Observing, Measuring, or Evaluating Courseware

Stephen W. Draper
Department of Psychology
University of Glasgow
Glasgow G12 8QQ U.K.
email: steve@psy.gla.ac.uk
WWW URL: http://www.psy.gla.ac.uk/~steve

Preface

These views originate in earlier work on evaluation in Human Computer Interaction done jointly with many colleagues including Keith Oatley and Paddy O'Donnell. Their application and adaptation to educational settings was done during my involvement in the TILT project (directed by Gordon Doughty), which is an institutional project funded under the TLTP programme. Consequently these ideas have been enormously influenced by the other members of the TILT evaluation group, principally Margaret Brown, Fiona Henderson, and Erica McAteer. But in writing these notes, I have found myself constantly thinking of remarks by Philip Crompton, who is the organiser of the ELTHE self-help group for evaluation, and represents to me the foot soldiers in evaluation for whom these essays are intended.

Those interested is pursuing the debate in this field can contact ELTHE (Evaluation of Learning Technology in Higher Education - self-help group). Contact Philip Crompton at pc4@stir.ac.uk or to join the group subscribe to ELTHE at mailbase@mailbase.ac.uk. The group also has a world wide web page, URL: http://annick.stir.ac.uk/elthe/

A version of this is to appear in the LTDI handbook.

Introduction

Numerous people are involved in some way in introducing learning technology into teaching, whether in acquiring and using some software developed elsewhere or in authoring new software. Having put in considerable effort during a project, we generally wish (or are required by others) to be able to show something about the results. Simply delivering the software on a disk is seldom felt to be enough: what can we do to pull together and present further evidence?

I shall refer to all such further evidence as "evaluation", and to the teaching material being evaluated as "courseware". In principle the same issues apply to all teaching methods from lectures and textbooks to computer software, multimedia, and advanced telecommunications. My views have grown from work in higher education, but may well apply in other areas of education. In what follows I offer an introduction to the basic issues of evaluating courseware in higher education, and an overview of some useful distinctions.

The simplest evidence is to list the functions of the software, or to list the number of people who bought or used the software. Such evidence is weak however because purchase, acquisition, and use depend as much on opportunity, available money, and advertising as on the quality of the courseware. Better evidence comes from inquiring about the effects, and there is a great range of methods to choose from: from asking informally how the teacher felt it went, to running a big controlled experiment.

What is the question?

As many writings and "methods" of evaluation say, the apparently obvious place to begin is with identifying the goal or purpose of evaluation: if you don't have a question you don't know what to do (to observe or measure), if you do then that tells you how to design the study. Many studies begin with questions like "Do the students learn more with the new software?". But you must ask yourself whether you are sure the question given you is the right question. After all, many questions are not. You could ask what colour a lecture was, bring in a spectroscope and take measurements during a lecture, but none of that would make the question sensible or get over the false presupposition that lectures have a colour. Similarly many people have talked as if "are computers good for learning?" was a sensible question, even though they would probably not have asked "are books good for learning?". Only if you are sure you know what the question is, that it is sensible, and that no surprises are possible, is it safe to base a study simply on making measurements that answer the given question. That is why including open-ended observations and questions is so important as part of most studies.

On the other hand, it is seldom helpful to approach a study with a blank mind. One place people go to for help is experts. Among other things, expertise gives a person experience of what the important issues and questions are likely to be. Every past problem can be turned into a question to check in future, although of course there is no guarantee that new problems will not emerge in new projects. Machell & Saunders (1991) in fact is basically a large, structured collection of questions, and novices to the field of evaluation find this very useful as a way of getting started. However it is important to recognise the present (and probably permanent) state of the field of education: no-one has a precise predictive theory of teaching or learning. Experts' experience allow their estimates to be of more value than novices, but it is not very accurate all the same. This has two consequences: that you must continue to ask whether your question is the right one and to make open-ended observations that may alert you to unforeseen issues, and that estimates no matter how expert are not going to be as accurate as actual measurements i.e. observing real students learning will always be more informative than consulting teachers and other experts, although it is usually more difficult and expensive. (Note that education is not so different from a lot of engineering in this respect: that is why testing is so important a part of most engineering projects, despite the expense.)

Planning and resource constraints

As with very many activities, no amount of expenditure guarantees getting what you really want, yet better quality results do require more resources. The maximum quality of the evaluation in many projects, and hence the quality of the lessons they can leave behind, and hence the long term usefulness of the projects as a whole, is effectively limited when they are set up. If time, money, and the skills for evaluation brought by hiring appropriate people, are not planned for and funded, then the outcomes are limited.

Yet planning is perhaps a more important limitation: provided evaluation is planned for from the start and kept high on the agenda, then useful results with modest resources are attainable. But without planning and management that keeps evaluation a high priority, it will not happen: evaluation cannot be effectively tacked on as an afterthought like writing an extra project report. This is most evident with projects centred on creating new materials. If testing and evaluation are not planned for as essential, then as the end of the project looms it is a rush to get any version at all finished, and the software will never be tested on learners. The chances of it being satisfactory are about those of a pedestrian walking across a motorway without injury, because we just cannot predict accurately whether and when students will learn. In fact such miracles have occurred, but few would conclude that that shows the procedure to be reasonable. Learners behave like motorists in such cases, and will avoid a disaster caused by others if they can: they will probably be very angry at having to work round the design faults, but since they want to learn they will do so even if it means going to the library afterwards to compensate for the deficiencies of the courseware. Allowing for testing is crucial, and even if relatively little time and money is spent on it, planning for it is cruical so that a working version of the software is ready in time: and that time is often determined by the availability of suitable test subjects. Furthermore, in development projects, more time after the test must be allowed for in which modifications suggested by the tests can be made. These are often not very lengthy to make, but they must be allowed for at the planning stage. Useful evaluation leads to action, and is largely wasted if it is done too late to make changes.

Planning at the project level, then, is the most important requirement. This is not only true in development projects, but also in projects centered on introducing courseware that is already finished. Here, evaluation will revolve around classroom trials, and these in turn are constrained by the availability of classes for the trials: often once a year at a time determined by the institution and not the project.

The main issues

Given at least some resources, and that planning was done in time, what might an evaluation consist of? The choices are enormous, and many of them are laid out in the references cited below. However there are perhaps two dimensions that turn out to be most important in understanding the space of choices.

Opinion, memory, and observation

The most convenient method for an evaluator is to ask someone else, preferably an expert, for a judgement. This is what journalists do almost entirely. It is obviously better than just recording their own opinions. However the opinions of (possibly interested) onlookers is not as informative as that of the learners themselves: that is why it is becoming standard practice to use student feedback questionnaires in teaching, rather than the teacher's own opinion of their performance, even though teachers are often aware of their own major strengths and weaknesses. However asking someone (a learner) retrospectively about a teaching episode, which is what all questionnaires do, is not nearly as informative as gathering on the spot information as it happens; although the difference in quality depends strongly on what is being asked about. For instance, when we ask students to tell us how long they spent on each learning resource (how long on an exercise, how long looking at the textbook) they have a lot of difficulty and are almost certainly very inaccurate. Similarly if you ask students to write down the worst feature of a piece of courseware, they can do this, but if you ask them to tell you about every problem they will forget most unless you ask them as they go along, when you will get perhaps five times as much information (at a cost of course, particularly to the student). This is because memory is much inferior to on the spot observation and recording. Questionnaires and interviews rely on memory and are therefore less valuable than on the spot observation, and the longer after the event they are, the less valauable they are.

Similarly an "experts' " opinion is less valuable than that of a teacher who has tried the materials on students, and a teacher's opinion is less valuable than those of actual learners. Learner's opinions however are often less trustworthy than behavioural tests (e.g. assessment scores): for instance men generally feel and express more confidence about what they have learned than women, while scoring no better on tests of what they actually learned. Again, cost and convenience run largely in the opposite direction (it is easier to ask opinions than to set and mark tests), and in practice a compromise must be decided.

In summary, although costs and opportunities may not often allow optimal methods, it is in general best to base evaluation on actual learning by representative students who really want to learn (not the opinions of onlookers or the performance of special subjects brought in for a trial); to test what they actually did learn, rather than asking whether they felt they learned; and if possible to observe them as they try to learn, and pick up as many observations from them as possible. Of course this is itself disruptive, and must often be avoided. The tradeoff here will be between getting the most useful information pointing to what changes to make to a design, and getting the most representative overall results. A development project might do well to decide to run some tests in a relatively disruptive mode as early as possible, and having refined the design run less disruptive tests to obtain evidence of final performance. Personal observation and interviewing gives better information than questionnaires, but on the other hand realistic classroom trials usually have all students learning at the same time, so questionnaires may be a sensible compromise in order to get data from the whole class with only one or two investigators.

Systematic surveys vs. surprise detection

The other major issue is that of the need for both answering systematically questions we are interested in in advance e.g. did all students learn the material up to some criterion, and detecting unexpected problems and issues. An analogy with visual perception may be useful. One thing that perception does is support specific tasks such as checking whether a particular friend's car drives past you: you scan all cars, make sure you don't miss any, and without bothering about irrelevant attributes of the cars e.g. how dirty they are, whether hub caps are missing, look at the identifying features (perhaps the registration number, or the colour and size). Another thing perception does however is allow you to notice completely unexpected things, such as a tiger walking down the street towards you, someone's umbrella which is just about to poke your eye out, or a street vendor offering vension which would do nicely for your dinner. It will do these things even though you did not plan to do them, and could not say that, for instance, you noticed everything on sale by street vendors.

Similarly with evaluation: it is important to cover both functions. Methods such as exam-type tests and questionnaires with fixed response categories will never warn you that something you did not anticipate is in fact important in the situation you are studying. Hence it is vital always to have some open-ended questions and preferably personal observation by the evaluator. In fact if at all possible it is best to run two studies, so that issues thrown up by the open-ended measures in the first can be used to do systematic surveys in the second. In this way, you can discover whether the 2 students who mentioned that the screens were hard to read in bright light were unusual, or in fact represented an issue that worried all the students. As this example shows, however, open-ended questions and observations are not a substitute for fixed questions: only by putting the same question or task to each learner and requiring the answers to be expressed using the same categories (or marked using the same coding or marking scheme) can you get comparative results that allow you to discover and report results such as what proportion of learners were affected by an issue.

Any evaluation study, then, should have both open-ended measures for detecting surprises, and fixed measures for generating comparative data that can answer specific questions. Without fixed measures you may not be able to say anything definite about the courseware: only an unstructured set of observations and opinions from individuals, which may or may not be shared by the other learners. Without open-ended measures you have no chance of detecting problems or anything you did not think of in advance, and it is from the unexpected that most important improvements stem.

Four types of evaluation classified by aim

When we consider possible approaches to educational evaluation, there are four general types described in the literature. We describe them in turn. They are not mutually wholly exclusive, but distinguishing them may be helpful before they are combined in individual cases.

Evaluation of CAL (computer assisted learning) is in fact intimately linked with the authoring and dissemination process. Thus approaches to evaluation reflect either what the authoring process seems to be before evaluation is considered, or else what the evaluators think it ought to be in order to make evaluation useful. Another way of putting this is that evaluation can be designed for different <u>purposes</u> or roles:

- <u>Formative evaluation</u>: to help improve the design of the CAL
- <u>Summative evaluation</u>: to help users choose which piece of CAL to use and for what
- <u>Illuminative evaluation</u>: to uncover the important factors latent in a particular situation of use
- <u>Integrative evaluation</u>: to help users make the most of a given piece of CAL

As far as I know the terms, though perhaps not the ideas, were introduced as follows: "formative" and "summative" by Scriven (1967) (see also Carroll & Rosson (1995) for their subsequent use in Human Computer Interaction); "illuminative" by Parlett & Hamilton (1972/77/87); "integrative" by Draper et al. (1996).

The consumer view: summative evaluation

The default "commonsense" view that tends to occur spontaneously to many people is that evaluation of CAL is rather like consumer reports on goods: the manufacturer designs and supplies them, then someone else does tests and produces reports to help purchasers decide which to buy. This view of evaluation is linked to a view that CAL is produced like textbooks and othe goods, and that evaluation is not expected to have any direct effect on the CAL itself by telling the authors how to improve it. Nor is it expected to help consumers in how to use the product: only which to buy. Thus this is a common view for perhaps these reasons: it fits the fact that a lot of CAL is produced like a lot of textbooks by a very small team of authors with no spare resources for testing; it fits with a tradition in the literature for comparative experimental testing (which can compare two sets of teaching materials well); it fits the needs of new CAL users to decide what to buy; and more broadly it is analogous to consumer reports and how we encounter most of the things we buy, which we are offered without being consulted about how we would like them designed.

Formative evaluation

One important use of evaluation is while it is being developed: testing it on learners while there are still resources for modifying it. This is the simplest way for evaluation to help authors (developers) is to try out the CAL material on users, preferably as similar as possible to the students it is intended for, and use open-ended methods to report the problems that arise and perhaps suggested amendments as well. Although often the time necessary for this is not allowed for in development plans, once a developer has experience of it, it is usually clear how useful this is. After all, testing is part of all engineering, and also feedback from students is used by almost all lecturers to adjust their lectures and handouts. The key point to realise when using it for CAL, is that such testing must be done in time to allow changes to the material in the light of the results before the end of the development period. This kind of testing is called formative evaluation, as it is used to modify ("form") the material.

The most realistic, and so most helpful, formative evaluation would use real students in their normal learning situation. This is likely to increase the time for the whole cycle of production, testing, and modification. Feedback to developers from other sites who are early users of the material is a helpful substitute that gets round this constraint. Although this practice really means that users are running poorly tested software, and in effect doing the testing that producers should have done themselves, it is better than having no way of catching problems and improving the software. It in fact corresponds to common processes in commercial software production, where producers keep track of users and collect performance reports in order to improve later releases of their software.

More information on planning this kind of evaluation can be found in Alessi & Trollip (1991), and in McAteer & Shaw (1994). As noted above the key constraint is planning to do the testing early enough that changes can be made. The reward is a significant improvement in quality of the end product. Thus the main added result will not be a report, but the modifications to the design actually done.

Illuminative evaluation

"Illuminative evaluation" refers to what might now be called loosely, and perhaps incorrectly, ethnography. The basic idea is for the investigator to hang out with the participants (students, teachers, etc.) to pick up how they think and feel about the situation, and what the important underlying issues are. For a more precise view and examples see Parlett & Hamilton (1972/77/87) and Parlett & Dearden (1977). Its importance is as an open-ended method that can detect what the important issues are, without which other methods often ask the wrong questions and measure the wrong things. For instance most studies still fail to measure motivation in any way, yet much CAL would never be used if it were not made compulsory by teachers or experimenters. However this is not a universal truth: in some cases students have a strong desire to use the CAL independent of coercion, in others they are indifferent and use it only under compulsion but without disliking it, in yet others they continue to express strong revulsion (even though educational tests show educational benefits). Another even simpler example concerns lectures: providing handouts and using slides were intended to augment the voice medium and make things easier for students, but it turned out from informants that this created a new problem for students of discovering from moment to moment what the connection between the three channels was (e.g. was the current slide on the handout or did they need to write it down?). Simply measuring the effectiveness of using the extra channels might have shown a reduced rather than an increased benefit, but without giving any clue about what the problem was. Illuminative evaluation is in effect a systematic focus on discovering the unexpected, using approaches inspired by anthropology rather than psychology.

TILT's "integrative" evaluation

The TILT project at Glasgow University has done many classroom studies of CAL. The kind of study they have concentrated on is of the real use of CAL as part of university courses, but with evaluators who can gather more and fuller information than a teacher alone can do through student verbal questions and standard course feedback questionnaires. They have begun to argue that these evaluations serve a rather different purpose than was first envisaged. They argue that for many teachers in practice, the question is no longer whether to use CAL or which package to use: this has often been decided already. Instead, for them the question is how to make the best use of CAL material they are already committed to using. Classroom evaluations typically give lots of information that can be used for this. For instance if all students complain about some issue, or score badly on a quiz item corresponding to an issue, then teachers immediately respond to the evaluation report by adjusting in some way e.g. making an extra announcement, or producing a supplementary handout. Thus a major use of classroom evaluations in practice is to be formative, not of the CAL itself, but of the overall teaching and learning situation. This of course can be and is responsive to local variations in how the CAL is used, and for whom. It can be a significant help in integrating CAL material into varying local situations and courses: Draper et al. (1996).

Approaches to method

The methods you use and questions you ask will depend partly on what you hope to use the evaluation results for (see previous section), partly on your views about methods.

Checklist approach

Machell & Saunders (1991) offers a structured approach to identifying the questions you are interested in from within a large space of possible concerns, pulling them together, and so perhaps generating a questionnaire for learners or a checklist for course organisers. This would lead to a report on courseware based on the pre-existing concerns of the evaluator, and largely relying on (memory for) experience of the courseware and its use.

What the participants feel

An alternative approach is not to rely on what the evaluator thinks, but to ask learners what they feel. A rather trivial form of this is common, in which a simple questionnaire asks learners whether they liked using the courseware — the "how was it for you?" approach. The problem with this is that it asks for opinions about enjoyment instead of measuring actual learning, and such feelings are strongly influenced by many things other than learning such as novelty or a desire to be polite to a concerned teacher. At the other extreme is a careful "illuminative" approach that identifies all the stakeholders (those affected by the courseware) and uses participant observation and in depth interviews rather than a short questionnaire. Parlett & Dearden (1977) and Murphy & Torrance (1987) illustrate work of this kind. In designing evaluations it may be best to avoid both ignoring and relying wholly on measurements of feeling: open ended observation of some kind, as argued above, is a crucial component of any evaluation; and learners' enjoyment and feelings are outcomes that it is as well to measure among others.

Addressing the whole situation

Courseware is generally only of interest if it promotes learning. However to the extent that it does, it only does so in conjunction with the wider teaching context in which it is used: how it is supported by handouts, books, compulsory assessment, whether the teacher seems enthusiastic about it, support among learners as a peer group, and many other factors. Major implications for evaluation follow from this. It is not possible to evaluate courseware by itself: you can only evaluate its effect together with that of the surrounding support it had in the situation studied. Evaluation must cover not just the courseware but the way and the situation in which it is delivered; and the results may only apply to that specific case.

Draper et al. (1994) is a rather pessimistic development of this point, concerned more with problems than solutions, but it does focus on the issues involved in looking at what actually determines learning in practice rather than only those issues most directly controlled by developers and distributors. In this it is in line with the emphasis above on the need for open-ended measures as well as systematic ones in order to detect issues that were not anticipated by the evaluator but which are important for how the courseware fares in practice.

However a focus on the specificity of the case can be a virtue: it allows evaluation to support teachers in getting the best out of a piece of courseware by optimising its integration into the particular local delivery situation. Although logically such reports do not tell you how the courseware would perform in other situations, building up a set of such detailed case studies complete with how successful they were and what teachers did to make them successful locally is obviously helpful information for other prospective users. Furthermore it accumulates information for teachers on how to use the courseware, which is still too seldom provided by the developers.

The experimental approach

The fourth, and grandest, kind of method is the experimental one. Here some educational intervention (such as a piece of courseware) will typically be tested by a direct comparison of its performance against that of some reasonable alternative (such as the traditional teaching it replaces). Educational journals have many examples of this approach to evaluation for research purposes.

This approach has two important characteristics. Firstly it is usually very expensive in time and researcher effort. A simple experiment comparing the performance of some new educational intervention against an alternative often consumes one or two person-years of research, without counting the input of teachers and and other research colleagues. This may be worth it to establish a new idea or theory, but not just to test one of the growing flood of new pieces of courseware. Secondly, any such experiment taken in isolation is open to all the criticisms sketched above that the learning outcomes in fact depend on many other factors besides the intervention being tested, many of which cannot be effectively controlled e.g. the enthusiasm the teachers and children feel about the methods being compared. Furthermore we are too ignorant of what these factors are to have any confidence that they are controlled in any experiment. Such experiments can be taken as establishing that it is now reasonable to take the new intervention seriously having performed well in one real test, but can seldom be taken as proof that it is inherently better or even necessarily effective by itself.

More comprehensive approaches

Above four roles for evaluation were introduced. However in practice more than one kind of evaluation can and should be done. Firstly, work done for one purpose may turn out useful for another (Draper et al.; 1996). Secondly, different types are appropriate at different stages in the development of an educational intervention (Scriven; 1967; Carroll & Rosson; 1995). In general, evaluation of one kind or another is useful before, during, and after development; and in well designed projects different kinds of evaluation should be done at different stages. One scheme for this has been developed by Diana Laurillard.

Laurillard's evaluation programme

Recently Diana Laurillard has presented a much more elaborate scheme for evaluation in various talks. In this approach, production stretches over years, and different evaluation techniques are used at different stages. For instance, before design begins a "phenomenographic" study (Marton; 1981) would be done of the main problems students experience in learning the topic from existing materials. This can identify both the starting point of students, and the main problems they are likely to encounter: essentially a pre-design analysis of needs. Evaluation in this approach continues through to full classroom trials of the CAL material used in the way specified by the developers.

In a talk in Nov. 1994, Laurillard outlined the following evaluation programme:

1. Pre-program design: Curriculum needs, Learning needs (phenomenographic study), Student access
2. Prototyping: Observation, Comparative trials
3. Formative evaluation: Observation, Pre/post tests, interviews, monitoring, questionnaires
4. Piloting: Observation, interviews, questionnaires
5. Summative evaluation: Questionnaires, interviews, tests, documentation

This method seems a good match for how the Open University teach courses, and also the larger packages produced by TLTP subject consortia for large classes of students in the first year. It seems unlikely to suit the development of CAL for final year options, where a long experience of teaching the topic does not exist, and the final number of students even nationally is unlikely to justify a big development effort. It also seems to ignore the widespread requirement to adapt CAL materials to local needs, where each application will be different and require separate evaluations that cannot be simply compared. This is because any classroom evaluation is really measuring the effect of the CAL material combined with all other components of the local situation e.g. announcements, integration with the rest of the course, etc. As conditions and indeed aims vary across institutions, so results will vary. Hence I would argue firstly for extending the above programme by a sixth step:

6. Integrative evaluation: tests, confidence logs, resource questionnaires (Brown et al. 1996).

I would also suggest that the relative emphasis and effort put into different stages will depend on the project and the size of the intended student population.

Postscript

As noted above, all of these types of evaluation could be done, each contributing something different. Two good books to begin with for further reading on this topic are Hamilton et al. (1977) and Murphy & Torrance (1987).

References

Alessi, S.M. & Trollip, S.R. (1991) Computer-based instruction: methods and development (New Jersey: Prentice Hall)

Brown,M.I., Doughty,G.F., Draper,S.W., Henderson,F.P., & McAteer,E. "Measuring learning resource use" Submitted to Computers and Education

Carroll,J.M. & Rosson,M.B. (1995) "Managing evaluation goals for training" Communications of the ACM vol.38 no.6 pp.40-48

Draper,S.W., Brown,M.I., Edgerton,E., Henderson,F.P., McAteer,E., Smith,E.D., & Watt,H.D. (1994) Observing and measuring the performance of educational technology TILT project, c/o Gordon Doughty, Robert Clark Centre, University of Glasgow [email: g.doughty@elec.gla.ac.uk]

Draper,S.W., Henderson,F.P., Brown,M.I., & McAteer,E. (1996) "Integrative evaluation: an emerging role for classroom studies of CAL" Computers and Education

Hamilton,D., Jenkins,D., King,C., MacDonald,B., & Parlett,M. (1977) (eds.) Beyond the numbers game: a reader in educational evaluation (Basingstoke: Macmillan)

Laurillard (no ref.). The views referred to above were expressed in talks in the period 1994-5 on her work on the TELL project. One lead would be to email D.Laurillard@open.ac.uk

Machell, J. & Saunders,M. (eds.) (1991) MEDA: An evaluation tool for training software Centre for the study of education and training, University of Lancaster [email: m.saunders@lancaster.ac.uk]

Marton, F. (1981) "Phenomenography - describing conceptions of the world around us" Instructional science vol.10 pp.177-200

McAteer,E. & Shaw,R. (1994) Courseware authoring guidelines: evaluation 1 - Developing and testing EMASHE project, c/o Gordon Doughty, Robert Clark Centre, University of Glasgow [email: g.doughty@elec.gla.ac.uk]

Murphy,R. & Torrance,H. (1987) (eds.) Evaluating education: issues and methods (Milton Keynes: Open University Press)

Parlett, M.R. & Hamilton,D. (1972/77/87) "Evaluation as illumination: a new approach to the study of innovatory programmes".
(1972) workshop at Cambridge, and unpublished report Occasional paper 9, Centre for research in the educational sciences, U. of Edinburgh.
(1977) D.Hamilton, D.Jenkins, C.King, B.MacDonald & M.Parlett (eds.) Beyond the numbers game: a reader in educational evaluation (Basingstoke: Macmillan) ch.1.1 pp.6-22.
(1987) R.Murphy & H.Torrance (eds.) Evaluating education: issues and methods (Milton Keynes: Open University Press) ch.1.4 pp.57-73

Parlett, M. & Dearden,G. (1977) Introduction to illuminative evaluation: studies in higher education (Pacific soundings press)

Scriven,M. (1967) "The methodology of evaluation" pp.39-83 in Tyler,R.W., Gagné,R.M. & Scriven,M. (eds.) Perspectives of curriculum evaluation (Rand McNally: Chicago).