studies were fraught with confusion, anxiety, and animosity. Evaluation as a field had little stature and no political clout.

Against this backdrop, the progress made by evaluators in professionalizing their field beginning in the 1970s is quite remarkable. Many universities now offer at least one course in evaluation methodology (as distinct from research methodology). A few—including the University of Illinois, the University of California at Los Angeles, the University of Minnesota, the University of Virginia, Claremont Graduate University, and Western Michigan University—have developed graduate programs in evaluation (LaVelle & Donaldson, 2010). Even so, the Western Michigan University program is the world's only interdisciplinary doctoral program in evaluation (Coryn, Stufflebeam, et al., 2010).

Increasingly, the field has looked to metaevaluation (Scriven, 1975; Stufflebeam, 1978, 2001c) as a means of ensuring and checking the quality of evaluations. In 1981 the Joint Committee issued standards for judging evaluations of educational programs, projects, and materials and established a mechanism by which to review and revise the standards and assist evaluators in using them. This review process has worked effectively, leading the Joint Committee to produce the second edition of *The Program Evaluation Standards* in 1994 and the third edition in 2011.[4] Moreover, publication of the Joint Committee's first edition of *The Personnel Evaluation Standards* in 1988, followed by the second edition in 2009, signaled advancement in methods for assessing systems for evaluating personnel. In addition, the Joint Committee

released *The Student Evaluation Standards* in 2003. Several other sets of standards with relevance for evaluation also have been published, the most important being AEA's *Guiding Principles for Evaluators* (2004) and the U.S. Government Accountability Office's *Government Auditing Standards* (U.S. General Accounting Office, 2002; U.S. Government Accountability Office, 2003, 2007). Many new techniques and methodological approaches have been introduced for evaluating programs, as described in Part Four of this book. The most comprehensive treatment of the state of the art in educational evaluation so far is the *International Handbook of Educational Evaluation* (Kellaghan & Stufflebeam, 2003).

The Age of Global and Multidisciplinary Expansion: 2005 to the Present

When the first edition of this book was being completed in 2006 (Stufflebeam & Shinkfield, 2007), it was realized that the evaluation field had already entered a new, recognizable age. Here, we label it the Age of Global and Multidisciplinary Expansion and arbitrarily have set its beginning as about 2005. As noted earlier, there are now over fifty professional evaluation societies in countries throughout the world (for example, France, Norway, Sri Lanka), many of which were established during this period. Moreover, the growing evaluation profession encompasses a wide range of disciplines and evaluators from various disciplinary perspectives and backgrounds who increasingly are exchanging information, studying in interdisciplinary degree programs, working on evaluation projects together, publishing together, and meeting together in broadly focused evaluation conventions and meetings. The last type of interaction is reflected in the Evaluation Conclave in southern Asia, which held its first conference in New Delhi, India, in 2010. As previously alluded to, the Canadian Evaluation Society initiated its Credentialed Evaluator (CE) designation in 2010, with designation meaning that the holder has provided adequate evidence of having obtained the education and experience required to be considered a competent evaluator.

Our own experience at Western Michigan University and elsewhere is applicable here, because in 2003 we established the first Interdisciplinary PhD in Evaluation program (Coryn, Stufflebeam, et al., 2010). This program's instructors, advisers, and students have backgrounds in such diverse disciplines as nursing, substance abuse treatment, sociology, social work, business, community development, economics, education, engineering, psychology, chemistry, public administration, statistics, and political science. The evaluation-related learning experiences of both students and faculty members are greatly enhanced by students' conducting fieldwork projects together. Also in 2004, under Scriven's leadership, the IDPE program established the open-access, online *Journal of MultiDisciplinary Evaluation,* originally modeled after the *Harvard Law Review*. This journal has been widely subscribed to across disciplines and internationally. Clearly, the evaluation profession is becoming increasingly pervasive in disciplines and nations across the world.

During this period, many evaluation sponsors in the United States and elsewhere have returned to requiring so-called evidence-based evaluation methods. Generally, these are patterned after the evidence-based practice model in medicine (that is, randomized controlled trials). This approach is now often required for evaluating both independent and federally sponsored initiatives charged with identifying effective interventions (U.S. Government

Accountability Office, 2009). The reemerging federal requirements for applying experimental design mirror similar requirements that were previously advocated by Campbell and others in the 1960s (Campbell & Stanley, 1966). Also during this period, many alternative evaluation models and approaches, developed and prescribed in earlier periods in the history of evaluation, have gained greater prominence, legitimacy, and application. Among these are transformative evaluation, appreciative inquiry, participatory evaluation, empowerment evaluation, and theory-driven evaluation (Coryn, 2009).

It also is notable that during this period many long-standing disagreements among members of the evaluation community have resurfaced. In particular, disagreements about appropriate methods for inferring cause-and-effect relationships between programs and their outcomes as well as the persistent quantitative-qualitative debate (T. D. Cook, Scriven, Coryn, & Evergreen, 2010; Donaldson & Christie, 2005; Donaldson, Christie, & Mark, 2009), which for a short period diminished, have again intensified in the field. Relatedly, many international organizations, such as the World Bank and similar entities, which historically have relied on experimental and econometric methods of evaluation, have slowly begun a shift toward participatory evaluation, theory-driven evaluation, self-evaluation, and other alternative models and approaches for evaluating their humanitarian efforts. Also, the U.S. Government Accountability Office (2007) intensified its position that audits and other evaluations of federal programs must meet requirements for independence and objectivity.

Summary

In this chapter we have made the following points:

- Societies and their institutions require formal, systematic evaluations, as distinguished from everyday, informal evaluations (which are inevitable and also often lacking in reliability).
- The definitions of formal evaluation we provided are keyed to values (merit, worth, and probity); assessment criteria; and needed interface/communication and technical tasks.
- Key criteria for judging programs include quality, accomplishments, side effects, responsiveness to assessed needs, cost-effectiveness, probity, safety, sustainability, transportability, and others.
- The main functions of evaluation are formative and summative.
- Basically, noncomparative approaches are appropriate for evaluating programs under development, whereas comparative approaches often are needed to evaluate completed programs.
- Evaluation is a profession that serves all other professions and draws from the full range of disciplines.
- Professionalism requires one to obtain and use evaluation to increase competence and improve services.
- The professionalization of evaluation over time has been tied closely to the field of education and has occurred across the Pre-Tylerian Period, the Tylerian Age, the Age of

Innocence, the Age of Realism, the Age of Professionalization, and the Age of Global and Multidisciplinary Expansion.

- Evaluations themselves must be assessed against the standards of the evaluation field—for example, those developed by the Joint Committee and the U.S. Government Accountability Office.

REVIEW QUESTIONS

1. List, contrast, and discuss the benefits and limitations of formal evaluations.
2. Explain and give examples of evaluation's symbiotic relationships with other fields.
3. Cite what you see as the pros and cons of defining evaluation as a process of comparing outcomes to objectives and, conversely, the pros and cons of defining evaluation as the systematic assessment of merit and worth.
4. Summarize this chapter's stated rationale for employing values clarification in program evaluations; list and explain key issues in clarifying the values held by a program's stakeholders; and then list steps that you see as potentially effective for clarifying stakeholder values and applying them to reach evaluative conclusions.
5. Describe an example of how members of U.S. society were put at risk or harmed due to the failure of responsible parties to heed and act on the findings of an evaluation.
6. Cite some reasons why evaluators should search for side effects.
7. Suppose you want to increase your competence to conduct program evaluations. List and give examples of the main categories of skills you would seek to acquire, and discuss how you believe you could best obtain these skills.
8. Define what is meant by the terms *merit* and *worth*. Then, from your experience, write an example of a program or other entity that possessed merit but not worth. Describe how merit and worth were assessed. Explain why assessments of worth are dependent on context.
9. Give examples of cases that require comparative evaluations. Give examples of other cases that require only noncomparative evaluations.
10. Compare and contrast the terms *formative evaluation* and *summative evaluation*. Give an example of each of these evaluation roles.

Group Exercises

This section is designed to support group discussion of key issues addressed in this chapter. Each exercise summarizes a particular case, then provides instructions for the group's analysis of and response to the case. After your group's members have read an exercise, engage in discussion to arrive at your group's response to the particular assignment.

Exercise 1

The head of a large state government department has found himself under political pressure to commission an evaluation of each of the four divisions of his department. None of these divisions has ever been evaluated except in the most cursory fashion, and then only sporadically. What is evident to stakeholders (the public) is that services of all four departments are costly but inadequate, and that the poor quality of delivery is causing growing frustration. Realistic financial provisions and timelines have been made available for this major evaluation, according to the head of the department. Suppose your group has been selected to conduct the evaluation. Outline the important early decisions you would need to make about key aspects of the evaluation; the kinds of initial understandings you would need to reach with the head of the department and division heads; and the kinds of assurances you would seek and give so that a successful evaluation can eventuate.

Exercise 2

A superintendent of a small school district is beset with problems relating to the introduction of a new state-mandated science program for grades 7 through 9. She has heard of both formative and summative evaluation processes, but has little grasp of their functions and possible benefits if applied to the new science program. Your services are engaged to give the superintendent a thorough understanding of what constitutes formative and summative evaluation. Outline the relevance of either form of evaluation to the superintendent's problems, suggest a circumstance under which formative evaluation might lead to summative evaluation, and state the kind of cooperation an evaluation team would find essential to completing a successful evaluation. What advice do you give the superintendent?

Exercise 3

As a group, identify two studies: one that meets the requirements of a sound research investigation but not those of a summative evaluation, and one that meets the requirements of a sound summative evaluation. Then construct a matrix that shows the main distinctions and similarities between the two types of studies. Subsequently, discuss whether the distinctions your group identified are real and important.

Exercise 4

As a group, list points for use in explaining the essential differences between informal and formal evaluation, and also between formative and summative evaluation.

Notes

1. The Joint Committee is a standing committee that was established in 1975. Its approximately eighteen members have been appointed by about fifteen professional societies in the United States and Canada that are concerned with improving evaluations in education. The committee's charge is to develop

standards for educational evaluations. So far, it has created standards for evaluations of educational programs, personnel, and students. This book's first author was the committee's founding chair.

2. Although the Joint Committee expanded its definition of evaluation in the third edition of *The Program Evaluation Standards* (2011, p. xxv), we prefer the 1994 definition and refer to it throughout the chapter.
3. This section on the history of educational evaluation is largely based on a previous account by Stufflebeam, Madaus, and Kellaghan (2000), which included Madaus's incisive analysis of the early history of educational testing and evaluation.
4. The initial 1981 edition was titled *Standards for Evaluations of Educational Programs, Projects, and Materials*. For convenience, however, throughout this book we refer to all three editions as *The Program Evaluation Standards* and by the year of publication.

Suggested Supplemental Readings

Fitzpatrick, J. L., Sanders, J. R., & Worthen, B. R. (2010). *Program evaluation: Alternative approaches and practical guidelines* (4th ed.). Upper Saddle River, NJ: Pearson.

Fournier, D. M. (1995). Establishing evaluative conclusions: A distinction between general and working logic. In D. M. Fournier (Ed.), *Reasoning in evaluation: Inferential links and leaps* (pp. 15–32). New Directions for Evaluation, no. 68. San Francisco, CA: Jossey-Bass.

Kellaghan, T., & Stufflebeam, D. L. (Eds.). (2003). *International handbook of educational evaluation*. Norwell, MA: Kluwer.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4th ed.). Thousand Oaks, CA: Sage.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Thousand Oaks, CA: Sage.

Scriven, M. (1993). *Hard-won lessons in program evaluation*. New Directions for Program Evaluation, no. 58. San Francisco, CA: Jossey-Bass.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Thousand Oaks, CA: Sage.

Shadish, W. R., & Luellen, J. K. (2005). History of evaluation. In S. Mathison (Ed.), *Encyclopedia of evaluation* (pp. 183–186). Thousand Oaks, CA: Sage.