



## CHAPTER ONE

---

# OVERVIEW OF THE EVALUATION FIELD

---

The term *evaluation* has been in the English language for centuries, and it has had diverse functions and meanings during that time. Only in recent decades, in particular the latter part of the twentieth century, has more precision been given to the term, including specificity to basic concepts and more explicit explanations about its aims as a functioning entity. This book's four central purposes are to help those who are studying, commissioning, and conducting evaluations to (1) become knowledgeable about the evaluation profession, including its vocabulary, functions, main methods, organizations, standards, guiding principles, specialized evaluation jobs, and sources of information about evaluation; (2) gain a perspective on evaluation theory; (3) understand and assess the major approaches for conducting program evaluations; and (4) acquire knowledge of procedural guidelines for applying evaluation approaches.

In this opening chapter, we are seeking answers to two questions: What is evaluation in today's terms? and What should it achieve? Basically, the chapter provides an overview of program evaluation. It discusses the relevance of evaluation to society and its pervasive character. It defines evaluation and other key concepts of the evaluation profession and contrasts the evaluation profession to other professions. Evaluations may involve multiple values of individuals, organizations or societies, and these may compete. However, we suggest that when there is a clear knowledge of the generic qualities underlying all good evaluation procedures and precepts, diverse elements may be satisfactorily encompassed in reports.

The chapter also identifies generic evaluative criteria, different ways of defining and applying program standards, and main uses of evaluation. It distinguishes between comparative and noncomparative evaluations and between informal and formal evaluations. It discusses the responsibilities of organizations and individual professionals and other service providers to evaluate and improve their services. It identifies major evaluative methodologies. It characterizes the evaluation profession and how it has developed, references guiding principles and professional standards for evaluations (discussed in detail in Chapter Three), and defines a range of evaluation-related jobs. In general, this chapter aims to acquaint readers with the basics of evaluation as a field of study and practice. The chapter is organized around key questions about evaluation that should be of interest to those relatively new to the field, seasoned evaluators who need to explain their profession to others, and evaluation instructors.

As do most other chapters, this one concludes with a set of review exercises and group discussion questions. After reading the chapter, we suggest you complete them as a means of assessing and confirming your mastery of the chapter. Also, you may find it useful to discuss your responses to the chapter's main organizing questions and review exercises with colleagues or fellow students. We suggest further that you enrich your understanding of this chapter's material by consulting the references listed at the end of the chapter and the Western Michigan University Evaluation Center Web site ([www.wmich.edu/evalctr](http://www.wmich.edu/evalctr)).

---

## Why Is Evaluation Important?

Evaluation arguably is society's most fundamental discipline. It is oriented to assessing and helping to improve all aspects of society. Proper objects for evaluation cover a wide range of entities: school programs, libraries, museums, hospitals, physicians, immunization programs, continuing medical education programs, courts, lawyers, judges, universities, schools, university curriculum, instructors, construction projects, ladders, food and other consumer products, telecommunication services, postal services, government agencies, transportation services, parks and recreation programs, agricultural extension services, environmental policies, disease prevention and control programs, national defense, border control, research plans and findings, theories, and many more. These examples illustrate that evaluation is ubiquitous. It permeates all areas of scholarship, production, and service 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 on 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 poor. Evaluation serves society by providing affirmations of worth, value, improvement (and how and when this should happen), accreditation, accountability, and, when necessary, a basis for terminating bad programs.

---

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

We refer to the object of an evaluation as the *evaluand* or (in the case of a person) the *valuee* (Scriven, 1981). Evaluands may be individuals, programs, projects, policies, products, equipment, services, concepts and theories, or organizations. Although this book concentrates on program evaluation, one can refer to a range of other areas of evaluation: personnel evaluation, product evaluation, institutional evaluation, student evaluation, and policy evaluation. The scope of evaluation applications broadens greatly when one considers the wide range of disciplines to which it applies. Among others, one can speak of educational evaluation, social services evaluation, arts evaluation, city planning and evaluation, real estate appraising, engineering testing and evaluation, hospital evaluation, drug testing, manufacturing evaluation, consumer products evaluation, agricultural experimentation, and environmental evaluation. Much of this book's material, including the concepts and methods in this chapter, is applicable across the full range of evaluation areas.

---

## 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 evaluation information. One clear example was the continued sales of the Corvair automobile after its developers and marketers knew of its rear-end collision fire hazard. 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.

We have thus aimed this book not only at evaluators or those studying to become professional evaluators, but also at the users of evaluations. We believe that everyone who plays a decision-making role in serving the public should obtain and act responsibly on evaluations of their services. Fulfilling this role requires each such decision maker to take appropriate steps to become an effective, conscientious, evaluation-oriented service provider. The production and appropriate use of sound evaluation is one of the most vital contributors to strong services and societal progress.

---

## How Does the Field of Evaluation Relate 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 of research on evaluation. Its research material evolves from, and is closely connected to, the wide range of evaluations conducted in all fields, such as those referred to earlier in this chapter. Evaluation has many professional organizations (including the American Evaluation Association and other state and national evaluation associations, with more than twenty in countries outside the United States) and university training programs (among them, the interdisciplinary doctoral evaluation program at Western Michigan University and other evaluation graduate programs at Claremont Graduate University, the University of Illinois, Ohio State University, the University of Minnesota, the University of North Carolina, the University of Virginia, and the University of California at Los Angeles). The field also has standards for evaluation services (including the Joint Committee 1988, 1994, and 2003 standards, respectively, for evaluating personnel, programs, and students) and the U.S. *Government Auditing Standards* (Government Accountability Office, 2003), plus the American Evaluation Association's *Guiding Principles for Evaluators* (2004). Moreover, as discussed in Chapter Two, the evaluation profession possesses concepts and tools required to examine both evaluation theory and evaluation practice.

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 make 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 professional-

ism. 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. In the latter instance, we refer to the evaluation of evaluation as *metaevaluation* (addressed in Chapter Twenty-Seven). On the other side, 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 relationship between evaluation and other fields of study and practice.

We believe that the more systematic, more thorough, and the more disciplined the evaluation procedures are, the more beneficial and enduring will be the changes in the evaluand. This is particularly the case with formal evaluations. As mentioned in this book from time to time, evaluators can only do their best; despite strenuous efforts to involve clients in evaluations, there is no certainty that clients will heed and act on sound evaluation findings. Later chapters refer to the necessity for evaluators to pursue disciplined design, information gathering, organization of information, analysis of information, and communicating outcomes in sound reports. Evaluations that lack these kinds of disciplines typically are fruitless, wasteful, and misleading. Again, as later chapters show, there is no single best method or model to carry out evaluations. Nonetheless, the alternative approaches all require vigorous and objective applications of sound methods to reach valid conclusions.

---

## What Is Evaluation?

Mainly because there have been different approaches to evaluation over the years, definitions of the term *evaluation* have themselves differed. In earlier times, for example, evaluation was commonly closely associated with assessing achievement against behavioral objectives or conducting norm-referenced testing. Then, particularly during the 1970s, emphasis was given to professional judgment. Since that time, an increasing number believe that evaluation is the collection and analysis of quality information for decision makers. These and other “definitions” of *evaluation* have elements of credibility, depending often on the type of evaluation study being undertaken. We take the stance in this book that no one method of evaluation is necessarily the best or most appropriate and that an eclectic approach may well be the most suitable to a particular context. It now remains to see whether it is possible to arrive at a definition, or definitions, that satisfactorily cover the often complex array of approaches and activities that constitute evaluation.

Evaluation has been defined in different ways. One of the earliest and still most prominent definitions states that it means determining whether objectives have been achieved. We reject this definition, because following it can cause evaluations to fail. One of its 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. The objectives might well be corrupt, dysfunctional, unimportant, not oriented to the needs of the intended beneficiaries, or mainly reflective of profit motives or other conflicts of interest of those in charge of the program. Another problem is that this definition steers evaluations in the direction of looking only at outcomes. Evaluations should also examine a program's goals, structure, and process, 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 side effects, which can be critically important in determining whether a product or other entity is safe. In addition to these deficiencies, evaluations employing the objectives-based definition provide feedback only at the end of a program. Evaluations also have an important role in helping to plan and guide programs toward successful outcomes. These defects and limitations are sufficient for us to reject the objectives-based definition of evaluation.

We also reject definitions that equate evaluation with any one methodology. Sometimes evaluations based on experimental designs 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, would be counterproductive, or would fail to address key questions about needs, goals, plans, processes, side effects, and other topics. Similarly, other useful techniques—such as surveys, standardized tests, site visits, or self-studies—are far too narrow in the information they yield to provide a sufficient basis for most program evaluations. Evaluation should not be equated with any one methodology. Instead, it should encompass all methods that are necessary and useful to reach defensible judgments of programs or other entities, and evaluators should selectively apply given methods.

In this book, we advocate a basic definition of evaluation<sup>1</sup> put forth by the Joint Committee on Standards for Educational Evaluation (1981, 1988, 1994, 2003). We present three variations of the definition. First, we present the definition as the Joint Committee stated it. Its definition is general. It calls for evaluations to be systematic and focused on determining an object's value. We subsequently 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 lay audiences and focusing their attention on the essence of evaluation. The second rendition can be helpful when discussing with clients the values that should be referenced when evaluating a particular program, institution, 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, consistent with common dictionary meanings of *evaluation*, and has the imprimatur of the prestigious Joint Committee on Standards for Educational Evaluation. This is the definition to use when discussing evaluation at a general level. Below we unpack the Joint Committee's definition to note and discuss its key concepts.

*Evaluation's* root term, *value*, denotes that evaluations essentially involve making value judgments. Accordingly, evaluations are not value free. They need to be grounded in some defensible set of guiding principles or ideals and should determine the evaluand's standing against these values. This truism presents evaluators with the need to choose the appropriate values for judging an evaluand. For example, in evaluating U.S. public services, evaluators should be true to, and sometimes specifically invoke, precepts of a democratic society such as freedom, equity, due process of law, and the need for an enlightened population. The Joint Committee's definition partially addresses the need to determine values by denoting that evaluations should assess merit or worth. Scriven (1991) points out the nontrivial differences between these two concepts and their important role in determining an evaluand's value. Table 1.1 summarizes the characteristics of these concepts, with further discussion below.

**Merit.** Generally 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 has good 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 his assessment of merit on published standards of effective teaching and a 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

TABLE 1.1. CHARACTERISTICS OF MERIT AND WORTH.

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

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.

**Worth.** An evaluand that has high merit might not be worthy. By *worth*, we refer to a program's combination of excellence and service in an area of clear need within a specified context. Suppose the middle school program was 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 had increased their production of competent middle school history teachers, and many of these new teachers could not find jobs. Arguably, the state would no longer need the special emergency program, because the state's universities were now supplying more qualified middle school history teachers than the schools could 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 the program's worth could be gauged only after an assessment had been made of the needs 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.

**Needs.** By a *need*, we refer to something that is necessary or useful for fulfilling a defensible purpose. We define a *defensible purpose* as a desired end that has been legitimately defined consistent with a guiding philosophy, set of professional standards, institutional mission, mandated curriculum, national constitution, or public referendum, for example. Other terms for defensible purpose are *legitimized mandates, goals, and priorities*. In the middle school illustration, presumably the state curriculum required that all students in the state be well educated in designated areas of history. This “defensible purpose” required further that school districts employ competent history teachers, which fits our definition of an entity that is necessary or useful for fulfilling the defensible purpose of sound history instruction—that is, a need. Since the state found this need was now being fulfilled by state colleges and universities, a proposal to retain this excellent special program would have met the criterion of merit but not the criterion of worth. In reaching judgments of something’s worth, evaluators should first identify needs and determine whether they are being met or are unmet in the context of interest.

Needs may be of either the outcome or treatment variety. 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 in order 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, though, the meeting of outcome needs is conditional on the meeting of treatment needs. For example, a dentist would be unlikely to check patients with no tooth decay for use of fluoridated water or toothpaste. Here the outcome need (for cavity-free teeth) is being met, and it is not necessary to check on the treatment need related to fluoridation.

**Needs Assessments.** In general, a *needs assessment* is a systematic assessment of the extent to which treatment or outcome needs are being met (Stufflebeam, McCormick, Brinkerhoff, & Nelson, 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 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, social, moral, and vocational areas, the achievement test scores would be insufficient to assess the full range of questions concerning students' outcome needs. To ensure that needs assessments are valid, they have to be keyed to the full range of intended outcomes. Some needs assessments will have a narrow scope and appropriately be assessed against a quite restricted construction of outcome needs.

Even in a narrowly focused program, 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 gauge examinations of program quality and outcomes against the assessed outcome and treatment needs of beneficiaries. Table 1.2 offers a convenient summary of key concepts related to needs assessment.

**TABLE 1.2. CONCEPTS RELATED TO 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 assessment of the extent to which treatment or outcome needs are being met	Examination of students' scores on national tests and evaluation of the involved teachers

**Evaluations Should Be Systematic.** Beyond its focus on merit and worth, the Joint Committee's 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 a countermeasure to careless 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**

While the Joint Committee's 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 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. Below we discuss each of the values noted in the extended definition of evaluation.

**Probity.** During the writing of this book, there was a rash of public scandals in which major U.S.-based corporations 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, it also was alleged to have distorted and covered up information to hide the company's unethical, unlawful practices. Here we see that the company cheated its shareholders, workers, and ultimately the public, and that the audit company was charged with aiding and abetting the fraud. On another

front, there have been despicable scandals across the globe in which clergy and teachers were found to be pedophiles.

Clearly, the public interest 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, evaluations should check on a program's uncompromising adherence to moral standards. However, when probity breaches are found, 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 part and parcel of many evaluations of programs and of evaluations of evaluations.

**Feasibility.** While a service might be of high quality, directed to an area of high need, and unimpeachable on ethical grounds, it still could fail 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 service should at least be modified in these areas to make it more feasible. Obviously a good evaluation of the service should speak to this issue and, where appropriate, provide direction for making the service easy to apply, efficient in the use of time and resources, and politically viable. Evaluation of a program's feasibility sometimes justifies a cancellation decision. This argument in favor of assessing feasibility seems applicable to all programs.

**Safety.** Many evaluations focus squarely on the issue of safety. Obvious cases are evaluations of new pharmaceutical products, laboratory equipment, meat and other food products, automobiles, stepladders, electrical equipment, and insecticides. Consumers are at risk to the extent that such commodities are manufactured and sold without rigorous safety checks and appropriate cautions. Moreover, many programs also require evaluations that examine the safety of facilities, equipment, activity regimens, crowd control, and others. To see the importance of safety evaluations in programs, one only need 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 escape exits and 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 products.

**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. When they have far-reaching implications, evaluators should look at and make judgments about the significance of their mis-

sion, structure, and outcomes. Such an assessment can be especially important in deciding whether and how far to disseminate lessons learned and helping interested parties make sound decisions concerning adopting and adapting a program or parts of a program. Evaluators should often consider the possibility that a 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. In the United States, an educational evaluation would be incomplete if it did not assess whether a public educational service is provided for, and made available to, public school students from all sectors of the society. This concept of equity is complex. It is not enough to say that public education services may be sought and used by all people. As Kellaghan (1982) has argued, when educational equality truly exists, there will be seven indications of equity:

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 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 welfare of all societal groups in their community environments.

## 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, we can see that the Joint Committee's 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 supplied information should reflect the full range of appropriate values.

We now expand the definition further 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 and technical steps. In the former, evaluators communicate with clients and other stakeholders in the interest of planning relevant evaluations; conveying clear, timely findings; and aiding and abetting use of the findings. The technical steps 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 interface and technical aspects of evaluation. This is best accomplished through formal courses and experiences in planning, conducting, and reporting a wide range of evaluations. We have characterized the work of evaluation in four steps: delineating, obtaining, reporting, and applying. Part Four of this book addresses these process steps in detail. We end by discussing the final major feature of our operational definition of evaluation, which concerns the nature of information included in evaluations.

***Delineating.*** The delineating step 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 govern the evaluation work. Basically, the delineating step encompasses effective, two-way 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, the ability to develop rapport, interviewing, content analysis, values clarification, conceptualization, proposal development, negotiating, contracting, and budgeting. The results of this step should set the stage for the ensuing data collection work. In fact, delineating activities extend throughout the evaluation in response to changing circumstances in the program, identifying new audiences, continuing interaction with stakeholders, and emerging needs for information. Moreover, a delineation process that is carried out thoroughly and professionally establishes a basis for essential trust and rapport between an evaluator and client group.

***Obtaining.*** The obtaining step 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 in order 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 step, the evaluator provides feedback to the client and other audiences. Typically such work includes preparing and delivering interim oral and printed reports, multimedia presentations, press releases, printed final reports, and executive summaries. 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 occur throughout and after completion of an evaluation. Particular areas of needed expertise are clear writing, the ability to format reports, competent editing, information technology, effective oral communication, and dissemination skills. Effective reporting sets the stage for applying the evaluation findings.

**Applying.** The applying step is under the control of clients 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, critique of the client group's plans to apply findings, or responses to questions from the client or other users. We have found that this kind of assistance from evaluators is highly regarded by clients. 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. Therefore, in starting an evaluation, the evaluator and client should consider the possibility of the evaluator's involvement in the applying 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 of principles and procedures of effective change and research on utilization of evaluation (see Alkin, Daillak, & White, 1979; Patton, 1997). Also, they need skills of communication, consulting, group process, and counseling.

**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 and how good it is.

This requires the evaluator to collect and report both descriptive and judgmental information.

*Descriptive Information.* A final evaluation report should describe the program's goals, plans, operations, and outcomes objectively. 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 such matters as 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 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, the evaluator needs to determine and report the extent to which the pertinent inoculations were administered to all the targeted children as planned. If 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 summarize judgments of the program. According to the above definition of evaluation, evaluations inevitably involve valuing, that is, judgment. Judgment-oriented feedback can be a vital, positive force when it is integral to development, directed to identifying strengths as well as weaknesses, and focused on improving the evaluand. Appropriate sources of judgments include program beneficiaries, program staff, pertinent experts, and, of course, the evaluator, among others.

---

## 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 projects to fund, based on their relative quality and importance compared with other possible uses of available funds. For a project already funded, the 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 acceptable

alcohol in a driver's blood. These examples are not just abstractions. They reflect true, frequent circumstances in society when 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, all protocols for such precise evaluative determinations are arbitrary, but they are not necessarily capricious. While many decision rules are set carefully in the light of relevant research and experience, the rules are human constructions, and their precise requirements arguably could vary. The arbitrariness of cut scores is also apparent in different alpha and beta levels that investigators invoke for determining statistical significance and statistical power. Typically, the alpha level is set, by convention, at .05 or .01, but might as easily be set at .06 or .02. (See Chapter Twenty-Five for a discussion of statistical significance.) 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 having juries of peers vote 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 funding proposals or job applicants, employing cut scores on students' achievement test scores, polling constituents, grading school homework assignments, contrasting students' tested performance to national norms, appropriating and allocating available funds across competing services, and charging an authority figure or engaging an expert panel to decide on a project's future. Although none of these procedures is beyond challenge, as a group they have addressed society's need for workable, decision-making tools.

Some of these procedures have advance setting of cut scores, standards, or decision rules in common. For example, in the United States, it is known in advance that all twelve 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 are also part and parcel of 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 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, 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 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, 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, evaluative criteria, and defensible levels of soundness for each one and for all as a group. So how do evaluators function when they have to make up their plans, identify criteria, and interpret outcomes without 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.

---

## **What Are Performance Standards? How Should They Be Applied?**

Often evaluation is characterized as comparing a performance to a standard. This concept 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 where 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 the participants 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 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 level of expectation. At the other end of the distribution, beneficiaries who are far below the standard may believe it is futile to reach the

cut score standard, and consequently they give up. Also, cut score standards have a tendency to narrow a program's focus; lock it in to predetermined objectives; and inhibit it from responding over time to emergent needs, developments, and insights.

An alternative to this narrow, preordinate approach to standards-based evaluation is to view a program standard as a requirement for continuing improvement. Deming (see Walton, 1986) sold this idea 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. In the education field, 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. It makes no sense to close the gap between high and low achievers, since 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 of different members of an evaluation audience is a necessary and difficult task in evaluations. We would argue that it is the shared and differential needs of the clients 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 a service.

Sometimes an evaluator should address the value conflict issue by separately interpreting process and outcome information against each main set of values or priorities of different segments of the stakeholder population. In other cases, the evaluator might serve better by seeking out and assessing multiple programs or services while considering that each might better address one set of values than another set.

In planning evaluations, evaluators should deal directly with the important matter of choosing and applying pertinent values. 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 the evaluator

to work within her or his basic philosophical convictions, that is, to act with integrity. Evaluators should also take into account a program's mission and the pertinent values, needs, and priorities of the program's leaders as well as the beneficiaries 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 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 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.

---

## Should Evaluations Be Comparative, Noncomparative, or Both?

Evaluations 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 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 find a better alternative by opening consideration to other providers.

In general, we think that evaluations should be comparative before one purchases a product or service or starts a program, noncomparative during program development or use of a service, and periodically comparative after development or sustained use in order to open the way for improvements or radical alternatives. Whether an evaluation should be comparative depends on the intended uses of the evaluation.

---

## 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*. Basically, they provide feedback for improvement. They are prospective and proactive. They are 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, in doing so, should pay close attention to the nature and needs of the consumers. Formative evaluations 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 all such formative evaluations is directed to 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, requiring random assignment of program participants to alternative program treatments and treatments to be kept stable and unchanging, 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 evaluation is to produce accountability or summative reports. These are retrospective assessments of completed projects, established programs, or finished products. Summative evaluations typically occur following development of a product, completion of a program, and 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 determining 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 sponsors 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 the audience in such situations. Field tests can make 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, or other evaluand. But even then, a comparative experiment is only a part of a sound summative evaluation. The full range of dimensions involved in summative evaluations is seen in Michael Scriven's Key Evaluation Checklist (accessible at [www.wmich.edu/evalctr](http://www.wmich.edu/evalctr) and outlined in Chapter Sixteen).

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

### **Relationship Between Formative and Summative Evaluations**

Both formative and summative evaluations are needed in the development of a product or service or, in the case of personnel, help in developing potential and gauging the extent to which required criteria for certification, tenure, promotion, and the like are met. Too often, summative evaluation is carried out only for judging programs or personnel. This restricts development processes and may lead to inadequate or even incorrect conclusions. Subjecting a trainee nurse to an accountability assessment while ignoring the obvious advantages of fostering improvement through formative methodologies is foolish. Similarly, when a wine-sealing machine is being built, the lack of formative information covering cost, efficiency, faults, and potential marketing would be disastrous to the manufacturers. Evaluations delayed until the near completion of a training or probationary period or a project's development may be too late to promote and assist successful outcomes.

The relative emphasis of formative and summative evaluations will change according to the nature and circumstances of the evaluand. In general, formative evaluation will be dominant in the early stages of a program and less so as the program matures. Summative evaluation will take over as the program concludes and certainly after it is completed. However, 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, nonetheless, that all evaluations fall into one or both categories. Many of the evaluation approaches depicted later in this book can be used for formative or summative purposes or both.

Robert Stake (1969) made an interesting observation, apropos the relationship between formative and summative evaluations, that formative evaluations are

**TABLE 1.3. FORMATIVE EVALUATION AND SUMMATIVE EVALUATION.**

<b>Descriptors</b>	<b>Formative Evaluation</b>	<b>Summative Evaluation</b>
Purpose	Quality assurance; improvement	Provide an overall judgment of the evaluand
Use	Guidance for decision making	Determining 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, and safety
Orientation	Prospective and proactive	Retrospective and retroactive
When conducted	During development or ongoing operations	After completion of development
Particular types of service	Assists 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
Variables	All aspects of an evolving, developing program	Comprehensive range of dimensions concerned 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
Evaluation plans	Flexible, emergent, responsive, interactive	Relatively fixed, not emergent or evolving
Typical methods	Case studies, observation, interviews, not controlled experiments	Wide range of methods including case studies, controlled experiments, and checklists
Reports	Periodic, often relatively informal, responsive to client and staff requests	Cumulative record and assessment of what was done and accomplished; contrast of evaluand with critical competitors; cost-effectiveness analysis
Relationship between formative and summative evaluation	Often forms the basis for summative evaluations	Compiles and supplements previously collected formative evaluation information

closely connected to “insiders,” that is, program developers, while 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 time lines, finance, 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.

Finally, formative evaluations often form the basis for summative evaluations. If this is to occur, those who commission the studies and the evaluators must make it clear to all involved that this will occur. It should also be recognized that on occasions, the merit or worth of a formative evaluation may be strengthened by the intervention of summative evaluations (usually carried out by external personnel) at critical points of a program’s development. Such procedures require sound professional collaboration, which is a hallmark of good evaluation practice. (For other dimensions of this topic, see the section entitled “Why Are Internal Evaluation Mechanisms Needed?” later in this chapter.)

### **Evaluations to Assist Dissemination**

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: alternative automobiles, insurance policies, mortgages, breakfast cereals, chain saws, refrigerators, 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 Part Two, we describe Scriven’s consumer-oriented evaluation approach. Note that many sound consumer-oriented evaluations do not employ experimental designs; however, under the right circumstances, experiments can be a useful means to meeting some of the information needs of consumers.

### **Evaluations to Foster Enlightenment**

The fourth use of evaluations is to foster enlightenment, that is, new understandings arising from revelations. Basically, evaluation and research are different enterprises. The former attempts to consider all criteria that apply in determining value, while the latter may be restricted to the study of selected variables that are of

interest in theory development or policy formation. 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 the results of evaluations 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 and policymakers. With some forethought, careful planning, and appropriate budgeting, evaluations may serve not only to guide operating programs, sum up their contributions, and disseminate 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?

To this point, it should be clear that program evaluation is a demanding field of practice. At the same time, everybody evaluates essentially all the time, whether making choices about the trivial or the critical. 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, while others require a rigorous, systematic approach, including an independent perspective.

### Informal Evaluations

Everybody performs informal evaluation 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, over-the-counter medications, 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, or try out something before deciding to keep it. These are all good and appropriate evaluative moves and fit within our

general concept of informal evaluation. However, the conduct of informal evaluations 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, 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 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 systematic, we refer to evaluations that are relevant, designed and executed to control bias, kept consistent with appropriate professional standards, 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 Three we summarize professionally developed guiding principles for evaluators, professional standards for program evaluations, and the 2003 U.S. Government Auditing Standards. Building on these principles and standards, this book is designed to help students and practicing evaluators attain the perspectives and basic level of proficiency required to implement defensible formal evaluations.

---

## How Do Service Providers and Their Organizations Meet Requirements for Public Accountability?

We cannot stress too much that society is dependent on sound evaluations to obtain safe, high-quality goods and services from a wide range of professionals and other service providers. Any operatives should deliver services that are of high

quality, up-to-date, safe, efficient, fairly priced, honest, and, in general, in the public interest. In order to meet accountability requirements, each profession, public service area, and society should regularly subject services to formal evaluations. Some of the evaluation work is appropriately directed at regulation and protection of the public interest. It 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 or commission. These internal or self-evaluations are an important aid to continually scrutinizing and improving services and also supplying data needed by the independent or external evaluators.

## Accreditation

A wide range of accrediting organizations periodically assess the performance of member organizations against formally established standards. Typical accrediting evaluations are grounded in clear accreditation criteria and guidelines for self-assessments. 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 and to what extent 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 should operate and use findings from an internal evaluation mechanism.

## 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. In order 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 the 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, 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.

### **Why Is Evaluation a Personal as Well as a Corporate Responsibility?**

Even if an institution has a strong evaluation unit, every professional in the institution needs further evaluation. There is no escaping the fact that evaluation is a personal as well as an institutional responsibility. Offices of evaluation and accrediting organizations can help meet an institution's major responsibilities for evaluation and accountability. Offices of evaluation can also provide in-service training and technical support in evaluation to the institution's staff. 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 for finding out and acting on what is going right and wrong. Moreover, conducting and acting on sound evaluation is part and parcel of what it means to be a professional: a member of an established profession who continually works to deliver better service. We hope this book will both inspire and assist individual professionals and other service providers as well as evaluation students and specialists to meet their needs for developing competence in and effectively carrying out systematic evaluations.

---

### **What Are the Methods of Formal Evaluations?**

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 communications, 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 have pertinence to systematic evaluation. The well-prepared evaluator will have a good command of concepts and techniques in all these areas and will keep informed about potentially useful technological developments. Evaluators who would exert leadership and help advance

their profession should contribute to the critique of existing methods and the development of new ones.

Over the years, many evaluators have chosen, even championed, the exclusive use of a few techniques. Some have equated evaluation with their favorite methods—for example, experimental design, standardized testing, questionnaires, or site visits. 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. A key point in the latter position is that use of multiple methods and perspectives enhances the dependability of inferences and conclusions and causes appropriate levels of circumspection.

We believe that evaluators should know about a wide range of pertinent techniques and how well they apply in different evaluative contexts. Then in each evaluative situation, they can assess which techniques are potentially applicable and which most likely would work best and in combination to serve the particular purposes of the given evaluation. 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, theorizing, interviewing, focus groups, survey research, case study, content analysis, observation, checklists, goal-free evaluation, advocacy 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 still has far to go. The evaluation field now has national and state professional societies of evaluators; annual conventions; a substantial literature, including professional journals and a wide range of theoretical and technical books; specialized Web sites; master's and doctoral programs; institutes and workshops on specialized evaluation topics; 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 compared with the field's status in 1964, when it had none of the above.

However, the evaluation field is still immature when compared with established professions, such as medicine, law, engineering, and accounting, and other service areas, such as those of master plumbers, licensed electricians, prosthetists, and dental hygienists. Especially, the evaluation field lacks some of the hallmarks of a mature profession. For example, membership in the American Evaluation Association is open to anyone regardless of training and expertise in evaluation. Furthermore,

the field has no mechanisms for certifying competent evaluators. 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 despite institutional, business, and other evaluations being widely recognized as essential to the health of any organization, acceptance of tertiary training to gain qualifications as an evaluator is lagging worldwide.

---

## What Are the Main Historical Milestones in the Development of the Evaluation Field?

The evaluation field evidences only modest efforts to systematically record and analyze its history. Any profession, in order to serve the needs of 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 evolved in the area of education.<sup>2</sup> We believe this is appropriate and will be instructive, since the profession of evaluation developed earliest and most heavily within the field of education. We provide only a brief historical sketch in order to note 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 five 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 the years 1958 to 1972; and (5) the Age of Professionalism, which includes developments from 1973 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 ex-

amination 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. Especially, 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 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 leadership to educational administrators in New York City, in 1895 Rice 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 in 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 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 continues today. 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 standardized testing, especially when high stakes are 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. Robert 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 criticisms, the use of objective achievement tests has continued to flourish. 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.

While program evaluation only recently has 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, Ralph 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 views 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 had been achieved. As a consequence 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, John 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 behavioristic psychology.

Tyler was drawn directly into this movement when he was commissioned to direct the research component of the now famous Eight-Year Study (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 program, Tyler was able to publicize what he saw as clear-cut advantages of his approach over others. Since 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. These 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 complacency regarding serious societal problems. As a consequence, we think this time is better referred to as the Age of Innocence, or even as the Age of Social Irresponsibility.

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 in order to provide the wide range of educational services that were common in larger school systems: mental and physical health services, guidance, food service, music instruction, expanded sports programs, business and technical education, and community education. Enrollments in teacher education programs ballooned, and college enrollments 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 education system. While 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, 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 pressures and 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 developments of evaluation methodologies would have to be grounded in concerns for accountability, usability, and relevance. Their rude awakenings 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, that would see 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 language 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 Tyler 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 (Jones, 2003). Third, the professional judgment approach was used to rate proposals and check periodically on the efforts of contractors. Finally, many evaluators undertook to evaluate curriculum development efforts through the use of 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 these 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 realize that their efforts were not succeeding.

This negative assessment was well reflected in a landmark article by 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 likely 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 the conceptualization and conduct of evaluations that were to be tested and found valid 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. Subsequently, President Lyndon Johnson 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, based largely on the use of standardized educational achievement tests and superficially on Tyler's objectives-oriented 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 the school or a particular program, these tests were at best indirect measures of learning, measuring much the same traits as general ability tests (Kellaghan, Madaus, & Airasian, 1982).

The use of standardized tests entailed another problem. Such an approach to evaluation 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 a dearth of 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 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, but they did not learn how to derive such objectives from information on the needs and problems of their students. Consequently, once they met the students to be served, the educators soon forgot or set aside as irrelevant the objectives they had written before the start of a 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 to fail. 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 highly dynamic, field-based, generalized assistance programs, especially in the course of development.

As a result of the growing disquiet with evaluation efforts and consistently negative findings, Phi Delta Kappa set up a 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 (1966).

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 Tyler model. Glaser (1963), Tyler (1967), and Popham (1971) pointed to criterion-referenced testing as an alternative to norm-referenced testing. Cook (1966) called for the use of system analysis techniques to evaluate programs. Scriven (1967, 1974), Stufflebeam (1967, 1971), and Stake (1967) introduced new models for evaluation that departed radically from prior approaches. These conceptualizations recognized the need to evaluate goals, look at inputs, examine implementation and delivery of services, as well as measure intended and unintended outcomes of the program. They also emphasized the need to make (or collect) judgments about the merit 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 that follow this one 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 the Present

Beginning in about 1973, the field of evaluation began to crystallize and emerge as a distinct profession 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 this development in the context of the field in the previous period.

At that time, 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 or specialized journals through which evaluators could exchange information about their work. Essentially no literature about evaluation existed except for unpublished papers that circulated through an underground network of scholars. There was a paucity of preservice 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 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 to professionalize their field beginning in the 1970s is quite remarkable. A number of journals emerged. 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 (WMU)—have developed graduate programs in evaluation. The WMU program is the world's only interdisciplinary doctoral program in evaluation.

Increasingly, the field has looked to metaevaluation (Scriven, 1975; Stufflebeam, 1978, 2001) as a means of ensuring and checking the quality of evaluations. A 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 the field in their use (Joint Committee on Standards for Educational Evaluation, 1981). This review process has worked effectively, producing the second edition of the *Program Evaluation Standards* in 1994. Moreover, the publication in 1988 of the Joint Committee's *The Personnel Evaluation Standards* signaled an advancement in the methods for assessing systems for evaluating personnel. In addition, the Joint Committee issued *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. General Accounting Office's *Government Auditing Standards* (2002, 2003). 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).

---

## What Are the Main Roles in Evaluation Work?

Thus far, we have given and discussed a basic definition of evaluation, identified a range of evaluative criteria and other evaluation concepts, identified some of the methods involved in evaluation work, and provided a historical overview of the evaluation field. We conclude this chapter by looking at evaluation roles.

Whether one or many persons conduct an evaluation, a number of roles typically have to be implemented, and when one considers the work entailed in supporting and promoting evaluation as a profession, this view of roles becomes even broader. Importantly, those who will be involved as either specialists or generalists in evaluations need a broad view of the pertinent roles to enhance their effective participation and collaboration in the evaluation work. Perhaps nothing is so destructive of the potential contribution of evaluation as the attitude that evaluation is an independent research pursuit of an individual investigator. On the contrary, evaluation at its best involves much collaboration.

In attempting to provide a broad view of the work involved in individual evaluations and in the evaluation profession in general, we have identified twelve basic roles. While other writers might identify a different set and assign different labels, we believe our roles encompass most, if not all, work responsibilities that evaluators might have to implement. Furthermore, in our experience, these twelve roles have been associated with actual job assignments. Of course, there would not always be a one-to-one correspondence between role and job, but we think these twelve roles should be useful for thinking about the broad array of evaluation activities and for considering possible careers in evaluation, designing evaluation training programs, and designing and staffing work of particular evaluations.

### Evaluation Client

Most worthwhile evaluations include a client who is integrally involved. This is true whether the study is conducted by one or more evaluators to serve a separate client or by a person to serve his or her own purposes. The client is the person or

group that will use the results for some purpose, such as program selection, program improvement, or accountability to a sponsor. The client group encompasses whoever commissioned the evaluation, as well as those who will attend to and use the results. In one way or another, the time and resources required for the evaluation are provided by the client or by one or more persons in the client group. The clients are also crucial sources of the questions to be addressed and of the criteria to be employed in interpreting the results; they also bear primary responsibility for applying the findings. Often an evaluation cannot be done if the client fails to set the stage politically and take steps to ensure full cooperation by those who will be involved.

A variety of key clients are found in educational, social, and health organizations. These include, for example, school district superintendents, hospital administrators, and agency directors; school principals, department chairpersons, and grants administrators; foundation presidents, military leaders, and corporation heads; and teachers, psychologists, and physicians. If evaluation is to contribute maximally to the provision of meritorious and worthy services to people across the wide range of evaluations, this full range of clients should be involved in and served by evaluations. Moreover, we believe they must be actively engaged in commissioning and planning evaluations in order to derive the information they need to do their jobs well.

Therefore, in order to fulfill these expectations, clients should be properly trained, as should all those persons who will be involved in implementing the other eleven roles. Perhaps the most important mission of evaluation training programs, both preservice and in-service, is to prepare clients of evaluation to carry out their evaluation responsibilities. If they do not have a sound concept of evaluation, they will likely avoid it, be passive in their associations with it, not direct it to serve their purposes, or find its products of little interest. But as knowledgeable clients, they can help make evaluation a powerful and useful tool for ensuring the quality and ongoing improvement of professional services. This point is so important that we believe evaluation plans and budgets often should provide for the evaluator to train the client and other stakeholders in the concepts, uses, and proper conduct of evaluation.

Almost without exception, evaluations require review panels during the course of a study. It is essential that clients (and stakeholders) are included in these panels for three main reasons. First, they gain insight into the direction that the evaluation is taking, together with some knowledge of ultimate conclusions. Second, the more knowledge they have about the potential outcomes of the evaluation, the more likely it is that the organization or program will benefit from better-quality services. Finally, by supplying valuable critiques of evaluation plans and draft reports, such panels can contribute to the evaluation's quality, clarity, and relevance.

## Evaluation Designer

One of the most fascinating aspects of evaluation work is that it should be viewed as a creative enterprise. Designing an evaluation to respond to a client's information requirements is far from a routine mechanical process. Among other responsibilities, one must conceptualize the questions to be addressed, size up the political scene, lay out a data collection and analysis plan, design the reports and the reporting process, project the staffing and financial requirements, determine how the standards of the evaluation profession will be met, and provide for giving clients and stakeholders a proper measure of evaluation training.

No models of past evaluations or published designs serve as all-purpose solutions to this complex set of planning decisions. On the contrary, evaluation design typically is a continuing conceptual process. To design a good evaluation, one should be a sensitive observer of social settings, a discerning interviewer and listener, and a willing and able learner about the substance of the program to be examined. The evaluator should also be knowledgeable about a wide range of inquiry and communication techniques and be able to draw together diverse pieces of specifications and form them into coherent guiding rationales and practical work plans.

Next to the role of the evaluation client, we see the role of evaluation designer as the most crucial in evaluation work. Those who become proficient in evaluation design are often employed full time as evaluators and are in great demand as consultants to assist evaluation clients in designing and commissioning evaluation studies.

The general evaluation conceptualizations presented in Parts Two and Three of this book provide crucial study material for those who wish to increase their facility in conceptualizing evaluation plans. The chapters in these parts of the book describe the general approaches developed by some of the evaluators who have become widely recognized as experts in designing evaluations. Beyond studying the creative contributions of others, we emphasize that development of one's evaluation design capabilities requires practice—one needs to be involved in designing a wide range of evaluations. This practice, we think, should include some individual work with a client. In addition, students can learn much by serving on evaluation planning teams. Moreover, the study of a wide range of evaluation designs prepared by others provides many good examples and insights into this complex work area. We believe that programs to train evaluators should include practicums and other forms of guided practice in the design of evaluations. And evaluators can increase their evaluation design skills by subjecting their evaluation designs to scrutiny by evaluation experts.

## Evaluation Coordinator

Many evaluations draw on the efforts of a number of participants, the use of considerable resources, and services to multiple clients. Hence, a need for considerable coordination of the work effort exists. This need is magnified when an organization has a number of evaluation projects under way. Not surprisingly, the profession includes many evaluation jobs with titles such as assistant superintendent for evaluation, coordinator of evaluation projects, director of the evaluation office, and staff director of the evaluation project.

One implication of this role is that those specializing in evaluation should obtain training in management functions. Another is that those who are assigned to manage evaluation work should possess or acquire basic proficiency in evaluation design. Finally, we believe that major education, health, and social institutions should often organize their evaluation efforts into a well-managed unit of the institution.

## Evaluation Caseworkers

Frequently, a crucial, and often separable, role in evaluation is that of evaluation caseworker or fieldworker. This role involves periodic interactions with clients, review of program documents, observation of program activities, interviewing and otherwise gathering information from participants, drafting reports, and assuming primary responsibility for writing up the findings. Sometimes the caseworker is in charge of implementing the complete study design. In other situations, he or she has a narrower responsibility. But this person is vitally involved in interacting with program personnel and helping to carry out the study design. Both technical and human relations skills are required to perform the caseworker role. Moreover, the person assigned to this role should possess or be able to develop credibility with the clients and with those who are carrying out the program.

## Evaluation Respondents

A role too often taken for granted is that of the evaluation respondent. Evaluation respondents are the people who fill out the forms, answer the test questions, respond to interview questions, submit their work products, and allow their work to be observed. In many cases, the quality of evaluation information obtained is heavily dependent on the respondents' willingness to cooperate and give their best effort. They are unlikely to do either of these tasks well if they are not informed about their role, promised at least an outline of outcomes, and convinced of the

importance of the study. Consequently, those in charge of evaluations should make special efforts to ensure that the study will be of some value to the respondents as well as to other clients. They should also be careful to provide activities by which to motivate and train respondents to respond appropriately.

### **Technical Support Specialists**

Most evaluations involve a certain amount of technical work that requires specialized expertise. For example, there may be a need to conduct and record the results of interviews; develop, administer, and score structured data collection instruments; analyze qualitative and quantitative information; set up and manage databases; and produce audiovisual presentations of the findings. In small-scale evaluations, a single evaluation generalist often has to do all such tasks without specialized assistance. The single evaluation agent is hard-pressed to coordinate and do the necessary fieldwork while also fulfilling all required technical specialties. This evaluation agent needs broad-gauged training in the wide range of relevant specialties. Often he or she must also be able to engage the assistance of qualified consultants.

Inevitably, large-scale studies need to engage specialists to perform these tasks so that the evaluation caseworkers can concentrate their efforts on fieldwork with the evaluation clients and evaluation respondents. This group of experts could include test development specialists, sampling specialists, computer specialists, statisticians, case study specialists, and technical writers.

### **Information Specialist**

One technical role that we have singled out for special attention is the information specialist. This role has an ongoing need for information that pertains to a series of evaluations as opposed to technical problems particular to given studies. The information specialist role is especially important in organizations that maintain evaluation offices. The evaluations conducted by such offices must take into account the institutional setting and its historical context, examine current and past findings to determine trends, and avoid needless duplication of data collection efforts. What is needed, at minimum, is a good system for indexing and storing evaluation reports so that the information they contain can be kept secure and easily accessed in future evaluations. In addition, many organizations would benefit by regularly collecting standardized information to help them track their inputs, processes, and outputs over time, as well as the attitudes of their constituents toward the organization's services.

Fulfillment of the information specialist role requires that the evaluator, or one or more members of the evaluation team, be proficient in information sciences and computer technology. Such proficiency requires knowledge and skills in such areas as communication, computer information systems, database design, computer programming, data processing, financial accounting, and long-range planning. Those who plan to specialize in evaluation should develop a general command of the information specialties, because a certain amount of work in this area is required even in a single-person evaluation. In addition, offices of evaluation often should assign one or more persons to fulfill this important function.

### Communication Specialist

All too often effective communication is missing in evaluation work. Evaluators and clients assume that evaluators know their business and somehow will obtain and report the right message; instead, evaluators often address the wrong questions or communicate the findings ineffectively, or both. We wish to make emphatically clear that effective evaluation includes not only collecting and analyzing appropriate information, but also ensuring through effective communication technology that the information will be presented in a useful form and subsequently used. Evaluators should not be held accountable for all misinterpretations and misuses of their reports, but, among other things, they should be held accountable for ensuring that their reports are relevant to the interests of audiences and communicated clearly.

To do so requires that evaluators be skilled in communication technology. They should be able to prepare clearly written reports that are directed to, and understandable by, the intended audience or audiences. Beyond being good writers, they need to be skilled in audience analysis and designing different reports for different audiences. They need facility in communicating findings through means other than printed material; for example, they need skills in public speaking, use of audiovisual techniques, use of public media, managing group processes, and utilizing Web sites. Finally, they need skills of political analysis and conflict management, since evaluations often are involved in controversy.

The communication specialist role has considerable implications for the training of evaluators, their use of consultants, and the staffing of evaluation offices. Clearly, evaluators should obtain training in a broad range of communication techniques. In their practice, they should, at a minimum, submit their draft reports to editors. And evaluation offices often should employ communication specialists or engage relevant consultants to assist in planning evaluations and presenting the findings. Such specialists also have an important role in evaluating

and helping to improve an evaluation office's performance over time in communicating its reports to its clients.

We acknowledge that evaluations often have to be done by single evaluators. To do their work well, they need to become proficient in the wide range of roles already identified. Probably the best means of encountering and learning how to address the full range of evaluation roles is by conducting a wide range of single-person evaluations under the tutelage of an evaluation expert. Participation in team evaluation efforts also provides excellent learning opportunities.

We have so far considered only evaluation roles that pertain to an actual evaluation. Now we turn to four others that serve in support roles in developing the evaluation profession.

### **Evaluation Trainer**

Throughout the preceding analysis, we have referred to the training needs of those who participated in evaluation. These needs encompass orientation to specific evaluation plans as well as the development of more generalized understandings and skills, and they include understandings that are acquired through examining one's practice. Moreover, the training needs pertain to the clients and respondents in evaluations, as well as to the evaluation specialists.

Evaluation, like any other profession, should ensure that the participants in evaluation are provided a wide range of sound evaluation learning opportunities. These include specialized training programs leading to graduate degrees in evaluation; service courses that are provided to assist those who are in training for a wide range of professional roles (such as superintendent, social worker, teacher, and physician) to evaluate their professional services; continuing education opportunities in evaluation for both evaluation specialists and generalists; and orientation sessions for the clients and respondents in specific evaluation studies. Such training should present up-to-date content that is relevant to the field problems that the trainees face. Clearly, evaluation training is a crucial role in maintaining and advancing quality evaluation services.

### **Evaluation Researcher**

The mission of any profession is to serve clients as well as possible, and the fulfillment of this commitment is always limited by the current state of knowledge about the field. Therefore, professions need a research ethos. Professionals must study the profession-related needs and problems of their clients, study their services to clients, examine their practices in the light of knowledge and principles

from relevant disciplines, examine their work in its historical context, theorize about the professional services, and move knowledge about their field forward through structured examination of the field's guiding assumptions and hypotheses. Chapter Two examines some of these issues.

Vast research programs are evident in mature professions such as medicine and law and in business and industry. In the former case, such research is driven by the professional ethic of delivering the best possible service. In the latter case, the primary motivation is profit and survival in a competitive world. Both areas are strongly aware that improved service and competitiveness depend heavily on a dynamic program of pertinent research.

Although evaluation is as old as civilization, it has existed as a formal field of practice for a very short time. Only since the early 1970s have serious efforts been made to record the history of the field, theorize about evaluation practice, and conduct pertinent descriptive and hypothesis-testing studies. But there is and should be a strong trend toward establishing and maintaining a research base for the field. This trend is especially evident in the exploding literature of the evaluation field, although much of this material does not qualify as research. In any case, we wish to emphasize that research on evaluation is a vital role in ensuring that evaluators progressively improve their services and impacts.

## Evaluation Developer

The role of evaluation developer has at its base the function of helping evaluators collectively attain and maintain the status of a profession. What is at issue in this role can be seen in a dictionary definition that characterizes a profession as “a calling requiring specialized knowledge and often long and intensive preparation including instruction in skills and methods as well as in the scientific, historical, or scholarly principles underlying such skills and methods, maintaining by force of organization or concerted opinion high standards of achievement and conduct, and committing its members to continued study and to a kind of work which has as its prime purpose the rendering of a public service” (*Webster's Third New International Dictionary*, 1966, p. 1811).

The evaluation field is beginning to manifest the characteristics of a profession that are contained in this definition. There are now training programs leading to master's and doctoral degrees in evaluation, as well as an extensive and burgeoning literature on evaluation. The evaluation literature prescribes content and experiences to be included in evaluation training programs and analyzes how the programs should draw from related disciplines, such as philosophy, psychology, sociology, and economics. Several professional organizations and sets of guiding principles and professional standards are available to evaluators. And

government, the public, and service organizations are continuing to demand evaluations of professional services, as well as provide funds to support evaluations.

## Metaevaluator

The final role we wish to emphasize is the overarching and pervasive role of metaevaluator. Involved with evaluating evaluation, the role extends to assessing the worth, merit, and probity of all that the profession is and does, evaluation services, use of evaluations, evaluation training, evaluation research, and organizational development. Metaevaluation invokes the accepted guiding principles and standards of the profession and assesses and tries to ensure that they are met. It also is involved when the standards themselves are examined to identify where they might need to be revised. In one way or another, metaevaluation needs to be incorporated into each activity of the evaluation profession. Sometimes it is a self-assessment activity, as when evaluators, evaluation trainers, evaluation researchers, or officials of evaluation societies scrutinize their plans, work efforts, and products against the standards of the evaluation profession and provide written attestations. In other cases, metaevaluation involves engaging an independent agent to assess professional evaluation services. The aims of metaevaluation are to ensure quality evaluation services, guard against or uncover malpractice or services not in the public interest, provide direction for improving the evaluation profession, and promote increased understanding of the evaluation enterprise. Clearly, all professional evaluators should maintain up-to-date knowledge of the field's standards, incorporate them in their work, be willing to serve as metaevaluator in relation to the work of others in the profession, and be proactive in evaluating and trying to improve the contributions of evaluation to society.

## Summing Up the Roles

With this brief discussion of metaevaluation we conclude our overview of evaluation roles. In a sense, we have dissected the role of the professional evaluator and have looked at its constituent parts and support groups. Rightly, we think, this has supported the impression that evaluation is often a team effort. But to conclude that evaluations must involve team efforts would be a misinterpretation. On the contrary, evaluation is necessarily an integral part of every professional's role. Our message is that this role is as difficult as it is important. The obvious fundamental conclusion is that all professionals should develop the capability to meet their evaluation responsibilities well, whether or not they can draw assistance from a team or must perform all of the evaluation functions individually.

One basic step in the right direction is to increase one's facility to conceptualize evaluation problems and approaches in order to facilitate the solution of

those problems. The remainder of this book has been prepared to guide this process. It discusses the need for a foundation of evaluation theory and standards, presents and examines in depth a number of alternative evaluation approaches, and provides guidelines for planning and carrying out sound evaluations.

---

## Review Questions

1. Describe an example, drawn from your experience, of how members of society were put at risk or harmed due to a lack of evaluation.
2. Describe an example of how members of society were put at risk or harmed due to a failure to heed and act on the findings of an evaluation.
3. Identify a job in society that should be staffed with an evaluation-oriented decision maker, and note some of the positive benefits associated with the person's possession of evaluation competence.
4. Explain and give examples of evaluation's ubiquitous place in society and its institutions, including its symbiotic relationship to other fields.
5. Cite what you see as the pros and cons of defining evaluation as the comparison of outcomes to objectives.
6. Cite some reasons that evaluators should search for side effects.
7. List the three definitions of evaluation given in the chapter. Then define the intended use of each definition. Subsequently, write an example of how you might use each definition.
8. Define what is meant by *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 context bounded.
9. Give examples of cases that require comparative evaluations and other cases that require noncomparative evaluations.
10. Compare and contrast the terms *formative evaluation* and *summative evaluation*. What points distinguish informal evaluation from formal evaluation? Is there necessarily any connection between these concepts and formative and summative evaluations?

---

## Group Exercises

This section is particularly relevant to group discussion where problem solving is required. It is designed to extend your insight into some of the issues outlined in this chapter.

*Exercise 1.* The head of a large state government department found himself under political pressure to commission an evaluation of each of the four divisions of his department. None of these divisions had ever been evaluated except in the most cursory fashion, and then only sporadically. What was evident to stakeholders (the public) was that services were costly but inadequate and that the poor quality of delivery was causing growing frustration.

Realistic financial provision and time lines were made available for this major evaluation, according to the head of the department.

Suppose your group was selected to conduct the evaluation. Outline the important early decisions you would need to make about the form of the evaluation; the kinds of initial understandings you would wish to have with the head of the department and division heads; and the kinds of assurances you would seek and give so that a successful evaluation could eventuate.

*Exercise 2.* A superintendent of a small school district was beset with problems relating to the introduction of a new state-mandated science program for grades 7 through 9. She had heard of both formative and summative evaluation processes, but had 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 evaluations. Outline the relevance to the superintendent's problems of either form of evaluation, 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 complete a successful study.

What advice do you give the superintendent?

---

## Notes

1. The Joint Committee on Standards for Educational Evaluation 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, and the second author assisted in the development of the original set of program evaluation standards.
2. This history of educational evaluation section is based on a previous account by George Madaus and Daniel Stufflebeam (Stufflebeam, Madaus, & Kellaghan, 2000, pp. 3–18).

## References

- Alkin, M. C., Daillak, R., & White, P. (1979). *Using evaluations: Does evaluation make a difference?* Thousand Oaks, CA: Sage.
- American Evaluation Association Ethics Committee. (2004). *Guiding principles for evaluators*. <http://www.eval.org/Guiding%20Principles.htm>.
- American Evaluation Association Task Force on High Stakes Testing. (2002). *Position statement on high stakes testing in pre-K-12 education*. Louisville, KY: Author.
- Cook, D. L. (1966). *Program evaluation and review technique: Applications in education*. Washington, DC: Government Printing Office.
- Cronbach, L. J. (1963). Course improvement through evaluation. *Teachers College Record*, 64, 672–683.
- Eisner, E. W. (1975, December). *The perceptive eye: Toward the reformation of educational evaluation*. Stanford, CA: Stanford Evaluation Consortium.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, 18, 519–521.
- Government Accountability Office. (2003, June). *Government auditing standards*. Washington, DC: Author. (GAO-03-673G)
- Guba, E. G. (1966, October). *A study of title III activities: Report on evaluation*. Paper presented at the National Institute for the Study of Educational Change, Indiana University.
- Hammond, R. L. (1967). *Evaluation at the local level*. Address to the Miller Committee for the National Study of ESEA Title III, Washington, DC.
- Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York: McGraw-Hill.
- Joint Committee on Standards for Educational Evaluation. (1988). *The personnel evaluation standards*. Thousand Oaks, CA: Corwin Press.
- Joint Committee on Standards for Educational Evaluation. (1994). *The program evaluation standards*. Thousand Oaks, CA: Corwin Press.
- Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards*. Thousand Oaks, CA: Corwin Press.
- Jones, L. V. (2003). National assessment in the United States: The evolution of a nation's report card. In T. Kellaghan & D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 883–904). Norwell, MA: Kluwer.
- Kellaghan, T. (1982). Los sistemas escalates coma objecto de evaluacion. In D. Stufflebeam, T. Kellaghan, & B. Alvarez (Eds.), *La evaluacion educativa*. Bogota: Pontificia Universidad Javeriana.
- Kellaghan, T., Madaus, G., & Airasian, P. (1982). *The effects of standardized testing*. Norwell, MA: Kluwer.
- Kellaghan, T., & Stufflebeam, D. L. (Eds.). (2003). *International handbook of educational evaluation*. Norwell, MA: Kluwer.
- Madaus, G. F. (2004). Ralph W. Tyler's contribution to program evaluation. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.
- Madaus, G. F., & Stufflebeam, D. L. (1988). *Educational evaluation: The classical writings of Ralph W. Tyler*. Norwell, MA: Kluwer.
- Metfessel, N. S., & Michael, W. B. (1967). A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs. *Educational and Psychological Measurement*, 27, 931–943.

- Patton, M. Q. (1997). *Utilization-focused evaluation: The new century text* (3rd ed.). Thousand Oaks, CA: Sage.
- Popham, W. J. (1971). *Criterion-referenced measurement*. Upper Saddle River, NJ: Educational Technology Publications.
- Provus, M. (1969). *Discrepancy evaluation model*. Pittsburgh, PA: Pittsburgh Public Schools.
- Sanders, W. L., & Horn, S. (1994). The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, 8(3) 299–311.
- Scriven, M. S. (1967). The methodology of evaluation. In *Perspectives of curriculum evaluation*. Skokie, IL: Rand McNally.
- Scriven, M. S. (1974). Pros and cons about goal-free evaluation. *Evaluation Comment*, 3, 1–4.
- Scriven, M. S. (1975). *Evaluation bias and its control*. Kalamazoo: Western Michigan University, The Evaluation Center.
- Scriven, M. (1981). *Evaluation thesaurus* (3rd Ed.). Point Reyes, CA: Edgepress.
- Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Thousand Oaks, CA: Sage.
- Smith, E. R., & Tyler, R. W. (1942). *Appraising and recording student progress*. New York: HarperCollins.
- Stake, R. E. (1967). The countenance of educational evaluation. *Teachers College Record*, 68, 523–540.
- Stake, R. E. (1969). Evaluation design, instrumentation, data collection, and analysis of data. In J. L. Davis (Ed.), *Educational evaluation*. Columbus, OH: State Superintendent of Public Instruction.
- Stufflebeam, D. L. (1966). A depth study of the evaluation requirement. *Theory into Practice*, 5(3), 121–133.
- Stufflebeam, D. L. (1967). The use and abuse of evaluation in Title III. *Theory into Practice*, 6, 126–133.
- Stufflebeam, D. L. (1971). The relevance of the CIPP evaluation model for educational accountability. *Journal of Research and Development in Education*, 5(1), 19–25.
- Stufflebeam, D. L. (1978). Metaevaluation: An overview. *Evaluation and the Health Professions*, 1(2), 146–163.
- Stufflebeam, D. L. (2001). The metaevaluation imperative. *American Journal of Evaluation*, 22(2), 183–209.
- Stufflebeam, D. L., Foley, W. J., Gephart, W. J., Guba, E. G., Hammond, R. L. Merriman, H. O., & Provus, M. M. (1971). *Educational evaluation and decision making in education*. Itasca, IL: Peacock.
- Stufflebeam, D. L., Madaus, G. F., & Kellaghan, T. (2000). *Evaluation models: Viewpoints on educational and human services evaluation*. Norwell, MA: Kluwer.
- Stufflebeam, D. L., McCormick, C. H., Brinkerhoff, R. O., & Nelson, C. O. (1985). *Conducting educational needs assessment*. Norwell, MA: Kluwer.
- Travers, R.M.W. (1983). *How research has changed American schools*. Kalamazoo, MI: Mythos Press.
- Tyler, R. W. (1967). Changing concepts of educational evaluation. In R. E. Stake (Ed.), *Perspectives of curriculum evaluation*. Skokie, IL: Rand McNally.
- U.S. General Accounting Office. (2002, January). *Government auditing standards: Amendment No. 3 Independence*. Washington, DC: Author.

- U.S. General Accounting Office. (2003, June). *Government auditing standards: 2003 revision*. Washington, DC: Author.
- Walton, M. (1986). *The Deming management method*. New York: Putnam.
- Webster's Third New International Dictionary*. (1966). Chicago: Encyclopedia Britannica.

