Medical Teacher, Vol. 28, No. 3, 2006, pp. 210–224

Taylor & Francis
Taylor & Francis Group

AMEE Education Guide no. 29: Evaluating educational programmes

JOHN GOLDIE
Department of General Practice, University of Glasgow, UK

ABSTRACT *Evaluation has become an applied science in its own right in the last 40 years. This guide reviews the history of programme evaluation through its initial concern with methodology, giving way to concern with the context of evaluation practice and into the challenge of fitting evaluation results into highly politicized and decentralized systems. It provides a framework for potential evaluators considering undertaking evaluation. The role of the evaluator; the ethics of evaluation; choosing the questions to be asked; evaluation design, including the dimensions of evaluation and the range of evaluation approaches available to guide evaluators; interpreting and disseminating the findings; and influencing decision making are covered.*

Introduction

Evaluation is integral to the implementation and development of educational activities, whether national programmes, an individual school's curriculum or a piece of work undertaken by a teacher with his/her students. In the UK the focus of evaluation in mainstream general education has seen a recent shift from development towards teacher accountability and appraisal (Kelly, 1989). While this has not occurred in higher education the possibility exists.

This guide aims to provide teachers and researchers in medical education with a framework for undertaking evaluations. While the focus is on evaluation in the context of medical education, the underpinning theories of evaluation draw on experience from the wider educational field and from other disciplines. Any examination of evaluation, therefore, requires consideration of the wider literature.

What is evaluation?

Evaluation is defined in the *Collins English Dictionary* (1994) as ''the act of judgement of the worth of...''. As such it is an inherently value-laden activity. However, early evaluators paid little attention to values, perhaps because they naively believed their activities could, and should, be value free (Scriven, 1983). The purpose(s) of any scheme of evaluation often vary according to the aims, views and beliefs of the person or persons making the evaluation. Experience has shown it is impossible to make choices in the political world of social programming without values becoming important in choices regarding evaluative criteria, performance standards, or criteria weightings (Shadish *et al.*, 1991). The values of the evaluator are often reflected in some of the definitions of evaluation which have emerged, definitions that have also been influenced by the context in which the evaluator operated. Gronlund (1976), influenced by Tyler's goal-based conception of evaluation, described it as ''the systematic process of determining the extent to which instructional objectives are achieved''. Cronbach (Cronbach *et al.*, 1980), through reflection on the wider field of evaluation and influenced by his view of evaluators as educators, defined evaluation as ''an examination conducted to assist in improving a programme and other programmes having the same general purpose''.

In education the term evaluation is often used interchangeably with assessment, particularly in North America. While assessment is primarily concerned with the measurement of student performance, evaluation is generally understood to refer to the process of obtaining information about a course or programme of teaching for subsequent judgement and decision-making (Newble & Cannon, 1994). Mehrens (1991) identified two of the purposes of assessment as:

1. to evaluate the teaching methods used;
2. to evaluate the effectiveness of the course.

Assessment can, therefore, be looked upon as a subset of evaluation, its results potentially being used as a source of information about the programme. Indeed student gain by testing is a widely used evaluation method, although it requires student testing both pre- and post-course.

History of evaluation

Planned social evaluation has been noted as early as 2200 BC, with personnel selection in China (Guba & Lincoln, 1981). Evaluations during the last 200 years have also been chronicled (Cronbach *et al.*, 1980; Madaus *et al.*, 1983; Rossi & Freeman, 1985). Modern evaluation theories and practices, however, have their intellectual roots in the work of Tyler (1935) in education, Lewin (1948) in social psychology, and Lazarfield (Lazarsfeld & Rosenberg, 1955) in sociology.

The main stimulus to the development of modern evaluation theories and practices, however, was the post-Second World War rapid economic growth in the Western world, particularly the United States, and the interventionist role taken by governments in social policy during the 1960s. With the increasing amounts of money being spent on social programmes there was the growing recognition that these programmes required proper evaluation, and mandatory evaluation was introduced.

Correspondence: Dr John Goldie, Department of General Practice and Primary Care, Community Based Sciences, University of Glasgow, 1 Horselethill Road, Glasgow G12 9LX, UK. Tel: 0141 330 8330; fax: 0141 330 8332; email: johngoldie@fsmail.net

ISSN 0142–159X print/ISSN 1466–187X online/06/030210-15 © 2006 Taylor & Francis
DOI: 10.1080/01421590500271282

At the same time there were a growing number of social science graduates who became interested in policy analysis, survey research, field experiments and ethnography, and turned their attention towards evaluation (Shadish *et al.*, 1991).

The earliest evaluation theorists, with little experience to reflect on, concentrated on methodology. They concentrated on testing the incorporation of new ideas into existing or new programmes. There was little reflection on the politics of how the methods could be applied to field settings. Reflection on increasing experience led to the diversification and change of evaluation theories and practice. There was no longer an exclusive reliance on comparative, outcome studies. The quality of programme implementation and the causal processes mediating programme impact were also considered (Sechrest *et al.*, 1979). This resulted in the greater use of qualitative methods in evaluation (Guba & Lincoln, 1981). Using policy-makers as both the source of evaluation questions and the audience for results gave way to consideration of multiple stakeholder groups (Weiss, 1983a, 1983b) Concern with methodology gave way to concern with the context of evaluation practice, and into fitting evaluation results into highly politicized and decentralized systems (Cronbach *et al.*, 1980). The development of evaluation theories has meant that they cover a wider field, have a better sense of the complexities that plague evaluation practice, and better integrate the diverse concepts, methods and practices that span the field (Cronbach, 1982a; Rossi & Freeman, 1985).

In education, evaluation for centuries had been mainly equated with student testing (Popham, 1988). Tyler, following his work with the Eight Year Study, a formal appraisal of the college performance of students prepared in 'progressive' high schools compared with the performance of students prepared in more conventional schools, came to view evaluation not as the assessment of students, but rather as the appraisal of an educational programme's quality (Tyler, 1949). He argued that a programme should be judged on the extent to which students obtained mastery of the programme's pre-stated objectives. Tyler's work, together with that of Bloom (1956) and Taba (1962) led to the development of the linear, hierarchical, objectives model of curriculum planning, with its structure of aims-learning experiences–content–organization of learning-evaluation. This 'industrial' approach to curriculum planning influenced many of the attempts at curriculum evaluation in the 1960s, and also influenced the development of formal evaluation strategies (Holt, 1981).

Cronbach, in his 1963 article 'Course improvement through evaluation', responding to the dissatisfaction felt by curriculum development staff who were finding little virtue in the existing methods for determining the effectiveness of their instructional materials, argued that if evaluation was to be of value to curriculum developers it should focus on the decisions they faced during the development phase of their curricula. He also argued that evaluation should deal less with comparisons between programmes, and more with the degree to which the programme promoted its desired purpose(s). He also stressed the importance of evaluation in helping refine a course when it was still sufficiently 'fluid' to make changes. His views failed to attract widespread interest outside the field of curriculum developers due to the lack of interest on the part of educators per se (Popham, 1988).

As with social programming in general, with the increasing sums being spent on educational programmes in the United States mandatory evaluation was introduced. The requirement for mandatory evaluation of curriculum innovation crossed the Atlantic and formal evaluation was made an essential requirement of all curriculum projects by such funding bodies as the Schools Council (Kelly, 1989). This stimulated activity in the field of educational evaluation. Until the 1967 essays by Scriven and Stake, few writers other than Tyler and Cronbach had tackled the overall conceptual nature of educational evaluation. By the early 1970s the field had grown rapidly, and many formal evaluation models were proposed. There was growing belief in the power of evaluation to transform poor educational programmes into highly effective programmes, and of the importance of evaluation results to decision-makers. However, this optimism of the early 1970s did not last. Experience showed that most educational decisions of importance, like most important decisions in the field of social policy, continued to be taken in a political, interpersonal milieu, where evidence plays a minor role (Popham, 1988). Educational decision-makers typically made their choices without waiting for the 'definitive' results of evaluations. Moreover, when the results were obtained they rarely provided conclusive.

With the realization of the political nature of the decision-making process, educational evaluators began to embrace Cronbach's view of the evaluator as an educator, in that he/she should rarely attempt to focus his/her efforts on satisfying a single decision-maker, but should focus those efforts on ''informing the relevant political community'' (Cronbach, 1982b). They also realized that, while many of their attempts at evaluation did not work, some did and when they worked programme quality improved to varying degrees. Improvement, even when modest, was recognized to be valuable (Popham, 1988).

The field of educational evaluation, as in the wider field of evaluation, has diversified and changed as a result of reflection on experience. It is a practice-driven field which, in the last 40 years, has emerged as an applied science in its own right.

Effecting programme evaluation

There are a number of steps to be taken in planning and implementing programme evaluation.

Initiation/commissioning

The initial stage of evaluation is where the institutions or individuals responsible for a programme take the decision to evaluate it. They must decide on the purpose(s) of the evaluation, and who will be responsible for undertaking it. There are potentially numerous reasons for undertaking evaluation. Muraskin (1997) lists some of the common reasons for conducting evaluations and common areas of evaluation activity (Table 1).

Table 1. Common reasons for understanding evaluation and common areas of evaluation activity (after Muraskin 1998).

- To determine the effectiveness of programmes for participants
- To document that programme objectives have been met
- To provide information about service delivery that will be useful to programme staff and other audiences
- To enable programme staff to make changes that improve programme effectiveness

Areas of evaluation activity

- Evaluation for project management
- Evaluation for staying on track
- Evaluation for programme efficiency
- Evaluation for programme accountability
- Evaluation for programme development and dissemination

Chelimsky & Shadish (1997) suggest that the purposes of evaluation, along with the questions evaluators seek to answer, fall into three general categories:

1. evaluation for accountability;
2. evaluation for knowledge;
3. evaluation for development.

These perspectives are not mutually exclusive. Each may be needed at particular times or policy points and not others, for example evaluation for knowledge may need to precede accountability. Table 2 shows the respective positions of the categories along five dimensions. The differences noted illustrate some of the current tensions in the evaluation field.

The potential cost of the evaluation often plays a major role in determining the scope of the evaluation and identity of the evaluator(s), as the cost will have to be met from the programme budget, or by seeking additional funding. The question of whether the evaluator should be internal or external to the programme's development and delivery is often considered at this point. In order to produce an effective educational evaluation, Coles & Grant (1985) point out that skills from many disciplines, for example psychology, sociology, philosophy, statistics, politics and economics, may be required. They rightly question whether one individual would have the competence to perform all these tasks, and whether an institution would necessarily have these skills in-house.

Defining the evaluator's role

The evaluator(s), having been appointed, must reflect on his/her role in the evaluation. This is important to establish as it will influence the decision-making process on the goals of the evaluation, and on the methodology to be used. It is at this point that the evaluator decides where, and to whom, his/her responsibility lies, and on the values he/she requires to make explicit. The questions to be asked in the evaluation, and their source of origin, will be influenced by these decisions.

The ethics of evaluation

Evaluators face potential ethical problems, for example, they have the potential to exercise power over people, which can injure self-esteem, damage reputations and affect careers. They can be engaged in relationships where they are vulnerable to people awarding future work. In addition, evaluators often come from the same social class and educational background as those who sponsor the evaluations. The ethics of an evaluation, however, are not the sole responsibility of the evaluator(s). Evaluation sponsors, participants and audiences share ethical responsibilities. House (1995) lists five ethical fallacies of evaluation:

1. Clientism—the fallacy that doing whatever the client requests or whatever will benefit the client is ethically correct.
2. Contractualism—the fallacy that the evaluator is obliged to follow the written contract slavishly, even if doing so is detrimental to the public good.
3. Methodologicalism—the belief that following acceptable inquiry methods assures that the evaluator's behaviour will be ethical, even when some methodologies may actually compound the evaluator's ethical dilemmas.
4. Relativism—the fallacy that opinion data the evaluator collects from various participants must be given equal weight, as if there is no bias for appropriately giving the opinions of peripheral groups less priority than that given to more pivotal groups.
5. Pluralism/Elitism—the fallacy of allowing powerful voices to be given higher priority, not because they merit such priority, but merely because they hold more prestige and potency than the powerless or voiceless.

To assist evaluators a number of organizations including the Joint Committee on Standards for Educational Evaluation (1994); the American Evaluation Association (1995); the Canadian Evaluation Society (1992); and the Australasian Evaluation Society (AMIE 1995) have issued guidance for evaluators undertaking evaluation. Other authors and professional organizations have also implicitly or explicitly listed ethical standards for evaluators, for example, the American Educational Research Association (1992), Honea (1992), and Stufflebeam (1991). Drawing on these, Worthen *et al.* (1997) have suggested the following standards could be applied:

1. Service orientation—evaluators should serve not only the interests of the individuals or groups sponsoring the evaluation, but also the learning needs of the programme participants, community and wider society.
2. Formal agreements—these should go beyond producing technically adequate evaluation procedures to include such issues as following protocol, having access to data, clearly warning clients about the evaluation's limitations and not promising too much.
3. Rights of human subjects—these include obtaining informed consent, maintaining rights to privacy and assuring confidentiality. They also extend into respecting human dignity and worth in all interactions so that no participants are humiliated or harmed.
4. Complete and fair assessment—this aims at assuring that both the strengths and weaknesses of a programme are accurately portrayed.
5. Disclosure of findings—this reflects the evaluator's responsibility to serve not only his/her client or

Table 2. Three perspectives and their positions along five dimensions (Adapted from Chelimsky & Shadish, 1997).

	Accountability perspective	Knowledge perspective	Developmental perspective
Purpose	To measure results or value for funds expended: to determine costs, to assess efficiency	To generate insights about public problems, policies, programmes & processes, to develop new methods and to critique old ones	To strengthen institutions to build agency or organizational capability in some evaluative area
Typical uses	Policy use, debate and negotiation, agency reform, public use	Enlightenment use, policy, research and replication, education, knowledge base construction	Institutional or agency use as part of the evaluative process, public and policy use
Evaluator role	Distant	Distant or close depending on evaluation design and method	Close, the evaluator is a 'critical friend' or may be part of a team
Advocacy	Unacceptable	Unacceptable, but now being debated	Often inevitable, but correctable through independent outside review
Position under policy debate	Can be strong (depending on leadership)	Can be strong (if consolidated and dissemination channels exist)	Uncertain (based on independence and control)

sponsor, but also the broader public(s) who supposedly benefit from both the programme and its accurate evaluation.
6. Conflict of interest—this cannot always be resolved. However, if the evaluator makes his/her values and biases explicit in an open and honest way clients can be aware of potential biases.
7. Fiscal responsibility—this includes not only the responsibility of the evaluator to ensure all expenditures are appropriate, prudent and well documented, but also the hidden costs for personnel involved in the evaluation.

However, the various educational backgrounds and professional affiliations of evaluators can result in them practising under several different and potentially conflicting ethical codes (Love, 1994). Given the pluralistic nature of those involved in evaluation, and the wider society, it is little wonder a consensus ethical code has not yet emerged.

To depend on codes ignores the values that the individual evaluator brings with him/her to the evaluation, and the importance of the individual being aware of his/her own values and being willing to change them in the light of changing knowledge. As with ethics in medicine a principles-based or a virtue-based approach, rather than adherence to external codes, may prove more desirable for practitioners.

Choosing the questions to be asked

The aims of the evaluation depend not only on the interests of individuals or groups asking them, and the purpose(s) of the evaluation, but also on the views of the evaluator as to his/her role. The work of Cronbach is perhaps the most far reaching in this area. He views the evaluator's role as educator rather than judge, philosopher-king or servant to a particular stakeholder group. In deciding which questions to ask, he advocates asking both all-purpose and case-specific questions. The all-purpose questions depend on the evaluator's assessment of the leverage associated with a particular issue, the degree of prior uncertainty about the answer and the degree of possible and desirable reduction in uncertainty in light of trade-offs among questions, methods and resources. This results in different types of issues prevailing in different programme contexts. His case-particular questions relate to the substantive theories underpinning programme design and investigate why a programme is, or is not, successful, knowledge that not all stakeholders are interested in as they may only desire outcome knowledge, or knowledge specific to their needs. His views place a heavy burden on the evaluator in terms of the methodology being complicated by the range of questions generated (Shadish *et al.*, 1991).

Shadish *et al.* (1991) supply a useful set of questions for evaluators to ask when starting an evaluation. These cover the five components of evaluation theory and provide a sound practical basis for evaluation planning (boxes 1–5).

Designing the evaluation

Having decided what needs to be done the evaluator has to design an appropriate plan to obtain the data required for the purpose(s) of his/her evaluation.

Dimensions of evaluation. Stake (1976) suggested eight dimensions along which evaluation methods may vary:

(1) Formative–summative: This distinction was first made by Scriven (1967). Formative evaluation is undertaken during the course of a programme with a view to adjusting the materials or activities. Summative evaluation is carried out at the end of a programme. In the case of an innovative programme it may be difficult to determine when the end has been reached, and often the length of time allowed before evaluation takes place will depend on the nature of the change.
(2) Formal–informal: Informal evaluation is undertaken naturally and spontaneously and is often subjective. Formal evaluation is structured and more objective.
(3) Case particular–generalization: Case-particular evaluation studies only one programme and relates the results only to that programme. Generalization may study one or more programmes, but allow results to be related to other programmes of the same type. In practice results may lend themselves to generalization, and the attempt to formulate rules for case

Box 1: Questions to ask about Social programming

(1) What is the problem to which the intervention is a response? For whom is it a problem? How big a problem is it, according to what criteria? Has the problem reached crisis proportions so that a short and timely evaluation needs to be done? Is the problem so trenchant that even a long-term evaluation will eventually be useful? Is the problem important enough to spend your time and evaluation resources studying it?
(2) Is this intervention just a minor variation on how things are usually done, or does it represent a major departure from common practice or common thinking about solutions to the problem?
(3) What are some of the alternative interventions to address this problem that have not been tried? Why have they not been tried? Would it be worth trying to locate, implement and evaluate some of these alternatives instead of the current intervention?
(4) Is this intervention a programme, a project or an element? How rapidly does it change in the natural course of things and does it have constituent parts that are quickly and easily changed? How big an impact could the intervention or its parts have, both as they are and if they were changed? Is the evaluation worth doing given the answers to these questions?

Box 2: Questions to ask about use

(1) What kind of use do you want to produce? Why? Given that it is hard to do an evaluation that facilitates all kinds of use equally well, what balance between instrumental and conceptual use would you like to produce?
(2) If you want to produce instrumental use, have you identified potential users? Have you talked with those users to find out what kind of information they want? Is the information they want to know about already available? Have you found out how potential users might use the results? When do they need results in order to use them, and can you provide the results by then? Will circumstances allow you to maintain frequent contact with users?
(3) If you want to produce enlightenment, who do you want to enlighten? How do you want to affect their thinking about the problem and the intervention? How do you reach them? Are the problem and intervention to be studied of lasting significance, or likely to be of transient interest? What do we know least about the problem and intervention?
(4) What are the characteristics of the programme being evaluated? Is it the whole programme or elements of it? What do its stakeholders want to know? Given these characteristics what questions would be most useful to ask?
(5) How should the results be communicated? Should interim results be reported periodically to users? In the final report should you include an executive summary? Action recommendations? Should oral briefings be used? Should reports of evaluation results be communicated in forms tailored to the specific information needs of different stakeholders? Can the results be disseminated through mass-media outlets?

Box 3: Questions to ask about knowledge construction

(1) What criteria are you going to use in deciding what constitutes acceptable knowledge? Will you use some traditional scientific set of standards, like the various constructs of validity? Which set, and why that set rather than others, such as fairness or credibility? How certain do you want the knowledge you construct to be? Are you willing to accept less than the most certain knowledge possible?
(2) What kind of knowledge does the client who paid for the evaluation want? What about programme stakeholders? How would they answer the questions you just answered about knowledge? Is there a serious mismatch between your standards for knowledge and those held by clients or stakeholders? If so, can you produce the kind of knowledge that stakeholders want, or at least educate them about the advantages and disadvantages of your respective opinions?
(3) What kind of knowledge, if any, do you think should be the most important in the evaluation? Knowledge about causation, generalization, implementation, costs, clientele, or something else? Why? How can you maintain a capacity for discovering things you did not think of at first?
(4) Can you produce the required knowledge, at the desired level of certainty, in the time available? Do you have sufficient acquaintance with the methodologies you need to produce this information? If not, can you build a team with such expertise?
(5) What arrangements will you make to carry out critical evaluation of your own evaluation? Can you do this prior to implementing the evaluation? Can outside experts or stakeholders critique your initial design or your final report?

Box 4: Questions to ask about valuing

(1) What would a thing like this intervention do to be good, and how would it accomplish those ends? What do other interventions like it do? What needs might it meet, and what harm would be done if those needs were not met? What do its stakeholders want it to do? What values does this intervention foster, and what values does it potentially harm? Might there be any negative side effects?
(2) How well does it have to do these things? Is it possible to construct any absolute standards in answer to this question? Are there any other interventions like the one being studied to which it could be compared? Are there any other interventions that could do what this intervention does, even if the two interventions do not seem much alike?
(3) How will you measure programme performance?
(4) At the end of the evaluation, do you plan to summarize all your results into a statement about whether the intervention is good or bad? If so, how will you weigh the different criteria in summing them to reflect which criteria are more or less important? Is it possible or desirable to construct a different value summary for each stakeholder group?

Box 5: Questions to ask about evaluation practice

(1) Why is the evaluation being initiated? How else could the money currently earmarked for this evaluation be spent? Is it worth spending time and money on this evaluation given other things one could do? Why?
(2) What purposes might the evaluation serve? To measure programme effects? To improve the programme? To influence the decision-makers? To judge programme worth? To provide useful information? To explain how an intervention, or ones like it, work? To help solve social problems? Why? How will you choose among these purposes?
(3) What role do you want to play in the evaluation? Methodological expert? Servant to the management or some larger set of stakeholders? Judge of the programme's worth? Contributor to programme improvement? Servant of the 'public interest'? Educator of the client paying for the evaluation? Why?
(4) Where could you get the questions? From clients, stakeholders or those who paid for the evaluation? From past research, theory or evaluations? From pending decisions or legislation? Why?
(5) What questions will you ask in this evaluation? Questions about real and potential clients and their characteristics and needs? About how the programme is implemented? About client outcome, and impacts on those with whom the client interacts? About the connections among clients, programme implementation and outcome? About costs and fiscal benefits? Why?
(6) What methods will you use? Why? Case-study methods like observation, interviews and inspection of records? Surveys? Needs assessments? Achievement testing? Meta-evaluation or meta-analysis? Causal modelling? Why? Do these methods provide good answers to the questions you are asking?
(7) How do you plan to facilitate the use of the evaluation? Is it your responsibility to do so? Why?
(8) Can you do all this within time and budget? If not, then what has the highest priority and why?
(9) What are your fallback options if something goes wrong with any of these matters?

study recognizes that generalizing requires greater control, and more regard to setting and context (Holt, 1981).

(4) Product–process: This distinction mirrors that of the formative–summative dimension. In recent years evaluators have been increasingly seeking information in the additional area of programme impact.
    (a) Process information: In this dimension information is sought on the effectiveness of the programme's materials and activities. Often the materials are examined during both programme development and implementation. Examination of the implementation of programme activities documents what actually happens, and how closely it resembles the programme's goals. This information can also be of use in studying programme outcomes.
    (b) Outcome information: In this dimension information is sought on the short-term or direct effects of the programme on participants. In medical education the effects on participants' learning can be categorized as instructional or nurturant. The method of obtaining information on the effects of learning will depend on which category of learning outcome one attempts to measure.
    (c) Impact information: This dimension looks beyond the immediate results of programmes to identify longer-term programme effects.

5. Descriptive—judgmental: Descriptive studies are carried out purely to secure information. Judgmental studies test results against stated value systems to establish the programme's effectiveness.
6. Pre-ordinate–responsive: This dimension distinguishes between the situation where evaluators know in advance what they are looking for, and one where the evaluator is prepared to look at unexpected events that might come to light as he/she goes along.
7. Holistic–analytic: This dimension marks the boundary between evaluations, which looks at the totality of a programme, from one that looks only at a selection of key characteristics.
8. Internal–external: This separates evaluations using an institution's own staff from those that are designed by, or which require to satisfy, outside agencies.

Choosing the appropriate design

A range of methods, from psychometric measurement at one end to interpretive styles at the other, has been developed. Table 3 provides a list of common quantitative and qualitative methods and instruments available to educational evaluators.

Shadish *et al.* (1991) advocate that evaluation theory can help tell us when, where and why some methods should be applied and others not. It can suggest sequences in which methods could be applied, ways in which different methods can be combined, types of questions answered better or less well by a particular method and the benefits to be expected from some methods as opposed to others. Cronbach (1982a) advises evaluators to be eclectic in their choice of methods, avoiding slavish adherence to any particular methods. Rossi & Freeman (1985) advocate the 'good enough' rule for choosing evaluation designs: ''The evaluator should choose the best possible design, taking into account practicality and feasibility . . . the resources available and the expertise of the evaluator'', a view echoed by Popham (1988).

Shadish (1993), building on Cook's (1985) concept that triangulation should be applied not only to the measurement phase but to other stages of evaluation as well,

Table 3. Common quantitative and qualitative methods and instruments for evaluation.

Quantitative methods	Qualitative methods
Experiments	Case studies
Pre–post test design	Action research approach
Post-test only design	Naturalistic and ethnographic approaches
Quasi-experiments	
Interrupted time-series design	
Non-equivalent comparison group design	
Regression–discontinuity design	
Surveys, longitudinal, cross-sectional and trend studies	
Delphi technique	
Q-sorts	
Cost-analysis	
Instruments	Instruments
Achievement testing	Interviews
Norm referenced semi-structured	
Criterion referenced unstructured	
Objectives-referencing discussion groups	
Domain-referencing	Observation
Attitude scales participant	
Rating scales spectator	
Questionnaires	Diaries/self-reports
Observation schedules	Documentary analysis
Interaction analysis	

advocates using critical multiplism to unify qualitative and quantitative approaches. He proposes seven technical guidelines for the evaluator in planning and conducting his/her evaluation:

(1) Identify the tasks to be done.
(2) Identify different options for doing each task.
(3) Identify strengths, biases and assumptions associated with each option.
(4) When it is not clear which of the several defensible options is least biased, select more than one to reflect different biases, avoid constant biases and overlook only the least plausible biases.
(5) Note convergence of results over options with different biases.
(6) Explain differences of results yielded by options with different biases.
(7) Publicly defend any decision to leave a task homogenous.

Approaches to evaluation

A plethora of evaluation models have been developed that can assist the evaluator in choosing the optimum method(s) for his/her particular evaluation. These range from comprehensive prescriptions to checklists of suggestions and as such are better described as approaches as many do not qualify for the term 'model'. Each approach comes with its built-in assumptions about evaluation and emphasizes different aspects of evaluation depending on the priorities and preferences of its author(s). Few come with careful step-by-step instructions practitioners can follow and even fewer are useful in settings and circumstances beyond those in which they were created (Worthen *et al.*, 1997). Atkin & Ellett (1985) contend that prescriptive evaluation models can be categorized along three dimensions: Methodology, Values, and Uses. The relative emphasis given to each area allows contrasts to be drawn between approaches.

With the explosion in the numbers of approaches in recent years, many of which overlap, a number of attempts have been made to categorize the different evaluation approaches. One of the most useful was developed by Worthen *et al.* (1997), influenced by the work of House (1976, 1983). They classify evaluation approaches into the following six categories:

(1) *Objectives-oriented approaches*—where the focus is on specifying goals and objectives and determining the extent to which they have been attained.
(2) *Management-oriented approaches*—where the central concern is on identifying and meeting the informational needs of managerial decision-makers.
(3) *Consumer-oriented approaches*—where the central issue is developing evaluative information on 'products', broadly defined, for use by consumers in choosing among competing products, services etc.
(4) *Expertise-oriented approaches*—these depend primarily on the direct application of professional expertise to judge the quality of whatever endeavour is evaluated.
(5) *Adversary-oriented approaches*—where planned opposition in points of view of different evaluators (for and against) is the central focus of the evaluation.
(6) *Participant-oriented approaches*—where involvement of participants (stakeholders in the evaluation) is central in determining the values, criteria, needs and data for the evaluation.

These categories can be placed along House's (1983) dimension of utilitarian to intuitionist-pluralist evaluation (Figure 1). Utilitarian approaches determine value by assessing the overall impact of a programme on those affected, whereas intuitionist-pluralist approaches are based on the idea that value depends on the impact of the programme on each individual involved in the programme.

Placement along the dimension is to some degree arbitrary. As evaluation is multifaceted and can be conducted at different phases of a programme's development, the same evaluation approach can be classified in diverse ways according to emphasis. The classification is based on what is seen as the driving force behind performing the evaluation. Within each category the approaches vary by level of formality and structure. To provide complete lists of the many different approaches to evaluation, which could appear under each category, is beyond the scope of this guide. Table 4 provides some typical examples of approaches that could appear under each category.

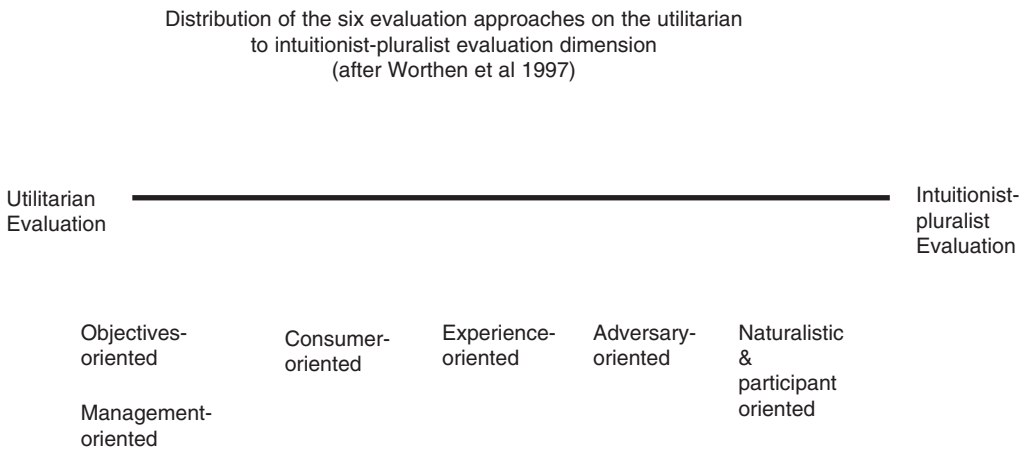

Figure 1. Distribution of the six evaluation approaches on the utilitarian to intuitionist–pluralist evaluation dimension.

Table 4. Examples of approaches that predominantly fit into Worthen *et al.*'s categories (1997).

Objectives-oriented
Tyler's industrial model (Smith & Tyler, 1942)
Metfessel & Michael (1967)
Provus's discrepancy model (1973)
Hammond (1973)
Scriven's goal- free evaluation (1972)
Management-oriented
The CIPP evaluation model (Stufflebeam, 1971)
The UCLA evaluation model (Alkin, 1969)
Provus's discrepancy model (1973)
Paton's utilization-focused approach (1986)
Wholley's approach to evaluation (1983, 1994)
Cronbach (1963, 1980)
Consumer-oriented
Scriven's concerns and checklists (1967, 1974, 1984, 1991)
Educational Products Information Exchange (EPIE) activities (Komoski)
CMAS (Morrisett & Stevens, 1967)
Expertise-oriented:
Accreditation bodies
Educational Connoisseurship (Eisner, 1975, 1991)
Adversary-oriented
Owens (1973)
Wolf (1975, 1979)
Levine *et al.* (1978)
Kourilsky (1973)
Participant-oriented
Stake's countenance framework (1967)
Parlett & Hamilton's illuminative model (1976)
Stake's responsive evaluation framework (1975)
Stake's preordinate evaluation approaches (1975)
Guba & Lincoln's writings on naturalistic enquiry (1981, 1989, Lincoln & Guba, 1985)

Similarly, to provide a complete description of each example would be beyond the scope of this guide. To assist readers in choosing which of the approaches might be most helpful for their needs, the characteristics, strengths and limitations of the six approaches are summarized in Table 5. These are considered under the following headings after Worthen *et al.* (1997):

(1) Proponents—individuals who have written about the approach.
(2) Purpose of evaluation—the intended use(s) of evaluation proposed by writers advocating each particular approach or the purposes that may be inferred from their writings.
(3) Distinguishing characteristics—key descriptors associated with each approach.
(4) Past uses—ways in which each approach has been used in evaluating prior programmes.
(5) Contribution to the conceptualization of an evaluation—distinctions, new terms or concepts, logical relationships and other aids suggested by proponents of each approach that appear to be major or unique contributions.
(6) Criteria for judging evaluations—explicitly or implicitly defined expectations that may be used to judge the quality of evaluations that follow each approach.
(7) Benefits—strengths that may be attributed to each approach and reasons why one might want to use this approach.
(8) Limitations—risks associated with use of each approach.

While some evaluators adopt or adapt proposed approaches, many evaluators conduct evaluations without strict adherence to any 'model'. However, they often draw unconsciously on their philosophy, planning and procedures through exposure to the literature. The value of the alternative approaches lies in their ability to present and provoke new ideas and techniques and to serve as checklists.

Interpreting the findings

Having collected the relevant data the next stage in evaluation involves its interpretation. Coles & Grant (1985) view this process as involving two separate, though closely related activities: analysis and explanation.

In analysing the findings, whichever method is chosen, it is important to establish the reliability and validity of the

Table 5. Comparative analysis of the characteristics, strengths and limitations of the six categories (after Worthen *et al.*, 1997).

	Objectives oriented	Management oriented	Consumer oriented	Expertise oriented	Adversary oriented	Participant oriented
Some proponents	Tyler Provus Metfesel & Michael Hammond Popham Taba Bloom	Stufflebeam Alkin Provus Wholley Cronbach	Scriven Komoski	Eisner Accreditation groups	Wolf Owens Levine Kourilsky	Stake Patton Guba & Lincoln Rippney MacDonald Parlett & Hamilton Cousins & Earl
Purpose of evaluation	Determining the extent to which objectives are achieved	Providing useful information to aid decision-making	Providing information about products to aid decisions about purchases or adoptions	Providing professional judgements of quality	Providing a balanced examining all sides of controversial issue; highlighting strengths and weaknesses	Understanding and portraying complexities of a programme's activities; responding to an audience's requirements for information
Distinguishing characteristics	Specifying measurable objectives; using objective instruments to gather data	Serving rational decision-making; evaluating at all stages of programme development	Using criterion checklists to analyse products; product testing	Basing judgements on individual knowledge & experience; use of consensus standards, team/site visitations	Use of public hearings, opposing points of view; decisions based on	Reflecting multiple realities; use of inductive reasoning and discovery
	Looking for discrepancies between objectives and performance		Informing consumers		Arguments heard during proceedings	First-hand experience on site
Past uses	Programme development; monitoring participant outcomes; needs assessment	Programme development; institutional management systems; programme planning; accountability	Consumer reports; product development; selection of products for dissemination	Self-study; blue-ribbon panels; accreditation; examination by committee; criticism	Examination of controversial programmes or issues; policy hearings	Examination of innovations or change about which little is known; ethnographies of operating programmes
Contributions to conceptualization of an evaluation	Pre–post measurement of performance; clarification of goals; use of objective tests and measurements that are technically sound	Identify and evaluate needs and objectives, consider alternative programme designs and evaluate them; watch the implementation of a programme; look for bugs and explain outcomes; see if needs have been reduced or eliminated; Meta-evaluation Guidelines for institutionalizing evaluation	Lists of criteria for evaluating educational products and activities; archival references for completed reviews: formative–summative roles of evaluation; bias control	Legitimization of subjective criticism outside self-study with verification standards	Use of forensic and judicial forms of public hearing; cross-examination of evidence through presentation of multiple perspectives, focus on and clarification of issues	Emergent evaluation designs; use of inductive reasoning; recognition ofmultiple realities; importance of studying context; criteria for judging the rigour of naturalistic enquiry

Criteria for judging evaluations	Measurability of objectives; measurement validity and reliability	Utility; feasibility; propriety; technical soundness	Freedom from bias; technical soundness; defensible criteria used to draw conclusions and make recommendations; evidence of need and effectiveness required	Use of recognized standards; qualifications of 'experts'	Balance; fairness; 'publicness'; opportunity for cross-examination	Credibility fit; audibility; confirmability
Benefits	Ease of use; simplicity; focus on outcomes; high acceptability; forces objectives to be set	Comprehensiveness; sensitivity to information needs of those in a leadership position systematic approach to evaluation throughout the process of programme development; well operationalized with detailed guidelines for implementation; use of a wide variety of information	Emphasis on consumer Information needs; influence on product developers; concern with cost-effectiveness and utility; availability of checklists	Broad coverage; Efficiency—ease of implementation, timing; capitalizes on human judgement	Broad coverage; close examination of claims aimed at closure or resolution; illumination of different sides of issues impact on audience; use of a wide variety of information	Focus on description and judgement; concern with context; openness to evolve evaluation plan; pluralistic; use of inductive reasoning; use of a wide variety of information; emphasis on understanding
Limitations	Oversimplification of evaluation and programmes; outcome-only orientation; reductionistic; linear; over-emphasis on outcomes	Emphasis on organizational efficiency and production model; assumption of orderliness and predictability in decision-making; can be expensive to administer and maintain; narrow focus on the concerns of leaders	Cost and lack of sponsorship may suppress creativity or not open to debate or cross-examination	Replicability; vulnerability to personal bias; scarcity of supporting documentation to support conclusion; open to conflict of interest; superficial look at context; overuse of intuition; reliance on the qualifications of 'experts'	Fallible arbiters or judges; high potential costs and consumption of time; reliance on investigatory and communication skills of presenters; potential irrelevancies or artificial polarization; limited to information that is presented	Non-directive; tendency to be attracted by the bizzare or atypical; potentially high labour-intensity and cost; hypothesis generating; potential for failure to reach closure

data. Growing consciousness of the fallibility of observation is reflected in the growth of multiple methods of data collection, in having multiple investigators analyse the same data set, and in doing data analysis in more of an exploratory than confirmatory mode (Glymour & Scheines, 1986). A philosophy of data analysis akin to Tukey's (1977) has arisen in quantitative evaluation, with issues being approached in several ways predicted on different methodological and substantive assumptions (Shadish *et al.*, 1991). An analogous development in the qualitative tradition is using multiple observers for each site instead of single observers. However, this can lead to observers struggling to reconcile the different interpretations that can arise (Trend, 1979).

When both qualitative and quantitative methods are used in the same study results can be generated that have different implications for the overall conclusion, leading to creative tension which may be resolved only after many iterations (Hennigan *et al.*, 1980). Whatever the data collection method, multiple tentative probes are the watchword, replacing conceptions based on theory-free observation, single definitive tests and crucial single studies (Shadish *et al.*, 1991). As mentioned previously, Shadish (1993) advocates using critical multiplism to unify qualitative and quantitative approaches.

Recognition of the social components of evaluation knowledge and the fallibility of evaluation methodologies has led to methods for critically scrutinizing evaluation questions and methods. These include: commentaries on research plans by experts and stakeholders; monitoring of the implementation of evaluations by government bodies and scientific advisory bodies; simultaneous funding of independent evaluations of the same programme; funding secondary analyses of collected data; including comments in final reports by personnel from the programme evaluated; and forcing out latent assumptions of evaluation designs and interpretations, often through some form of adversarial process or committees of substantive experts (Cook, 1974; Cronbach, 1982a). Such meta-evaluations assert that all evaluations can be evaluated according to publicly justifiable criteria of merit and standards of performance, and that the data can help determine how good an evaluation is. The need for meta-evaluation implies recognition of the limitations of all social science, including evaluation (Hawkridge, 1979). Scriven (1980) developed the Key Evaluation checklist, a list of dimensions and questions to guide evaluators in this task (Table 6).

Having analysed the data, the evaluator needs to account for the findings. In education the researcher accounts for the findings by recourse to the mechanisms embodied in the contributing disciplines of education (Coles & Grant, 1985). As few individuals have expert knowledge of all the fields possibly required, specialist help may be required at this point. This again has resource implications for the evaluation. Shadish *et al.*'s (1991) questions (boxes 2–5) also offer evaluators salient points to consider when interpreting the results of their evaluation.

Table 6. Key evaluation checklist.

(1)	Description: What is to be evaluated?
(2)	Client: Who is commissioning the evaluation?
(3)	Background and context of the evaluand and the evaluation
(4)	Resources available to or for the use of the evaluand, and of the evaluators
(5)	Function: What does the evaluand do?
(6)	Delivery system: How does the evaluand reach the market?
(7)	Consumer: Who is using or receiving the (effects of) the evaluand?
(8)	Needs and values of the impacted and potentially impacted population
(9)	Standards: Are there any pre-existing objectively validated standards of merit or worth that apply?
(10)	Process: What constraints/costs/benefits/apply to the normal operation of the evaluand?
(11)	Outcomes: What effects are produced by the evaluand?
(12)	Generalizability to other people/places/ times/versions
(13)	Costs: Dollar versus psychological versus personal; initial versus repeated; direct/indirect versus
(14)	immediate/delayed/discounted
(15)	Comparisons with alternative options
(16)	Significance: A synthesis of all the above
(17)	Recommendations: These may or may not be requested, and may or may not follow from the evaluation
(18)	Report: Vocabulary, length, format, medium, time, location, and personnel for its presentation need careful scrutiny
(19)	Meta-evaluation: The evaluation must be evaluated, preferably prior to (a) implementation, (b) final dissemination of the report. External evaluation is desirable, but first the primary evaluator should apply the Key Evaluation Checklist to the evaluation itself. Results of the meta-evaluation should be used formatively, but may also be incorporated in the report or otherwise conveyed (summatively) to the client and other appropriate audiences

Dissemination of the findings

Again Shadish *et al.*'s questions on evaluation use (box 2) are of value in considering how, and to whom, the evaluation findings are to be reported. Reporting will be in some verbal form, written or spoken, and may be for internal or external consumption. It is important for the evaluator to recognize for which stakeholder group(s) the particular report is being prepared. Coles & Grant (1985) list the following considerations:

(1) Different audiences require different styles of report writing.
(2) The concerns of the audience should be reviewed and taken into account (even if not directly dealt with).

(3) Wide audiences might require restricted discussion or omission of certain points.
(4) The language, vocabulary and conceptual framework of a report should be selected or clarified to achieve effective communication.

Evaluators require to present results in an acceptable and comprehensible way. It is their responsibility to persuade the target audience of the validity and reliability of their results. Hawkridge (1979) identified three possible barriers to successful dissemination of educational research findings:

(1) The problem of translating findings into frames of reference and language which the target audience can understand. However the danger in translating findings for a target audience is that the evaluator may as a result present the findings in a less than balanced manner.
(2) If the findings are threatening to vested interests, they can often be politically manoeuvred out of the effective area.
(3) The 'scientific', positivistic, approach to research still predominates in most academic institutions, which may view qualitative research methods and findings as 'soft', and be less persuaded by their findings. As qualitative methods receive greater acceptance this is becoming less of a problem.

A further problem concerns the ethics of reporting. As Coles & Grant (1985) suggested in their consideration of how—and to whom—to report, dissemination of information more widely may require to be censored, for example, information about a particular teacher would not usually be shared with anyone outside a select audience. The evaluator also has to be aware that the potential ramifications of a report may go wider than anticipated, for example into the mass media, where this may not be desired.

Influencing decision-making

As has been touched upon earlier, the initial enthusiasm of the 1970s educational evaluators became soured by the realization of the political nature of the educational decision-making process, and by the inconclusive results that were often obtained. Coles & Grant (1985) suggest the following ways in which evaluators can effect the educational decision-making process:

(1) involving the people concerned with the educational event at all stages of the evaluation;
(2) helping those who are likely to be associated with the change event to see more clearly for themselves the issues and problems together with putative solutions;
(3) educating people to accept the findings of the evaluation, possibly by extending their knowledge and understanding of the disciplines contributing towards an explanation of the findings;
(4) establishing appropriate communication channels linking the various groups of people involved with the educational event;
(5) providing experimental protection for any development, allocating sufficient resources, ensuring it has a realistic life expectancy before judgements are made upon it, monitoring its progress;
(6) appointing a coordinator for development, a so-called change agent;
(7) reinforcing natural change. Evaluation might seek out such innovations, strengthen them and publicize them further.

Conclusions

Evaluation has come of age in the last 40 years as an applied science in its own right, underpinned by evaluation theory.

Evaluators have to be aware of the political context in which many evaluations take place and of their own values and beliefs. They must decide where and to whom their responsibilities lie, and be aware of their ethical responsibilities while realizing evaluation sponsors, participants and audiences also have ethical responsibilities. Evaluators are often limited in the scope of the evaluation they can undertake owing to budgetary—and their own technical—limitations.

In performing evaluations, evaluation theory can help evaluators with all aspects of the process. Previously adopted approaches often present and provoke new ideas and techniques, and provide useful checklists. However, evaluators should be aware of the limitations of individual evaluation approaches and be eclectic in their choice of methods. The 'good enough' rule is worth remembering. As with all research findings, the validity and reliability of the data obtained are important to establish. When using quantitative and qualitative approaches in the same evaluation it is important to unify the different approaches. Recognition of the social components of evaluation knowledge and the fallibility of evaluation methodologies has led to the need for meta-evaluation.

In disseminating the findings, evaluators need to present results in an acceptable and comprehensible way for sponsors, the various stakeholder groups and the wider society. Further consideration of the political context is often required at this point, particularly when attempting to influence decision-making.

On reviewing the results of his/her endeavour it is important for the educational evaluator to remember the lesson history teaches: that improvement, even when modest, is valuable.

Notes on contributor

John Goldie, MB ChB MMEd FRCGP DRCOG Dip Med Ed ILTM, is Senior Clinical Tutor in the Department of General Practice, University of Glasgow. For the last eight years he has been lead researcher in the evaluation of ethics learning in Glasgow University's new medical curriculum.

References

Alkin, M.C. (1969) Evaluation theory development, *Evaluation Comment*, 2, pp. 2–7.
American Educational Research Association (1992) Ethical standards, *Educational Researcher*, 21, pp. 23–26.
American Evaluation Association (1995) Guiding principles for evaluators, in: W.R. Shadish, D.L. Newman, M.A. Scheirer &

C. Wye (Eds), Guiding principles for evaluators, *New Directions for Program Evaluation*, 34, pp. 19–26.

Amie, M. (1995) The Australasian Evaluation Society, *Evaluation*, 1, pp. 124–125.

Atkin, M.C. & Ellett, F.S. (1985) Evaluation models: Development, in: *International Encyclopedia of Education: Research and Studies* (Oxford, Pergamon Press).

Bloom, B.S. (1956) *The Taxonomy of Educational Objectives* (London, Longman).

Canadian Evaluation Society (1992) Standards for program evaluation in Canada: a discussion paper, *Canadian Journal of Program Evaluation*, 7, pp. 157–170.

Chelimsky, E. & Shadish, W.R. (1997) *Evaluation for the 21st Century: A Handbook* (Thousand Oaks, CA: Sage).

Coles, C.R. & Grant, J.G. (1985) Curriculum evaluation in medical and health-care education. ASME Medical Education Research Booklet 1, *Medical Education*, 19, p. 405.

Collins English Dictionary (1994) *Collins English Dictionary*, 3rd ed (Glasgow, Harper Collins).

Cook, T.D. (1974) The potential and limitations of secondary evaluations, in: M.W Apple, M.J. Subkoviak & M. Kamrass (Eds) Educational Evaluation: Analysis and

Cook, T.D. (1985) Postpositivist critical multiplism, in: R.L. Shotland & M.M. Mark (Eds) *Social Science and Social Policy* (Beverly Hills, CA: Sage).

Cronbach, L.J. (1963) Course improvement through evaluation, *Teachers College Record*, 64, pp. 672–683.

Cronbach, L.J. (1982a) In praise of uncertainty, in: P.H Rossi (Ed.) *Standards for evaluation practice*, pp. 49–58 (San Francisco, Jossey-Bass).

Cronbach, L.J. (1982b) *Designing Evaluations of Educational and Social Programs* (San Francisco, Jossey-Bass).

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F. & Weiner, S.S. (1980) *Towards Reform of Program Evaluation* (San Francisco, Jossey Bass).

Eisner, E.W. (1975) *The Perceptive Eye: Toward the Reformation of Educational Evaluation*, Occasional Papers of the Stanford Evaluation Consortium (Stanford, CA, Stanford University).

Eisner, E.W. (1991) Taking a second look: Educational connoisseurship revisited, in: M.W. McLaughlin & D.C Philips (Eds), *Evaluation and Education: At Quarter Century*, Ninetieth Yearbook of the National Association for the Study of Education, Part 2 (Chicago, University of Chicago Press).

Elliot, J. & Adelman, C. (1973) Reflecting where the action is, *Education for Teaching*, 92, pp. 8–20.

Glymour, C. & Scheines, R. (1986) Causal modelling with the TETRAD program, *Synthese*, 68, pp. 37–63.

Gronlund, N.E. (1976) *Measurement and Evaluation in Teaching*, 3rd edn (New York, Macmillan).

Guba, E.G. & Lincoln, Y.S. (1981) *Effective Evaluation: Improving the Usefulness of Evaluation Results through Responsive and Naturalistic Approaches* (San Francisco, Jossey-Bass).

Guba, E.G. & Lincoln, Y.S. (1989) *Fourth Generation Evaluation* (Thousand Oaks, CA, Sage).

Hammond, R.L. (1973) Evaluation at the local level, in: B.R. Worthen & J.R. Sanders (Eds) *Educational Evaluation: Theory and Practice*, pp 157–169 (Belmont, CA, Wordsworth).

Hawkridge, D. (1979) Persuading the dons? *British Journal of Educational Technology*, 10, pp. 164–174.

Hennigan, K.M, Flay, B.R & Cook, T.D. (1980) 'Give me the facts': some suggestions for using social science knowledge in national policy making, in: *Advances in Applied Social Psychology* (Hillsdale NJ, Lawrence Erlbaum).

Holt, M. (1981) *Evaluating the Evaluators* (Sevenoaks, Hodder & Stoughton).

Honea, G.E. (1992) *Ethics and public sector evaluators: nine case studies, unpublished doctoral dissertation* (Department of Educational Studies, University of Virginia).

House, E.R. (1976) Justice in evaluation, in: G.V. Glass (Ed.) *Evaluation Studies Review Annual*, Vol. 1 (Beverly Hills, CA, Sage).

House, E.R. (1983) Assumptions underlying evaluation models, in: G.F. Madaus, M. Scriven & D.L. Stufflebeam (Eds) *Evaluation Models: Viewpoints on Educational and Human Services Evaluation* (Boston, Kluwer-Nijhoff).

House, E.R. (1995) Principled evaluation: A critique of the AEA Guiding Principles. In: W.R. Shadish, D.L. Newman, M.A. Scheirer & C. Wye (Eds) *Guiding Principles for evaluators. New Directions for Program Evaluation*, 66, pp. 27–34 (San-Francisco, Josey Bass).

Joint Committee On Standards For Educational Evaluation (1994) *The Program Evaluation Standards*, 2nd edn (Thousand Oaks, CA, Sage).

Kelly, A.V. (1989) *The Curriculum: Theory and Practice*, 3rd edn (London, Paul Chapman Publishing).

Kourilsky, M. (1973) An adversary model for educational evaluation, *Evaluation Comment*, 4, pp. 3–6.

Lazarsfeld, P.F. & Rosenberg, M. (1955) *The Language of Social Research* (Glencoe, IL, Free Press).

Levine, M., Brown, E., Fitzgerald, C., Goplerud, E., Gordon, M.E., Jayne-Lararus, C., Rosenberg, N. & Slater, J. (1978) Adapting the jury trial for program evaluation: a report of an experience, *Evaluation and Program Planning*, 1, pp. 177–186.

Lewin, K. (1948) *Resolving Social Conflicts: Selected Papers on Group Dynamics* (New York: Harper & Brothers).

Lincoln, Y.S. & Guba, E.G. (1985) *Naturalistic Inquiry* (Beverly Hills, CA, Sage).

Love, A.J. (1994) Should evaluators be certified? in: J.W. Altschuld & M. Engle (Eds) *The Preparation of Professional Evaluators: Issues, Perspectives, and Programs*, New Directions for Program Evaluation, 62, pp. 29–40 (San Francisco, Jossey-Bass).

Madaus, G.F., Scriven, M.S. & Stufflebeam D.L. (1983) *Evaluation Models: Viewpoints on Educational and Human Services Evaluation* (Boston, Klewer-Nijhoff).

Mehrens, W.A. (1991) *Measurement and Evaluation in Education and Psychology*, 4th edn (New York, Holt, Rinehart & Winston).

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, pp. 931–943.

Morrisett, I. & Stevens, W.W. (1967) *Steps in Curriculum Analysis Outline* (Boulder, University of Colorado, Social Science Education Consortium).

Muraskin, L. (1997) *Understanding Evaluation: The Way to Better Prevention Programs* (Rockville, MD, Westat Inc.).

Newble, D. & Cannon, R. (1994) *A Handbook for Medical Teachers*, 3rd edn (Dordrecht, Kluwer).

Owens, T.R. (1973) Educational evaluation by adversary proceeding, in: E.R. House (Ed.) *School Evaluation: The Politics and Process* (Berkeley, CA, McCutchan).

Parlett, M. & Hamilton, D. (1976) Evaluation as illumination: a new approach to the study of innovatory programs, in: G.V. Glass (Ed.) *Evaluation Studies Review Annual*, Vol. 1 (Beverly Hills, CA, Sage).

Patton, M.Q. (1986) *Utilization-focused Evaluation*, 2nd edn (Beverly Hills, CA, Sage).

Popham, W.J. (1988) *Educational Evaluation*, 2nd edn (Englewood Cliffs, NJ, Prentice Hall).

Provus, M.M. (1973) Evaluation of ongoing programs in the public school system, in: B.R. Worthen & J.R. Sanders (Eds) *Educational Evaluation: Theory and Practice* (Belmont, CA, Wadsworth).

Rossi, P.H. & Freeman, H.E. (1985) *Evaluation: A Systematic Approach*, 3rd edn (Beverly Hills, CA, Sage).

Scriven, M. (1967) The methodology of evaluation, in: R.W. Tyler, R.M. Gagne & M. Scriven (Eds) *Perspectives of Curriculum Evaluation*, pp. 39–83 American Educational Research Association Monograph Series on Evaluation 1 (Chicago, Rand McNally).

Scriven, M. (1972) Pros and cons about goal-free evaluation, *Evaluation Comment*, 3, pp. 1–7.

Scriven, M. (1974) Goal-free evaluation, in: E.R. House (Ed.) *School Evaluation: The Politics and Process*, pp. 319–328 (Berkley, CA, McCutchin).

Scriven, M. (1980) *The Logic of Evaluation* (Inverness, CA, Edgepress).

Scriven, M. (1983) The evaluation taboo, in: E.R. House (Ed.) *Philosophy of Evaluation*, pp. 75–82 (San Francisco, Jossey-Bass).

Scriven, M. (1984) Evaluation ideologies, in: R.F Connor, D.G. Altman & C. Jackson (Eds) *Evaluation Studies Review Annual*, Vol. 9 (Beverly Hills, CA, Sage).

Scriven, M. (1991) Key evaluation checklist, in: M. Scriven, *Evaluation Thesaurus*, 4th edn (Beverly Hills, CA, Sage).

Sechrest, L., West, S.G., Phillips, M., Redner, R. & Yeaton, W. (1979) Some neglected problems in evaluation research, *Evaluation Studies Review Annual*, 4, pp. 15–35.

Shadish, W.R. (1993) Critical multiplism: a research strategy and its attendant tactics, in: L. Sechrest (Ed.) *Program Evaluation: A Pluralistic Enterprise*, New Directions for Program Evaluation, 60, pp. 13–57 (San Francisco, CA, Jossey-Bass).

Shadish, W.R., Cook, T.D. & Leviton, L.C. (1991) *Foundations of Program Evaluation: Theories of Practice* (Newbury Park, CA, Sage).

Smith, E.R. & Tyler, R.W. (1942) *Appraising and Recording Student Progress* (New York, Harper & Row).

Stake, R.E. (1967) The countenance of educational evaluation, *Teachers College Record*, 68, pp. 523–539.

Stake, R.E. (1975) *Program Evaluation, particularly Responsive Evaluation*, Occasional paper No. 5 (Kalamazoo, Western Michigan University Evaluation Centre).

Stake, R.E. (1976) *Evaluating Educational Programmes: the Need and the Response* (Menlo Park, CA, CERI/OECD).

Stufflebeam, D.L. (1971) *Educational Evaluation and Decision Making* (Itasca, IL, F.E. Peacock).

Stufflebeam, D.L. (1991) Professional standards and ethics for evaluators, in: M.W. McLaughlin & D.C Philips (Eds) *Evaluation and Education: At Quarter Century*, Ninetieth Yearbook of the National Association for the Study of Education, Part 2 (Chicago, University of Chicago Press).

Taba, H. (1962) *Curriculum Development: Theory and Practice* (New York, Harcourt, Brace and World).

Trend, M.G. (1979) On the reconciliation of qualitative and quantitative analyses: a case study, in: T.D Cook and C.S Reichardt (Eds) *Qualitative and Quantitative Methods in Evaluation Research*, pp. 68–86 (Beverley Hills, CA, Sage).

Tukey, J.W. (1977) *Exploratory Data Analysis* (Reading, MA, Addison-Wesley).

Tyler, R.W. (1935) Evaluation: A challenge to progressive education, *Education Research Bulletin*, 14, pp. 9–16.

Tyler, R.W. (1949) *Basic Principles of Curriculum and Instruction* (Chicago, University of Chicago Press).

Weiss, C.H. (1983a) The stakeholder approach to evaluation: origins and promise, in: A.S. Bryk (Ed.) *Stakeholder-based Evaluation*, pp. 3–14 (San Francisco, CA, Jossey-Bass).

Weiss, C.H. (1983b) Towards the future of stakeholder approaches in evaluation, in: A.S. Bryk (Ed.) *Stakeholder-based Evaluation*, pp. 83–96 (San Francisco, CA, Jossey-Bass).

Wholley, J.S. (1983) *Evaluation and Effective Public Management* (Boston, Little, Brown).

Wholley, J.S. (1994) Assessing the feasibility and likely usefulness of evaluation, in: J.S. Wholley, H.P. Hatry & K.E. Newcomer (Eds) *Handbook of Practical Program Evaluation* (San Francisco, CA, Jossey-Bass).

Wolf, R.L. (1975) Trial by jury: a new evaluation method, *Phi Delta Kappan*, 57, pp. 185–187.

Wolf, R.L. (1979) The use of judicial evaluation methods in the formulation of educational policy, *Educational Evaluation and Policy Analysis*, 1, pp. 19–28.

Worthen, B.L., Sanders, J.R. & Fitzpatrick, J.L. (1997) *Program Evaluation: Alternative Approaches and Practical Guidelines*, 2nd edn (New York, Longman).