Educational programs (and other publicly funded programs) have continued to increase in size and expense. Not surprisingly, taxpayers and public officials have increasingly urged that these programs be made more accountable to their publics. Indeed, "accountability" for expenditures of public funds has become the hue and cry of an ever-increasing number of economy-minded social reformers. In several countries, policy makers at both national and local levels now routinely authorize funds to be used for the express purpose of evaluating educational programs to determine their effectiveness. Thus, "program evaluation" has come into being as both a formal educational activity and as a frequently mandated instrument of public policy. Many private educational enterprises have similarly turned to program evaluation as a means of answering questions about the benefits received from monies expended on various educational programs.

To define program evaluation, it is necessary to define its component parts. In an educational context, a program can be thought of as any educational enterprise aimed at the solution of a particular educational problem or the improvement of some aspect of an educational system. Such a program would typically be sponsored by public or private funds, possess specified goals, and exhibit some structure for managing the procedures, materials, facilities, and/or personnel involved in the program.

Evaluation can be defined most simply as the determination of the worth of a thing. In its simplest form, therefore, program evaluation consists of those activities undertaken to judge the worth or utility of a program (or alternative programs) in improving some specified aspect of an educational system. Examples of program evaluations might include evaluation of a national bilingual education program, a university's preservice program for training urban administrators, a ministry of education's staff development program, or a local parent education resource center. Evaluations may be conducted for programs of any size or scope, ranging from an arithmetic program in a particular school to an international consortium on metric education.

A curriculum evaluation may qualify as a program evaluation if the curriculum is focused on change or improvement, as implied in the previous definition of "program." Program evaluations, however, often do not involve appraisal of curricula (e.g., evaluation of a computerized student recordkeeping system or evaluation of the extent to which funds from a national program for the hearing impaired are actually used to provide services to children with hearing impairments). For this reason, the closely related but more specialized topic of curriculum evaluation is not discussed further in this section (see Curriculum Evaluation).

1. Purposes of Program Evaluation

Most program evaluators agree that program evaluation can play either a formative purpose (helping to improve the program) or a summative purpose (deciding whether a program should be continued). Anderson and Ball (1978) further describe the capabilities of program evaluation in terms of six major purposes (which are not necessarily mutually exclusive). They are:

(a) to contribute to decisions about program installation;
(b) to contribute to decisions about program continuation, expansion, or "certification";
(c) to contribute to decisions about program modifications;
(d) to obtain evidence to rally support for a program;
(e) to obtain evidence to rally opposition to a program;
(f) to contribute to the understanding of basic psychological, social, and other processes (only rarely can this purpose be achieved in a program evaluation without compromising more basic evaluation purposes).

2. The History of Program Evaluation

The informal practice of program evaluation is not new, dating back at least to 2000 BC, when Chinese officials were conducting civil-service examinations, and con-
tinuing down through the centuries to the beginnings of school accreditation in the late 1800s. The first clear evidence of formal program evaluation, however, appears to be Joseph Rice's 1897-1898 comparative study of spelling performance of 33,000 students in a large United States school system. Few formal evaluations of educational programs were conducted in the next few decades, with Tyler and Smith's Eight-year Study of the 1930s being the next notable effort to evaluate the outcomes of an educational program. During the late 1950s and early 1960s (the post-Sputnik years), cries for curriculum reform led to major new curriculum development programs and to subsequent calls for their evaluation. The relatively few evaluation studies that resulted revealed the conceptual and methodological impoverishment of the field—or perhaps more accurately, the "nonfield"—of evaluation in that era. In many cases, the designs were inadequate, the data invalid, the analyses inaccurate, and the reports irrelevant to the important evaluation questions which should have been posed. Most of the studies depended on idiosyncratic combinations and applications of concepts and techniques from experimental design, psychometrics, curriculum development and, to a lesser extent, survey research. Theoretical work related to educational evaluation, per se, was almost nonexistent. Few scholars had yet turned their attention to the development of generalizable evaluation plans which could be adopted or adapted specifically to educational evaluation studies. In the absence of a "formal subject matter" or educational evaluation, evaluators of educational programs were left to glean what they could from other fields to help them in their work.

Since a large number of persons serving in evaluation roles during the late 1950s and 1960s were educational and psychological researchers, it is not surprising that the experimental tradition quickly became the most generally accepted evaluation approach. The work of Campbell and Stanley gave enormous impetus to predominance of experimental or quasi-experimental approaches to program evaluation. Although some evaluators cautioned that correct use of the experimental model may not be feasible, the elegance and precision of this model led most program evaluators to view the experimental method as the ideal model for program evaluations.

Not all program evaluators were enamored with the use of traditional quantitative methods for program evaluations, however, and their dissatisfaction led to a search for alternatives. Qualitative and naturalistic methods, largely shunned by program evaluators during the 1960s as unacceptably "soft," gained wider acceptance in the 1970s and thereafter as proposals for their application to program evaluations were made by Parlett and Hamilton, Stake, Eisner, Guba and Lincoln, and others. Sharp disagreements developed between proponents of the newer qualitative approaches and adherents to the more broadly accepted quantitative methods and the 1970s were marked by polemics as the two schools of thought struggled for ascendancy. The late 1970s and the early 1980s saw the dialogue begin to move beyond this debate as analysts accelerated their discussions of the benefits of integrating both types of methods within a program evaluation (for instance see Cook and Reichardt 1979, Worthen 1981, and the especially useful summary by Madey 1982).

Concurrent with program evaluators' struggle to sort out the relative utility of quantitative and qualitative methods, a separate but closely related development was taking place. Beginning in the late 1960s, several evaluation writers began to develop and circulate their notions about how one should conduct educational evaluations; these efforts resulted in several new evaluation "models" being proposed to help the practicing program evaluator. Although these seminal writings in educational evaluation (discussed in Sect. 3) were doubtlessly influenced by the quantitative-qualitative controversy and some proved more comfortable companions with one or the other methodological persuasion, several were broader in conceptualization, providing guidelines for conducting program evaluations that could use either quantitative or qualitative data. As these frameworks for planning evaluation studies were applied and refined, program evaluators began to turn to them as promising sources of guidance. Collectively, these writings, the so-called evaluation models, represent the formal content of program evaluation and are discussed in the following section.

3. Alternative Approaches to Program Evaluation

Because of space restrictions, only some of the more popular current approaches used in conducting program evaluations can be presented in this section. Many of these (and other) approaches to program evaluation are summarized in Worthen and Sanders (1987) and the work of authors mentioned but not referenced herein (and in other sections of this entry) can be found in that source. For convenience, these conceptual frameworks for evaluation are clustered into five categories, although some of the frameworks are sufficiently multifaceted that they could appear in more than one category. Most of these "models" have focused broadly on program evaluation, although some are focused more specifically on curriculum evaluation. It should be noted that these frameworks deal with methods, not techniques: discussion of the many techniques which might be used in program evaluations is beyond the scope of this article.

3.1 Performance-Objectives Congruence Approaches

This approach to program evaluation was originally formulated by Ralph Tyler, who conceived of evaluation as the process of determining the extent to which the educational objectives of a school program or curriculum are actually being attained. He proposed a process in which broad goals or objectives would be
established or identified, defined in behavioral terms, and relevant student behaviors would be measured against this yardstick, using either standardized or evaluator-constructed instruments. These outcome data were then to be compared with the behavioral objectives to determine the extent to which performance was congruent with expectations. Discrepancies between performance and objectives would lead to modifications intended to correct the deficiency, and the evaluation cycle would be repeated.

Tyler's rationale was logical, scientifically acceptable, readily adoptable by program evaluators (most of whose methodological upbringing was very compatible with the pretest–posttest measurement of student behaviors stressed by Tyler), and had great influence on subsequent evaluation theorists. Hammond's EPIC evaluation model followed Tyler's model closely, adding only a useful program-description "cube" which elaborates instructional and institutional variables often overlooked in previous evaluations. Provens' discrepancy model of program evaluation is clearly Tylerian and gains its name from the constant setting and juxtaposition of program standards against program performance to yield "discrepancy information" needed for program improvements. Popham's instructional objectives approach also clearly stems from Tyler's earlier conceptions.

Useful as this approach to evaluation is viewed by its many adherents, critics such as Guba and Lincoln (1981) have noted that it lacks a real evaluative component (facilitating measurement and assessment of objectives rather than resulting in explicit judgments of worth), lacks standards to judge the importance of observed discrepancies between objectives and performance levels, and depends on a highly utilitarian philosophy, promoting a linear, inflexible approach to evaluation.

3.2 Decision-Management Approaches
The most important contributions to a decision-oriented approach to program evaluation are Stufflebeam's Context, Input, Process, and Product (app) evaluation model and Alkin's Center for the Study of Evaluation model, which follows a similar logic to the Context, Input, Process, and Product model but distinguishes between program implementation and program improvement, two subdivisions of what Stufflebeam terms process evaluation. In both models, objectives are eschewed as the organizer for the study and the decision to be made by program managers becomes pivotal. Stufflebeam has provided an analysis of types of programs implemented, program objectives are required to make and proposes a different type of evaluation for each type of decision. In both of these decision-oriented models, the evaluator, working closely with the program manager, would identify the decisions the latter must make and collect sufficient information about the relative advantages and disadvantages of each decision alternative to enable the decision maker to make a judgment about which is best in terms of specified criteria. Thus, evaluation became an explicitly shared function dependent on good teamwork between evaluators and decision makers.

This approach has proved appealing to many evaluators and program managers, particularly those at home with the rational and orderly systems approach, to which it is clearly related. It was viewed by others, however, as failing to determine explicitly the program's worth and being dependent on somewhat unrealistic assumptions about the orderliness and predictability of the decision-making process.

3.3 Judgment-oriented Approaches
This general approach to evaluation, which historically has been the most widely used evaluation approach, is dependent upon experts' application of professional expertise to yield judgments about a program being observed. For example, the worth of a program would be assessed by experts (in the view of the evaluation's sponsor) who would observe the program in action, examine its products or, in some other way, glean sufficient information to render their considered judgment about the program. Site visits initiated by funding agencies to evaluate programs they support and visits by accrediting agencies to secondary schools and universities are examples of judgment-oriented program evaluations. Scriven, in his article The Methodology of Evaluation (Worthen and Sanders 1973), stressed judgment as the sine qua non of evaluation and, in his insightful examination of educational evaluation, did much to rescue this approach from the disrepute into which it had fallen in evaluation's headlong rush to gain respectability as a science. He stunned orthodox objectives-oriented evaluators by his suggestion that evaluators might go beyond measuring a program's performance to also evaluate the program's goals and later compounded the shock still further with his suggestion that evaluators should do "goal-free" evaluations, in which they not only ignore the program's goals but actually make every effort to avoid learning what those goals are. Thus, judgments about programs were based on the actual outcomes of the program, intended or not, rather than on the program's objectives or on decisions faced by program managers.

Another important judgment-oriented evaluation model is Robert Stake's Countenance Model, in which he suggests that the two major activities of formal evaluation studies are description and judgment (the "two countenances") of the program being evaluated. Within the description phase, Stake follows Tyler's rationale of comparing intended and actual outcomes of the program. However, he argued that in the judgment phase standards and procedures for making judgmental statements must be explicated to ensure the publicness of evaluative statements, although he failed to provide any suggestions as to how to weight or combine individual standards into overall judgments about the program.
Eisner's "connoisseurship model" casts program evaluators as educational critics whose refined perceptual capabilities (based on knowledge of what to look for and a backlog of relevant experience) enable them to give a public rendering of the quality and significance of that which is evaluated. In this model, the evaluator is the "instrument," and the data collecting, analyzing, and judging that Stake tried to make more public are largely hidden within the evaluator's mind, analogous to the evaluative processes of art criticism or wine tasting.

Collectively, these judgment-oriented approaches to evaluation have emphasized the central role of judgment and human wisdom in the evaluative process and have focused attention on the important issues of whose standards (and what degree of publicness) should be used in rendering judgments about educational programs. Conversely, critics of this approach suggest that it often permits evaluators to render judgments that reflect little more than figments of fertile imaginations. Others have noted that the presumed expertise of the evaluators is a potential weakness and worse, strong arguments can be made that serious disadvantages can accrue if a program is evaluated only by content experts (Worthen and Sanders 1987). Finally, many program evaluators are disinclined to play the single-handed role of educational judge (which they feel smacks of arrogance and elitism) proposed by some of these approaches.

3.4 Adversarial Approaches
Adversarial evaluation is a rubric that encompasses a collection of divergent evaluation practices which might loosely be referred to as adversarial in nature. In its broad sense, the term refers to all evaluations in which there is planned opposition in the points of view of different evaluators or evaluation teams—a planned effort to generate opposing points of view within the overall evaluation. One evaluator (or team) would serve as the program's advocate, presenting the most positive view of the program possible from the data, while another evaluator (or team) would play an adversarial role, highlighting any extant deficiencies in the program. Incorporation of these opposing views within a single evaluation reflects a conscious effort to assume fairness and balance and illuminate both strengths and weaknesses of the program.

Several types of adversarial proceedings have been invoked as models for adversary evaluations in education, including judicial, congressional hearings, and debate models. Of these, most of the sparse literature in this area has focused on adaptations of the legal paradigm, providing insights into how concepts from the legal system (for instance, taking and cross-examination of human testimony) could be used in educational evaluations. Owens, Wolf, and others have adapted the legal model to educational evaluations, while Worthen and Rogers have described use of the debate model in an adversary evaluation and have discussed pitfalls and potentials of the legal and other forensic paradigms in conducting program evaluations.

Despite the publicity given this approach to evaluation, as yet there is little beyond personal preference to determine whether program evaluations will profit most from being patterned after jury trials, congressional hearings, debates, or other arrangements.

3.5 Pluralist-Intuitionist Approaches
Ernest House has used this descriptor to characterize several evaluation models, contrasting them with more "utilitarian" models. In this approach to evaluation, the evaluator is a portrayer of different values and needs of all the individuals and groups served by the program, weighing and balancing this plurality of judgments and criteria in a largely intuitive fashion. Thus, the "best program" is largely decided by the values and perspectives of whomever is judging (an obvious fact nonetheless ignored in most other evaluation approaches). Examples of pluralist-intuitionist evaluation "models" are those proposed by Stake, Parlett and Hamilton, Rippey, and MacDonald's democratic evaluation. There are unique contributions of each of these proposals. Stake urges program evaluators to respond to the audience's concerns and requirements for information, in terms of their value perspectives, and argues that the evaluation framework and focus should emerge only after considerable interaction with those audiences. Parlett and Hamilton draw on the social anthropology paradigm (and psychiatry and sociology participant observation research) in proposing progressive focusing of an evaluation whose purpose is to describe and interpret (not measure and predict) that which exists within an educational system. Rippey focuses on the effects of programs on the program operators and views evaluation as a strategy for conflict management. MacDonald views evaluation as primarily a political activity whose only justification is the "right to know" of a broad range of audiences. Yet a common thread runs through all these evaluation approaches—value pluralism is recognized, accommodated, and protected, even though the effort to summarize the frequently disparate judgments and preferences of such groups is left as an intuitive process which depends heavily on the sagacity and impartiality of the evaluator.

Critics of this approach to program evaluation discount it as hopelessly "soft headed" and argue that few if any program evaluators are such paragons of virtue and wisdom as to be skillful in wielding the seductively simple, yet slippery and subtle tools this approach requires. Champions of pluralistic, responsive approaches reply that they can be readily used by any sensitive individual and that they are infinitely richer and more powerful than other approaches and, indeed, can subsume them, since they are flexible and do not preclude the use of other approaches within them, should that be desired by the evaluator's sponsor.

Collectively, the writings reviewed briefly in Sect. 3, the so-called evaluation models, represent the formal content on which educational program evaluators draw. It is, therefore, appropriate to ask how useful they are. The answer is "very useful, indeed," even though they collectively have not moved evaluation very far toward becoming a science or discipline in its own right (a dubious aspiration, nonetheless sought by many evaluators). In a recent analysis, Worthen and Sanders (1987) suggested that (a) the so-called evaluation models fail to meet standard criteria for scientific models, or even less rigorous definitions of models, and (b) that which has come to be referred to as the theoretical underpinnings of evaluation lack important characteristics of most theories, being neither axiomatic nor deductive, having no real predictive power, and being untested and unvalidated in any empirical sense. That same analysis, however, suggested that these conceptions about how evaluations should be conducted--the accompanying sets of categories, lists of things to think about, descriptions of different evaluation strategies, and exhortations to which one might choose to attend--influence the practice of program evaluation in sometimes subtle, sometimes direct, but always significant ways. Some program evaluators design evaluations which adopt or adapt proposed models of evaluation. Many evaluators, however, conduct evaluations without strict adherence (or even intentional attention) to any "model" of evaluation, yet draw unconsciously in their evaluation philosophy, plans, and procedures on that which they have internalized through exposure to the literature of program evaluation. So the value of the "models" lies in their ability to help us to think, to provide sources of new ideas and techniques, to serve as mental checklists of things we ought to consider, or remember, or worry about. Their value as prescriptive guidelines for doing evaluation studies seems much less.

5. Impediments to Improving Program Evaluation

Despite the advances made in program evaluation, there is obviously room for a great deal of improvement. In this section, four areas that need improvement for educational evaluation to reach its full potential are discussed briefly.

5.1 Evaluation Lacks an Adequate Knowledge Base

Since the early 1970s, Stufflebeam, Worthen and Sanders, Smith, and others have issued a call for evaluation to be researched to develop an adequate knowledge base to guide evaluation practice. That call is still largely unanswered, despite some promising research which has been launched on evaluation methods and techniques. A program of research aimed at drawing from other disciplines new methodological metaphors and techniques for use in educational evaluation existed at the Northwest Regional Educational Laboratory for nearly a decade and has introduced program evaluators to promising new metaphors and techniques drawn from areas such as architecture, philosphic analysis, investigative journalism, and literary and film criticism. A second National Institute of Education-sponsored research effort at the University of California at Los Angeles focused largely on descriptive studies of evaluation practices in educational agencies. In addition, a few research studies aimed at generating knowledge about either particular evaluation strategies and procedures or factors affecting evaluation utilization have begun to appear.

These positive developments notwithstanding, there is still little empirical information about the relative efficacy of alternative evaluation plans or techniques or many evaluation components germane to almost any model. For example, virtually no empirical information exists about the most effective way to conduct a needs assessment or weight criteria in reaching a summative judgment. Little is known about the extent to which various data collection techniques interfere with ongoing educational phenomena. Techniques for identifying goals are developed anew with every evaluation, since there is no evidence that any one way of conducting these activities is more effective than any other. Elaborate systems are developed for providing evaluative feedback, but there is little research evidence (as opposed to rhetoric and position statements) about the relative effectiveness of feedback under differing conditions and scheduling. One could go on to create an exhaustive list of phenomena and procedures in evaluation which badly need to be systematically studied, but the above should suffice to make the point. Smith (1981) has summarized the needs for research on evaluation as requiring more knowledge about (a) the contexts within which evaluation is practiced, (b) the nature of evaluation utility, and (c) the effectiveness of specific evaluation methods. Nearly a decade later those needs still remain largely unmet.

5.2 Evaluation Studies are Seldom Evaluated

The necessity of "meta-evaluation" has long been apparent to evaluators and completion of the Standards for Evaluations of Educational Programs, Projects, and Materials (Joint Committee on Standards 1981) marked a welcome milestone. Although many evaluation writers had proposed their own sets of meta-evaluation criteria, none carried the profession-wide weight reflected in the comprehensive standards so carefully prepared by the Joint Committee. These standards include criteria within each of the following categories: utility standards; feasibility standards; propriety standards; and accuracy standards.

Despite the wide acceptance and availability of these standards, however, there is no evidence that program evaluations are being subjected to any closer scrutiny than was the case before their publication. Even casual inspection reveals that only a small proportion of eval-
ulation studies are ever evaluated, even in the most per-
functory fashion. Of the few meta-evaluations which do
occur, most are internal evaluations done by the evalua-
tor who produced the evaluation in the first place. It is
rare indeed to see evaluators call in an outside expert to
evaluate their evaluation efforts. Perhaps the reasons are
many and complex why this is so, but one seems particu-
larly compelling—evaluators are human and are no more
ecstatic about having their work critiqued than are pro-
fessionals in other areas of endeavor. Indeed, it can be a
profundely unnerving experience to swallow one's own
prescriptions. Although the infrequency of good meta-
evaluation might thus be understandable, it is not easily
forgivable, for it enables shoddy evaluation practices to
go undetected and worse, to be repeated again and again,
to the detriment of the profession.

5.3 Program Evaluators Fail to Understand the Political
Nature of Evaluation

Cronbach and co-workers (1980) have presented the
view that evaluation is essentially a political activity.
They describe evaluation as a "novel political institu-
tion" that is part of governance of social programs.
They assert that evaluators and their patrons pursue un-
realistic goals of finding "truth" or facilitating "right"
decisions, rather than the more pertinent task of simply
enlightening all participants so as to facilitate a demo-
cratic, pluralist decision-making process. While some
may reject this view as overstated, it underscores the fact
that program evaluation is inextricably intertwined with
public policy formulation and all of the political forces
involved in that process. Evaluators who fail to under-
stand this basic fact expend unacceptably large amounts
of human and financial resources conducting evaluations
that are largely irrelevant, however impeccably they are
designed and conducted.

5.4 Program Evaluators are too Narrow in their Choice
of Evaluation Approaches and Techniques

It may be that innocence about the political nature of the
evaluation enterprise contributes to the naive hope that
evaluation will one day grow into a scientific discipline.
That day, if attainable, would seem far off. Education
itself is not a discipline but rather a social process field
which draws its content from several disciplines. It
seems unlikely that educational program evaluation,
which also borrows its methods and techniques from
many disciplines will gain the coherence that would re-
sult in it becoming a discipline in its own right. Perhaps
that is just as well, for much of the richness and potential
of educational program evaluation lies in the depth and
breadth of the strategies and tools it can employ and in
the possibility of selectively combining them into
stronger approaches than when used singly (Worthen
1981). Yet eclectic use of the evaluator's tools is a la-

mentably infrequent occurrence in program evaluations.
Disciple-prone evaluators tend to cluster around their
respective evaluation banners like vassals in a form of
provincial bondage. For program evaluation to reach its
potential, such intellectual bondage must give way to
more mature and sophisticated approaches that draw ap-
propriately on the richness and diversity of the many
approaches, models, and techniques that characterize
program evaluation today.

Bibliography

Anderson S B, Ball S 1978 The Profession and Practice
of Program Evaluation. Jossey-Bass. San Francisco,
California.

Cook T D- Reihehardt C T 1979 Qualitative and Quantita-
tive Methods in Evaluation Research. Sage, Beverly
Hills, California

Cronbach L J, Ambron S, Dornbusch S, Hess R, Hornik
R, Phillips D, Walker D, Weiner S 1980 Toward Re-
form of Program Evaluation: Aims, Methods and In-
stitutional Arrangements. Jossey-Bass, San Fran-
cisco, California

Guba E G, Lincoln Y S 1981 Effective Evaluation: Im-
proving the Usefulness of Evaluation Results
through Responsive and Naturalistic Approaches.
Jossey-Bass, San Francisco, California

Joint Committee on Standards for Educational Evalua-
tion 1981 Standards for Evaluations of Educational
Programs, Projects, and Materials. McGraw-Hill,
New York

Madey D L 1982 Some benefits of integrating qualita-
tive and quantitative methods in program evaluation,
4(2): 223-36

Smith N L 1981 Developing evaluation methods. In:
Smith N L (ed.) 1981 Metaphors for Evaluation:
Sources for New Methods. Sage, Beverly Hills, Cali-
fornia

Worthen B R 1981 Journal entries of an eclectic evalua-
Strategies for Curriculum Evaluation. Association
for Supervision and Curriculum Development, Al-
exandria, Virginia

Worthen B R, Sanders J R 1973 Educational Evaluation:
Theory and Practice. Jones, Worthington, Ohio

Worthen B R, Sanders J R 1984 Content specialization
and educational evaluation: A necessary mar-
University Evaluation Center, Kalamazoo, Michigan

Worthen B R, Sanders J R 1987 Educational Evalua-
tion: Alternative Approaches and Practical Guide-
lines. Longman, New York