Evaluation: improving practice, influencing policy

David Wall

KEY MESSAGES

- Evidence from educational evaluation is essential to enhance professional practice and to achieve the best medical education for students, trainees and doctors engaged in continuing professional development.
- Well-constructed evaluation is rigorous and defensible.
- The direct use of evaluation evidence in educational policy decisions is unusual.
- Good evaluation of medical education will ultimately improve patient care.

Introduction

This chapter covers the wide role of evaluation in medical education from micro to macro, from the evaluation of individual teaching episodes to entire curricula, for the purposes of improving pedagogy to influencing national policy.

We begin with some definitions and a discussion of the purposes of evaluation. Some conceptual models are then described including the evaluation cycle, a task-orientated model of evaluation and versions of the 'Kirkpatrick hierarchy'. This first, theoretical, section of the chapter concludes by exploring the differences between evaluation and research.

The chapter goes on to describe potential sources of evaluation evidence and a number of evaluation methods, including questionnaires, interviews, focus groups, site visits and some group evaluation activities. As much educational evaluation involves questionnaires, the next section describes the design and use of questionnaires in evaluation. It includes advice on writing your own evaluation tool, including writing the statements and the response options, putting the questionnaire together, using paper and online formats, response rates, piloting and field testing.

Finally, the all-important issue about whether evaluation results are acted upon is addressed. The paper discusses the role of evaluation in change management and the relationship between evaluation and educational policy. The concept of 'enlightenment' – a gradual seeping in and acceptance of new ideas, rather than a logical step-by-step incorporation of evaluation evidence – is introduced, along with some more covert purposes of evaluation, such as control. The paper concludes with examples of the range of evaluation tools that have been brought to bear on the important topics of curriculum, faculty development and educational climate.

What Is Evaluation?

Evaluation in medical education has been defined as 'a systematic approach to the collection, analysis and interpretation of information about any aspect of the conceptualisation, design, implementation and utility of educational programmes'.(1) Evaluation is generally understood to mean 'the process of obtaining information about a course or programme of teaching for subsequent judgement and decision making'.(2)

Evaluation questions we might therefore legitimately ask would include the following:

- What did the specialty registrars think of the cross-cultural communication skills course?
- What is the educational climate like for medical students in the operating theatre?
- Was the six-day faculty development course effective?
- How reliable was our shortlisting and interviewing for paediatric trainees?
- What do students and junior doctors think about the career advice provided?

Evaluation is then much more wide-ranging than merely about handing out questionnaires to students and trainees at the end of teaching sessions.

Definitions: Evaluation, Assessment and Appraisal

In everyday life the terms *evaluation*, *assessment* and *appraisal* are often used interchangeably. This confusion is compounded by international differences in

Understanding Medical Education: Evidence, Theory and Practice. Edited by T Swanwick.

definitions. In North America, for instance, the word 'evaluation' is equated with the UK term 'assessment', to mean measurement of learners' skills.(2) An example of this is in the mini-clinical evaluation exercise – actually an 'assessment' tool for testing junior doctors' history-taking and examination skills.(3) Since this is primarily a UK publication I shall be using the English and European meanings for evaluation, assessment and appraisal. Evaluation has already been defined and stipulative definitions of assessment and appraisal are provided below.

Assessment is defined as 'the processes and instruments applied to measure the learner's achievements, normally after they have worked through a learning programme of one sort or another.(1) Assessment then is about testing the learners. This also appears to be the way that the word 'assessment' is used in mainland Europe.

Appraisal is 'a two-way dialogue focussing on the personal, professional and educational needs of the parties, which produce agreed outcomes'.(4) In the UK National Health Service all general practitioners and consultants now have annual appraisals, with an appraiser, using a structured format to the appraisal documentation based on the General Medical Council's (GMC) *Good Medical Practice* areas.(5)

Purposes of Evaluation

There are many purposes of evaluation. One way to conceptualise this is to think in terms of curriculum, teaching and learning, and assessment.

Evaluation is vital for curriculum development, in measuring if the curriculum is fit for purpose and in terms of curriculum outcomes.

Evaluation may also be used to ensure that the education provided is meeting the learners' needs. Did the learners learn from the teaching programme? It is often used to identify areas where the teaching (in its widest sense) needs to improve. It is used to see if an educational programme is of an acceptable standard so it may be approved for training and accreditation purposes. It may be used to give feedback to teachers, to managers and to faculty on the programmes being run in the organisation. It may be used as part of the information presented at the annual appraisal process for medical teachers, and for promotion and career development. In terms of assessments, it may be used to gather outcome measures on pass rates for qualifying and professional examinations.

And increasingly, evaluation is used in the area of assessment, evaluating the development, uses, reliability and validity of assessment tools. Often this has reached very sophisticated levels, using complex psychometrics and statistics (e.g. generalisability theory) to ascertain if assessment tools are fit for purpose. In addition, evaluation may be used to determine future educational policy in either curriculum, teaching and learning, or assessment. It may also be used as a tool to implement centrally determined policy through a number of covert and controlling processes (Box 23.1).

BOX 23.1 Purposes of evaluation

Curriculum

- Curriculum development
- Fitness for purpose
- Curriculum outcomes

Teaching and learning

- Meeting learners' needs
- Identifying poor teaching
- Approval of teaching programmes
- Feedback to teachers and the organisation
- Annual appraisal for teachers
- Promotion and career development

Assessment

- Outcome of assessments
- Development and use of assessment tools
- Appropriateness of assessment strategies

Policy

- Determining future medical education policy

Control

- Gaining compliance
- Instilling values
- Surveillance
- Management

Conceptual Models in Evaluation

Conceptual models and frameworks may help to further our understanding of evaluation. Three of these, the 'evaluation cycle', Musick's 'task orientated' approach(6) and Kirkpatrick's hierarchy, are described below.

The evaluation cycle

A simple four-step model of the evaluation cycle is presented in Figure 23.1.(7) It begins with *planning and preparation*, then the *teaching and learning activity*, then the *collection of data about the activity*, and then *reflection and analysis*.

Task-orientated evaluation

Musick(6) provides a task-orientated model of programme evaluation, a model with five steps, outlined in Box 23.2.

Kirkpatrick hierarchy

The Kirkpatrick hierarchy was first described by Donald Kirkpatrick in 1967, as a series of levels of

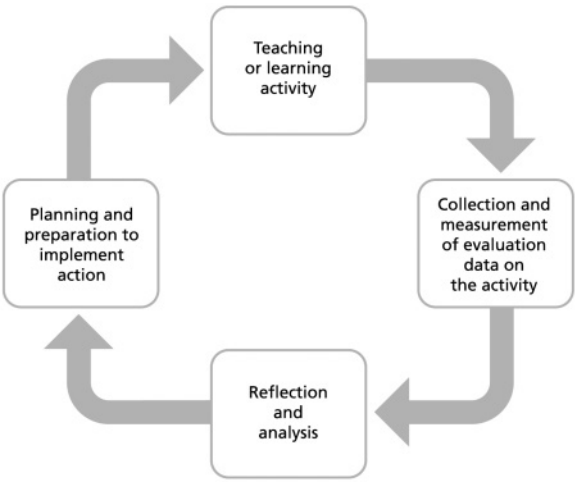

Figure 23.1 A simple four-step model of the evaluation cycle.

evaluation on which to focus questions. These were, at the base (the lowest level), satisfaction with the teaching and learning, next if learning has taken place, then behaviour change and finally, at the top, results of the impact on society (*see* Figure 23.2).

Kirkpatrick's hierarchy(8) has been applied for use in medical education.(9) Here, the lowest level was defined as participation in or completion of the learning, followed by the reaction or satisfaction of participants, the learning or knowledge gained, the changes in health professionals' behaviours, performance or practice, and finally, at the top of the hierarchy, healthcare outcomes (*see* Figure 23.3). Many educational evaluations are at the lower levels of this adapted model. Belfield *et al*.(9) found that in a study of 305 papers, only 1.6% had looked at healthcare outcomes.

The Best Evidence Medical Education (BEME) collaboration has used this model in its various projects

BOX 23.2 Task-orientated model of programme evaluation in graduate medical education(6)

- Examine the evaluation need
  Why are you doing it, and for whom?
- Determine the evaluation focus
  What is to be evaluated?
- Determine the evaluation methodology
  When, where, how to be done and what analyses?
- Present the evaluation results
  Who are the key stakeholders to review results, and when should these be presented?
- Document the evaluation results
  How will the results be documented and used for programme improvement?

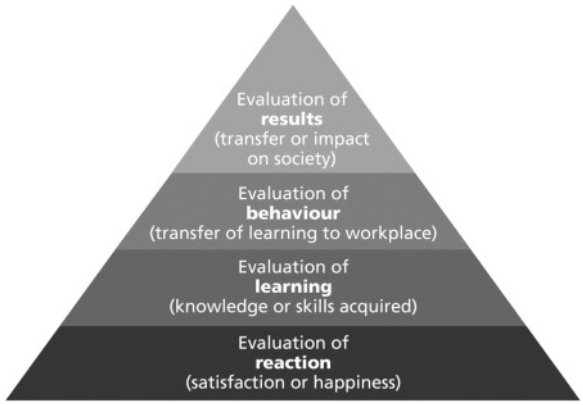

Figure 23.2 Kirkpatrick's hierarchy of evaluation.

to evaluate particular educational strategies. Further information may be found on the BEME web site (http://www2.warwick.ac.uk/fac/med/beme), together with summaries of projects undertaken and in progress, coding sheets for evaluation, published papers and a wealth of other material. Several working groups, all involving multinational cooperation, have now reported.

Evaluation or Research?

Evaluation and research in education are similar activities, and may use the same research methods. Morrison(10) recently made a distinction between the two, claiming that research was aimed at producing generalisable results for publication in peer-refereed literature and needed ethical approval, whereas evaluation was carried out for local use. She also stated that evaluation was a continuous process whereas research may not be continuous if an answer were to be found. My personal view is that this is unhelpful. It is very important to produce well-designed evaluations for

Figure 23.3 Kirkpatrick's hierarchy in medical education.(12)

publication, otherwise we may all be making the same mistakes over and over again in our local situations. For example, if we are to adopt a new assessment tool for 360-degree assessment, this needs to be piloted and properly evaluated in terms of validity, reliability, practicability and so on. Users will want to know how many raters need to be used for reliability, how many fails in any one domain is significant, who are the most appropriate people to fill in the form and so on. To evaluate the new instrument properly we may need to use interviews, questionnaires and focus groups, and analytical tests such as Cronbach's alpha(11) and generalisability theory(12) to answer these evaluation questions. This is exacting work, and not low-key activity purely for local use. Evaluation should be embedded in the culture of an organisation and used to provide feedback at all levels, but at its best, and arguably most useful, it should be conducted with the rigour of the best research project.

The overlap between evaluation and research causes most problems in the area of ethics. Bedward *et al*.(13) in their editorial in *Medical Teacher* in 2004, discussed the lack of clarity on what requires ethical approval, the reliance on one procedure for all applications, confusion over the scope of responsibilities within trusts and the scale of the work involved. In the UK, the National Research Ethics Service has published guidance differentiating research, audit and evaluation advising that only the former requires official research ethics committee approval. These guidelines can be found at http://www.nres.npsa.nhs.uk.

Despite the official view, there are some significant ethical issues in evaluation. In a recent review, Goldie(2) described some of these issues, including:

- doing what the client wants
- adhering to the written contract
- following a certain methodology
- giving equal weight to all opinion without bias
- elitism – allowing more powerful voices to be given greater weight.

Goldie cites seven ethics standards areas for evaluators, drawn from a number of national bodies by Worthen *et al*.(14) (*see* Box 23.3). Ultimately though, it is the individual evaluator's responsibility to work ethically to bring to their work a principles- or virtue-based approach rather than to slavishly follow external codes.

BOX 23.3 Ethical standards in evaluation(14)

- Service orientation
- Formal agreements
- Rights of human subjects
- Complete and fair assessment
- Disclosure of findings
- Conflicts of interest
- Fiscal responsibility

Formative and Summative Evaluation

As with assessment, evaluation may be formative or summative in nature. *Formative* evaluation may be carried out by faculty to improve and shape the teaching, to gather evidence and to feed it back to those planning and carrying out the teaching in order to effect improvements.

Summative evaluation is used – often by powerful bodies with regulatory responsibility – for accreditation and reaccreditation purposes, and for promotion of individuals. These are often high-stakes decisions, such as the approval of the undergraduate curriculum of a new medical school, the accreditation of a training programme or the award of a professorship.

Regardless of the purpose of evaluation, the sources of evidence and methods used will largely be the same. And it is to these issues that we turn next.

Sources of Evaluation Evidence

In terms of evaluating teaching and learning, Berk(15) commented that student ratings have dominated as the primary measure of the evaluation of teaching effectiveness for the past 30 years. Berk(15) made a plea for '… a variety of sources to define the construct and to make decisions about performance …'. Like him, I make a plea for the broadening and deepening of the evidence base with other sources of evidence, so that in any one evaluation we are able to use multiple sources of evidence. He suggested drawing on at least three sources of evidence, so that the strengths of one may compensate for the weaknesses of others (e.g. the different biases of peer ratings and student ratings). He suggested 13 sources of evidence, using a wide range of evaluation methods. These are listed in Box 23.4 and each is discussed in more detail below.

Student (learner) ratings

Here, students or learners evaluate the teaching programme, often by means of evaluation questionnaires. These have been used for many years and are a necessary and essential part of evaluating teaching effectiveness. Research confirms that student ratings are an excellent source of evaluation evidence.(16) However, they should not be relied on by themselves when making important decisions.

Peer ratings

With peer ratings, colleagues look at one's teaching programme, either by sitting in and observing, or reading and reviewing the teaching materials. Peer

BOX 23.4 Sources of evaluation evidence(14)

What	How	Who	Why
Student (learner) ratings	Rating scale	Students	Formative and summative
Peer ratings	Rating scale	Peers	Formative and summative
External expert ratings	Rating scale	Experts	Formative and summative
Self-ratings	Rating scale	Self	Formative and summative
Video recording of teaching	Rating scale	Self and peers	Formative and summative
Student interviews	Rating scale	Students	Formative and summative
Exit and alumni ratings	Rating scale	Graduates	Formative and summative
Employer ratings	Rating scale	Graduates' employer	Programme
Administrator ratings	Rating scale	Administrator	Summative
Teacher scholarship	Review	Administrator	Summative
Teaching awards	Review	Administrator	Summative
Learning outcomes measures	Exams	Administrator	Formative
Teaching portfolios	All above	Students' peers, administrator	Summative

review of teaching and peer review of teaching materials are useful, and cover aspects that the students as learners will not be in a position to evaluate. Since teaching is a scholarly activity, it should be subject to peer review (as with research). There is evidence that the results of peer review compare well with student ratings,(17) so peer ratings complement student ratings. Peer ratings are best used for formative rather than summative evaluation.(16)

External expert ratings

Here, an external expert will look at the teaching programme, as with peer evaluation. Such individuals will be very experienced and highly skilled with the ability to give really good and helpful feedback. This may be of great benefit to younger teachers, although in my own experience this is not always the case.

Self-ratings

Self-evaluation is an important part of the evaluation of a teaching programme. However, much like assessment, where there are differences between teacher and student assessment,(18) superior teachers provide a more accurate self-evaluation than do less highly rated teachers.(19) A highly effective, but rarely executed, evaluation strategy is to triangulate the three sources of evidence from students, peers and self.

Video recording of teaching

Berk(15) suggests that a video is one of the best sources of evidence for formative evaluation decisions. But who should evaluate it, and what methods and criteria should be used? A video recording of a teaching tutorial has been a requirement in UK general practice trainer approval for many years now. However, in hospital practice clinical teachers still seem to be wary of the video method, both in clinical work and in evaluation.

Student interviews

Interviews with groups of students are useful for evaluation purposes and may be considered more accurate and reliable than individual student ratings. Braskamp and Ory(20) suggested three types of group student interview:

- a quality control circle (perhaps what we would call a junior doctors' forum) that meets regularly to review the teaching and feedback comments and suggestions
- classroom group interviews (with someone other than the usual teacher)
- exit interviews with those who have completed the programme.

Exit and alumni ratings

A great strength of the alumni rating is that it can give valuable feedback but from a different perspective to that of current students. An example of the sort of question exit or alumni data can inform might be 'Was your medical school course for purpose?' We did precisely this study and found, perhaps not surprisingly, that it was good in some parts but not in others.(21) The young doctors and their consultant educational supervisors thought they were good at communication skills but less good at diagnosis, decision making and prescribing. Elsewhere, graduates' opinions have been shown to correlate very highly with current students' views, even up to four years later.(22)

Employer ratings

Is the young graduate fit to do the job? What is their performance like once qualified? Studies in this area can evaluate the strengths and weaknesses of the programme, and again can give valuable feedback from another, different perspective. In the study above(21) we also included the consultants. One common theme was that young doctors knew no anatomy, were good at communicating but lacked the basic knowledge with which to communicate with patients.

Administrator ratings

Deans and those in senior positions may evaluate faculty in terms of criteria for promotion, merit awards, discretionary points and so on, as well as for teaching excellence awards. Often they will rely on secondary evidence, rather than direct observation themselves.

Teaching scholarship

By this I mean the academic contribution that people make to the growing body of knowledge in medical education, in terms of development of teaching programmes, research, published papers, books and presentations at conferences. This can discriminate the key educators or star performers from others.(23)

Teaching awards

An example of reward for teaching excellence is the system of awards at Toronto General Hospital by Posluns *et al*.(24) The authors developed a programme to recognise teaching excellence, using evaluations from trainees and departmental administrators. Awards were presented to teachers at a ceremonial event.

Learning outcomes measures

Claims are often made that success in professional examinations are a measure of how good the teaching was in a particular hospital. Superficially this might seem attractive, in view of the present emphasis on outcomes-based curricula.(25) But can we really infer teaching excellence from student performance? In fact, the correlation is low to moderate, only 0.4 in meta-analyses.(26) The big problem is how we isolate teacher input as the sole cause of students' performance at the end of their programme of study. There are many variables in addition to the teaching (such as student ability, motivation, examination standards) that also have an impact on the achievement of outcomes. So in evaluating teaching effectiveness, learning outcome measures should be used with great caution.

Teaching portfolios

Currently portfolios seem to be everywhere in medical education, and increasingly these are 'electronic' and 'online'. Established practitioners use portfolios as collations of evidence towards an annual appraisal; doctors in the foundation and specialty training programmes use them as repositories for personal reflections and workplace-based assessments, and some medical schools have adopted portfolios as an integral part of the course. A teaching portfolio would ideally contain one's best work as a teacher, with evaluations from a variety of sources, and a reflective analysis on the different parts, together with appendices detailing some of the evidence. But how should such a portfolio of evidence be assessed? There are many approaches, with the assessment strategy chosen linked clearly to the portfolio structure. See Chapter 7 for further details.

Evaluation Methods

This section includes brief descriptions of several methods that may be used in educational evaluation:

- questionnaires
- interviews
- focus groups
- site visits
- examples of group methods.

In larger scale evaluations, such as those of nationwide educational programmes, several methods may be used serially, for example, in the case of a focus group being used to generate items for a questionnaire. Or they may be used in parallel where multiple methods are adopted to tap into different data sources in order to build the richest possible picture of the educational initiative under study.

Questionnaires

There are several advantages of using a questionnaire to evaluate a programme of teaching. The questionnaire is feasible and economical in terms of time and effort to collect a range of views from the whole population to be studied, rather than sampling some parts of the population. Questionnaire data (especially for closed rating scale questions) may be analysed using statistical testing for significance and associations between different data, including data reduction techniques such as factor analysis.(27,28) In addition, it may be necessary to use statistical testing to assess reliability of the data, for example, Kappa, Wilcoxon signed-rank test and repeated measures analysis of variance to assess test–re-test reliability, and Cronbach's alpha to assess internal consistency.(29) The questionnaire may also allow a search for new patterns, by two methods. One of these is the use of open-ended questions and free comments, analysed by qualitative methods, where one may catch the 'gem' of information that may be missed by the closed question.(30) The other is the method of principal component factor analysis, a data reduction method to reduce the quantitative data from the Likert questions to a small number of factors with common characteristics.

There are, however, some disadvantages to the questionnaire method. There is a well-recognised problem with pre-coded responses, which may not be sufficiently comprehensive to accommodate all answers, forcing the candidate to choose a view that does not represent their views correctly.(31) We make assumptions that all respondents will understand the questions in the same way, and there is no way of clarifying the question as in a one-to-one interview. Non-response affects the quality of the data and thus the generalisability of the results. Responders may differ from non-responders, in that non-responders may be of lower social class,(32) or older and more ill than responders.(31)

To overcome these problems, an evaluation questionnaire may be designed to include both open questions (to get at the gems), closed questions (yes/no), tick box questions with specified categories, scale items such as the Likert rating scale (agree to disagree points on a scale) and the opportunity for free comments to attempt to catch any other gems there may be.

Questionnaire design is covered in greater detail later on in this chapter.

Interviews

Cohen *et al*.(30) defined a research interview as a 'two-person conversation initiated by the interviewer, for the purpose of gathering research relevant information'.

The interview has several uses within educational evaluation and research. It can be used:

- as a way of gathering information about the evaluation questions
- to develop ideas for new hypotheses
- in conjunction with other evaluation methods in a study
- to validate
- to go deeper and explore new themes generated from other evaluation methods
- to test hypotheses that have already been generated.(30)

Also, the interview method is a powerful way to gain internal validity in case study work, to go deeper and explore new themes generated from other educational evaluation methods in this work.

Focus groups

The focus group is a form of group interview in which discussion and interaction within the group is part of the methodology.(33)

People are encouraged to talk, exchange ideas, tell stories, comment on each other's ideas and ask each other questions. The method is useful in evaluation in exploring learners' knowledge and experiences, and also in determining what they think about the course and why. The idea of a focus group is that it may help to clarify ideas and views that might be less accessible in a one-to-one interview.

Kitzinger(33) gives advice on group composition, running the group (with four to eight people as an ideal number), analysis and writing up. I believe it is essential to record such group activity and transcribe it for detailed analysis. Digital recording using a digital recorder and boundary microphone (for 360-degree capture of what people say) will give good sound quality. In addition, audio files may be stored on computer, burned onto compact disc and sent to participants and colleagues for further comments.

Site visits

This has been a traditional way to evaluate hospital posts and general practice training posts for many years. In general practice, Pereira Gray(34) advocated this in 1981 and remarked that '… some advisers and organisers have now visited as many as a hundred different practices and thus have built up a considerable experience of the characteristics of satisfactory and unsatisfactory training practices …'. The visit to the organisation, speaking with the trainers and the trainees face to face, and seeing the working environment does still have a valuable place, but it is important the process is robust and reproducible, and that visitors are trained properly in their role.

Group consensus techniques

A number of consensus techniques have been developed for the evaluation of educational events involving medium to large groups. Two commonly used approaches are described here.

Snowball review

This is a group-based evaluation(1) that uses a series of steps where comment and opinion are suggested, discussed, shared and agreed, before going on to the next step, until a final list of good and not so good points about a course has been agreed. The steps are as follows.

- Each individual alone lists, say, three good and three not so good points about the course.
- Participants form pairs and discuss their suggestions then come to an agreement as a pair.
- Each pair then forms a group of four, and again debates the views and comes to an agreement.
- Two groups of four join up to form a group of eight. Again they debate and agree their conclusions.
- A reporter presents the group's views to the whole course.

This is a good method in that it involves everyone and ideally reaches a consensus and a conclusion, but it does take time.

Nominal group technique

This is another group-based consensus method. It differs from the snowball review (above) in that each person gives their views and then all the views are collected up and voted on. The steps are as follows.

- Each individual is asked in turn for feedback on the best and least good aspects of the course.
- Comments are collected and listed (once) on a flip chart, that is, if two members of the group thought that the catering was not very good, this is only listed once.
- The facilitator continues to go round the group until all (unique) comments have been exhausted.
- Group members are then allowed a set number of votes to distribute among the items listed.
- The result is a scored and ranked list of feedback comments to which all members of the group have contributed.
- An optional stage is to put people into groups to discuss some or all of the items derived from the voting.

Group consensus techniques carry a health warning. Both methods may achieve answers of good face validity but are of low order as far as Kirkpatrick's hierarchy is concerned and the results do not lend themselves to further analysis in terms of generalisability.

To defend what we do as educators, it is very important to conduct rigorous evaluations that are robust, well designed and measure what they are supposed to measure in a reliable way. The method that lends itself most easily to such an approach is the questionnaire, and we will now look at the design and use of questionnaires in greater detail. You may be fortunate to have a ready-made, valid and reliable tool available 'off the shelf', but you may need to design your own. If you do, here is some evidence-based advice to help you do so.

Designing Questionnaires for Evaluation

If you have decided on an evaluation questionnaire, then how should you proceed? You may know that there is already an existing evaluation questionnaire in use, validated, published in the medical education literature and in the public domain for you to use. You may wish to buy a commercially available rating scale from an organisation if this fits your purposes, and you have the funds to do so. Such scales have been developed and validated, and have good validity and reliability. A number are cited by Berk.(15) Or you may wish to pick statements already available in an item bank. One such source is Purdue University in the US, which has an item bank of over 600 evaluation question items. It may be accessed at the Purdue Instructor and Course Evaluation Service on http://www.cie.purdue.edu/data/pices.cfm.

Caution is needed when taking an off-the-shelf questionnaire and using it in your own organisation. The questionnaire may have been developed in another country, and terms used may have different meanings. The questionnaire may not really focus on what you need to evaluate in your own situation. If this is the case, you might have to develop your own questionnaire, as nothing already available will fit your purpose.

Developing your own evaluation tool

The process of developing your own evaluation instrument involves several steps. You will need to:

- decide what you are going to measure
- choose the right questions to ask
- write the statements
- design the response options (the anchors)
- put it all together and do field tests
- analyse the results.

Berk(15) describes all these steps in great detail. Similar information is given in the classic work on questionnaire design by Oppenheim.(27) The great detail in these texts is outside the scope of this chapter, but here are some general principles that may be of help.

Writing the statements

- Each statement should describe a particular idea.
- Use clear, simple language.
- Each statement should be grammatically correct, have the correct punctuation and avoid double negatives.
- Keep to one concept per question – do not ask two questions in one sentence.
- Avoid unfamiliar words.
- Avoid statements that suggest a single answer.
- Avoid abbreviations, as they may mean completely different things to different people.

Berk(15) provides a checklist of 20 questions to ask in relation to testing your rating scale statements.

Writing the response options – the anchors

One you have the statements, the next step is writing the response options. These may be structured in a number of ways as follows.

- *Likert scale:* a bipolar scale with positive and negative feelings towards the statement, 'strongly agree' to 'strongly disagree'. They may have even numbers, a two-, four- or six-point scale, or odd numbers, a three-, five- or seven-point scale with a middle point meaning 'not sure' or 'uncertain'.
- *Quality scale:* using terms such as 'poor' to 'excellent'.
- *Frequency scale:* asking how often something happens, such as daily, weekly, every two weeks or once a month.
- *Quantity scale:* using sometimes vague terms such as 'all', 'very much', 'often', 'a little' or 'not at all'. All points on the scale should be described.

How many points on the scale should there be? Increasing the number of points increases the reliability, but there is a point of diminishing return after five points,(35) so the nine-point scale we see on some current instruments may be difficult to justify.

Should there be a midpoint – an undecided or uncertain category? A midpoint score gives us little or no information. Berk(15) recommends a four- or six-point scale as these require the responder to come to a decision one way or another. There are inevitably biases in rating scales. A *halo effect* is seen where an overall impression will bias the score on each statement. *End aversion* is where responders avoid the extreme ends of the scale for a question. *Extreme response* bias is the opposite of the previous bias. *Positive response* bias is where responders agree irrespective of the question being asked. Also there may be *gender*, *age* and *racial* biases. Some or all of these may need to be addressed.(15)

Assembling the questionnaire

It is helpful to have a framework to follow when putting together a questionnaire. A list of the main points to consider is provided in Box 23.5.

Paper-based or online?

Although some prefer a paper-based questionnaire, the use of online questionnaires is increasing. Advantages for an online format include less cost(36) and some evidence of a higher response rate than in paper-based studies. There are good correlations between paper-based and online format.(37)

Pre-pilot testing

A pre-pilot is essential. Show the instrument to a few of your friends and colleagues. Can they understand the questions? A questionnaire I designed to evaluate our bursary scheme for university educational qualifications went through four versions before a pilot study was eventually undertaken, as problems were ironed out.

BOX 23.5 How to: Construct a questionnaire

Heading: what it is the questionnaire about?

Statement: what is the purpose of the evaluation?

Directions: for filling the questionnaire in.

Identification: of the course, the specialty, the grade of post (or other relevant identifiers), without which the evaluation may be meaningless.

Questions: a mixture of yes/no questions, tick box categories (such as gender, specialty, years qualified), Likert-type statements, open free text questions and free comment. It is usual to put personal information towards the end, and simple yes/no or tick box categorical questions near the beginning.

Return of the questionnaire: instructions for where to send the completed questionnaire, and by when.

Pilot study

The next step is a pilot study on a small sample, perhaps 5–10 people. Modifications may still need to be made after this stage. This is also an opportunity to check if you can code in the responses and put these into a form for analysis, perhaps with SPSS, a powerful statistical program used worldwide for statistical calculations, and NVivo, a programme for the qualitative analysis of free text (both written and spoken) into themes and concepts.

Field testing

Next would be a field test on a larger sample. Following field testing the questionnaire is now ready to be launched.

Response rates

The response rate is an important source of bias, so efforts need to be made to maximise the number of respondents. Methods known to increase the response rates to postal questionnaires include:

- designing and using a short, simple questionnaire
- providing a covering letter on headed notepaper
- including a reply paid envelope
- following up non-responders with a further questionnaire, covering letter and reply paid envelope.(31)

These procedures were much in line with a recent Cochrane review of the evidence by Edwards *et al.*,(38) which indicated that a short questionnaire, first class post, personalised letters, follow-up contact with a further questionnaire and a letter on university or other headed notepaper were all of value in increasing the response rate.

The judgement about what constitutes a good response rate is not clear cut. Suggestions about response rate in the educational literature vary somewhat. Edwards and Talbot(39) suggested a response rate over 60%. Cohen *et al.*(30) suggested at least 40%, and ideally up to 70–80%. Bowling(31) suggested that a response rate of 75% was generally accepted as good, since above this figure of response the biases from the non-responders assumed much less effect on the results as a whole.

Evaluation Results: Implementation and Influence

No matter how beautifully designed the evaluation strategy, the key question for evaluators is can the results be fed back into the system and acted upon? Put another way, how do we manage change in a medical education system? There is a vast amount of literature in the area of change management that has been well summarised elsewhere.(40,41) Much of the early literature suggested a controlled way of proceeding through the change process in a controlled and

logical way. For example, using a problem-solving approach we might identify a problem, agree the problem, suggest solutions, agree solutions, implement solutions and evaluate the changes. This is very logical and rational, but as recent educational reforms in the UK have demonstrated, rational concepts and models have not been the only ones at work.

The UK policy initiative *Modernising Medical Careers* (42,43) began as an attempt to improve the senior house officer grade. In 2005, a new Foundation Programme for newly qualified doctors was launched, and 2007 saw the introduction of new curricula and assessments for all specialties with a new run-through training structure and junior doctors applying through a national untried online application system. All this has happened with very few evaluations of pilots to test the new ideas, with the exception of the Foundation Programmes. Perhaps as a result, the introduction of reform of specialty training proved a disaster for thousands of junior doctors, and we have yet to experience the effects on patient care.

To try to explain such events it is helpful to rethink the linear model of change referred to above. Mintzberg(44) suggested that change can either be thought of as a pre-planned and logical series of steps or, perhaps more realistically, as an open-ended, ongoing and unpredictable process that aims to align and realign in response to change within society. Political power, the Royal Colleges, the Postgraduate Medical Education and Training Board, the GMC and the British Medical Association were all active players in the *Modernising Medical Careers* process, a massive piece of educational change underpinned [with one or two small exceptions(45)] by almost no evaluation evidence.

So what then is the relationship between evaluation and public policy? Weiss(46) and Swanwick(47) both discuss this interface and highlight that although researchers would wish for a direct and logical link, this is rarely the case. Policy makers rarely base their new policies directly on evaluation decisions. Organisations do not often use knowledge directly to help organisational change. In fact, policy makers tend only to cite evaluation studies that suit their particular line of policy. Direct influence of evaluation on policy is unusual. Instead, a more gradual process of change, the slow incorporation of ideas and methods into policy, is more common, a process that Weiss calls 'enlightenment'. New ideas seep in, percolate through the system, gain ground and may eventually become mainstream. Through this softer process, evaluation can have real and powerful consequences, challenging old ideas, offering new perspectives and helping to influence the policy agenda. Finally, evaluation may be used by policy makers for the less overt purpose of control. Here, evaluation can be used to ensure compliance with a policy, to force a set of values on the medical profession, for surveillance and to manage the whole process. Harland(48) described four aspects to this control function of evaluation: compliance, patterning, surveillance and management.

Pitfalls in Educational Evaluation

There are several common pitfalls to try to avoid in educational evaluation, some of which have been mentioned earlier. Here are some common problems I have encountered over a career of conducting educational evaluation.

Only measuring what is easy to measure

This may be related to following a certain methodology. It might be quite easy to send out a questionnaire and more difficult to do case studies where problems of bullying, negative feedback and so on are the main problem. It may also relate to gathering only one source of evidence, rather than trying to triangulate the evidence from three different areas.

Low response rates

This has already been discussed above. The lower the response rate, the greater the bias of what the non-responders think. This may have a serious effect. So it is important to get a good response rate, of perhaps more than 70%, and to state what the response rate is in any analysis of evaluations carried out.

Poor reliability

Some evaluation instruments have not been properly tested for reliability, or if they have, have been found wanting. The latter situation is very rare, and a lack of proper testing, or even lack of knowledge as to what should have been done, is far more common. It may then be impossible to derive any conclusions, or the conclusions obtained may be unreliable, which is worse when important conclusions are drawn from them. Recently, when I was explaining about reliability of shortlisting procedures, many consultants told me they had never heard of measures of reliability being applied to such situations and were unable to understand what was involved.

Taking your eye off the ball

In many evaluations of previously good hospital posts, by the next time we visited the situation had changed for the worse. This may be because of a change of programme director, changes in the hospital trust organisation and government targets (such as the four-hourly accident and emergency targets).

Taking too much notice of the individual with a misplaced grievance

I have certainly seen examples of this, even on national quality assurance visits. Everyone is happy with the teaching except for one disaffected and sometimes

dysfunctional individual. The visiting team takes all of the disaffected one's concerns seriously, to the exclusion of all other evidence. All this appears in the report and slants the conclusions one way. Further investigations later show much of this to be factually incorrect, the individual never turned up to the teaching, scored very poorly on their 360-degree assessment and later turned out to have a severe mental illness. However, all this is found out after the event – and by then it is too late.

Ethical issues

A number of ethical issues commonly arise, including:

- always doing what the client wants (including producing the desired results)
- slavishly adhering to the written contract
- allowing bias to creep in
- ignoring minority reports
- allowing powerful voices to be given greater weight.

Adherence to standards of good practice, such as those produced by the UK Evaluation Society (*see* Box 23.6), will help to overcome some of these temptations.

Expecting your findings to be implemented widely

Change happens slowly, organically and often unexpectedly. Your evaluation findings may take some time to seep into the ground water of educational practice, or they may be picked up by an enthusiastic policy maker and used as justification for wholesale system reform. Just do not expect rational step-by-step decisions to follow from your evaluation, no matter how robust the design and analysis.

BOX 23.6 Focus on: Good practice in evaluation

A number of organisations have defined standards for good evaluation practice. The European Evaluation Society lists the standards laid out by a number of European countries. The UK version spells out what constitutes good practice for:

- evaluators
- participants
- commissioners
- institutions conducting self-evaluation.

The UK Evaluation Society Guidelines for Good Practice in Evaluation can be downloaded from http://www.evaluation.org.uk

The American Evaluation Association has also produced guiding principles; these are built around five areas.

Systematic inquiry: evaluators conduct systematic, data-based inquiries about whatever is being evaluated.

Competence: evaluators provide competent performance to stakeholders.

Integrity/Honesty: evaluators ensure the honesty and integrity of the entire evaluation process.

Respect for people: evaluators respect the security, dignity and self-worth of the respondents, programme participants, clients and other stakeholders with whom they interact.

Responsibilities for general and public welfare: evaluators articulate and take into account the diversity of interests and values that may be related to the general and public welfare.

The full guidance can be viewed at http://www.eval.org/Publications/GuidingPrinciples.asp

Conclusion

Evaluation is an essential part of medical education, needs to be carried out rigorously and correctly, and is certainly not an inconsequential activity as some have claimed. A good knowledge of many different areas is needed if we are to be able to carry out meaningful evaluations, which can be fed back to improve learning and teaching. At the end of the day, the aspiration is that our evaluations will produce improvements in patient care, an outcome right at the top of Kirkpatrick's hierarchy.

Examples of Evaluations in Medical Education

Example 1: evaluation of a course or curriculum

A widely known method for looking at course or curriculum content in medical education is what many know as Harden's Ten Questions.(49) These questions can be used both in planning a course or curriculum, and to evaluate the course in a systematic way. The 10 questions are described below.

1. What are the needs in relation to the product of the training programme?

   Is the programme or curriculum fit for purpose? To evaluate this, one may consult experts, errors in practice, critical incident reports, task analysis, morbidity and mortality figures, opinions and beliefs of star performers, looking at existing curricula and views of recent students.
2. What are the aims and objectives?

   What will the student be able to do at the end of the course of study? Is this borne out by evaluation data?
3. What content should be included?

   Content should be put into a course if it:
   - directly contributes to the course objectives
   - forms a building block of skill or knowledge needed to tackle a later part of the course

- allows development of intellectual abilities such as critical thinking
- aids the understanding of other subjects on the course.

4 How should the content be organised?

This relates to the order in which subjects are taught, and a theoretical plan of why the order is organised as it is.

5 What educational strategies should be adopted?

'Strategies' relate to the curriculum model being used, such as a spiral curriculum, an objectives model, a process model or an outcome-based model. Sometimes there is no obvious model at all.

6 What teaching methods should be adopted?

Student grouping may be one way, with, for example, whole class teaching by the lecture method, small groups, one-to-one bedside teaching and distance learning. Another way may be by teaching and learning tools, such as computer packages, web-based learning, simulators, skills laboratories and role play. The choice of methods needs to reflect the course aims and objectives. One would not choose to teach communication skills and breaking bad news to patients in a lecture-based course.

7 How should assessment be carried out?

This includes the choice of assessments used [such as essays, projects, portfolios, multiple-choice questions, Objective Structure Clinical Examinations (OSCEs), oral examinations, and long and short cases].

- Who will assess the work?
- Are there external examiners?
- Is there self-assessment?
- Will assessment be continuous throughout the course or at the end?
- What are the standards to be achieved?
- Are the assessment standards criteria referenced or norm referenced?
- How is the course evaluated? Is it by the students?
- Is there internal and external evaluation?

8 How should details of the curriculum be communicated?

Details have to be communicated to those teaching the course, the students attending the course, potential students and other bodies. How is this done? How do the subjects relate to each other and the final product?

9 What educational environment or climate should be fostered?

Does the environment encourage cooperation between students, students and teachers, scholarship, probity and support, or is it hostile, with teaching by humiliation, sexism and bullying?

10 How should the process be managed?

- Who is responsible for planning, organising and managing the process?
- Can changes be made?
- How does this course relate to others?
- Is there student representation?
- Do the teaching staff know what is going on?
- Is there a course committee?

Example 2: evaluation of faculty development programmes

Does faculty development – teaching the teachers – really work? What is the evidence for this? This summary illustrates many different methods of evaluation that have been carried out seeking to answer this question. There is now good evidence that these courses do improve teachers' ability to teach, and do improve learning by trainees. Researchers in medical and dental education have shown that teaching the teachers initiatives really do work.

One of the few studies that tried to do this was undertaken by Whitehouse,(50) who followed up participants who had attended a six-day educational teaching the teachers course. Participants did use the lessons learnt on the course and put many of these ideas into their own education practice. Indeed, the group formed on this course continues to meet more than 10 years on from the original course.

In Turkey, Yolsal *et al*.(51) reported the impact of their Training of Trainers courses since 1997. They used a questionnaire at the end of their six- to nine-day courses, and at six months post-course. Seventy-two per cent of their medical teachers said they had implemented the knowledge and skills acquired on the courses, and that students had given better feedback on their teaching. Many stated they now enjoyed their teaching more, and that they had set up a network of keen teachers after being on courses together. This study and these results are very similar to Whitehouse's study and results.(50)

Steinert *et al*.(52) at McGill University in Canada described a year-long faculty development programme to develop leaders in medical education. The course included educational knowledge and skills, in protected time, while maintaining the participants' other clinical, teaching and research responsibilities. A year after completing the programme, the authors found in a follow-up survey of 22 faculty members that many had joined new educational committees, taken up new leadership roles in medical education and developed new courses for students and doctors in training, and two had pursued further studies to Masters level. However, all three studies relied on self-reporting by the candidates rather than other evidence.

There are more recent studies that used different methodologies and end-point assessments to make similar points. Godfrey *et al*.(53) from Sheffield asked whether a teaching the teachers course did in fact develop teaching skills. They used a quasi-experimental design and compared a group of consultant teachers who underwent a three-day teaching the teachers

course with a control group taken from the course waiting list. However, the candidates were not assigned to the taught group or the control group randomly. A questionnaire of teaching skills was applied to participants and controls before the course and at 8 to 10 months afterwards. Up to 63% of participants and 51% of controls replied to all aspects of this study. Candidates who had attended the course did significantly better and reported significantly greater improvements in self-reported teaching skills. These authors did discuss self-reporting as a source of bias, but argued that the control group in this research design did help overcome these issues.

In a Californian study, Morrison *et al*.(54) used an experimental design with an intervention group and a control group of medical residents to explore whether they could improve residents' teaching skills. Their research question was: 'Do trained teachers perform better then untrained control residents?' They ran a 13-hour teaching the teachers course over six months. This was a rigorously designed and carried out study. The two groups were similar in terms of gender, specialty and academic performance, and people were assigned randomly to each group, either to attend the course or be a control. The outcome measures included an eight-station, three-and-a-half-hour OSCE – previously validated – in which the participants were assessed by trained medical students. A subset of the participants were interviewed one year later by two educational researchers, who did not know in advance who had attended the course and who was a control. Those residents in the teaching the teachers course group did significantly better in all eight stations of the OSCE, and people who were required to attend did as well as those who volunteered. Also, the interviews showed that the taught group showed greater enthusiasm for teaching, used learner-centred approaches, had more elaborate understanding of pedagogic principles and planned to teach after finishing their training.

In Alberta, Panderchuck *et al*.,(55) in another quasi-experimental study, looked at the effects of a two-day teaching the teachers workshop. They compared a group of medical teachers who had attended the workshops with controls (who had not) by means of ratings by medical students on their teaching abilities before and after the workshops, using the standard University of Alberta student evaluation questionnaire. Their study used data from 1993 to 2002. Students' ratings of teachers' teaching abilities increased significantly for the teachers after the workshops, but remained unchanged for the control group, who had not attended the workshops.

Most recently, in November 2006, new evidence from the BEME review of faculty development initiatives by Steinert *et al*.,(56) which reviewed 2777 papers and selected 53 of these papers for detailed analyses, showed:

- overall satisfaction with programmes was high
- participants reported positive changes in attitude to teaching
- increased knowledge of educational principles and of gains in teaching skills
- changes in teaching behaviours
- greater educational involvement and establishment of collegiate networks.

Key strategies for effective interventions to improve teaching effectiveness in medical education included:

- experiential learning
- feedback
- effective peer and colleague relationships
- interventions that followed the principles of teaching and learning
- the use of a wide variety of educational methods.

In summary, much of the earlier literature is based on opinion from distinguished and senior individuals within the profession, both individually and in terms of organisations such as expert government committees, Royal Colleges and the GMC.(57,58) More recently there are examples of good evaluation studies that try to examine the effects of teaching the teachers courses by means of follow-up and self-reporting of changes, and of the effects of such initiatives on the learners, the medical students themselves. Here there is evidence that teaching the teachers courses do improve teaching ability, widen teachers' pedagogic approaches to teaching and increase enjoyment of teaching.

Example 3: evaluation of the educational climate

The educational environment variously referred to as climate, atmosphere or tone is a set of factors that describe what it is like to be a learner within that organisation. Chambers and Wall(59) considered the educational climate in three parts. These were:

- the physical environment (safety, food, shelter, comfort and other facilities)
- the emotional climate (security, constructive feedback, being supported, and absence of bullying and harassment)
- the intellectual climate (learning with patients, relevance to practice, evidence-based, active participation by learners, motivating and planned education).

The good clinical teaching environment(60) should ensure the teaching and learning is relevant to patients, has active participation by learners, and shows professional thinking and behaviours. There should be good preparation and planning of both the structure and content, reflection on learning, and evaluation of what has happened in the teaching and learning. Spencer also goes on to describe some of the common problems with teaching and learning in the clinical environment, including lack of clear objectives, focus on knowledge rather than problem-solving skills, teaching at the wrong level, passive observation,

little time for reflection and discussion, and teaching by humiliation.

Teaching by humiliation, bullying and harassment is a big problem in the UK and other countries. Much of this relates to teachers' lack of awareness of educational skills and knowledge(61,62) and inability to promote a good, supportive educational climate for trainees in which to learn.(57,58) Lowry(61) described disenchantment with medicine in the words of a young doctor as 'It could have been such a wonderful thing to be a doctor – but it's not. It's just a disaster.'

Sadly, there are many examples in the literature of bullying and harassment of junior doctors and medical students. Such studies show the practice is widespread, and on the individual level illustrate how destructive to confidence and well-being bullying and harassment can be. Wolf *et al*.(63) carried out a questionnaire study of medical students in the Louisiana State University School of Medicine. Of these, 98.9% reported mistreatment, with shouting and humiliation being most frequent. Over half reported sexual harassment, which was reported mainly by women students. There was a high level of remarks degrading doctors and medicine as a profession. Increased mistreatment was associated positively with a perceived increase in cynicism.

In Manchester, Guthrie *et al*.(64) used a postal questionnaire to measure the psychological morbidity and the nature and sources of stress in first-year medical students. Half the students reported a stressful incident, most of which related to medical training and the styles of medical teaching. They described being upset by the attitudes of tutors, including humiliation, shouting, ridicule, ignorance exposed and a confrontational nature.

Metcalfe and Matharu(65) used a postal questionnaire study of medical students in Manchester to investigate student perceptions of good and bad teaching. Good examples were when there was active learning, and teachers let students 'run the session'. Bad examples were mainly examples of bad behaviour of staff towards students, such as humiliation, sexism and ridicule.

Is this all in the past? In the 21st century, are all these problems behind us? The answer is probably no. The medical literature still reports bullying and harassment. Doctors are still disillusioned. Bullying and humiliation are still common within the medical profession. An anonymous author(66) described repeated bullying of a junior surgeon by a consultant. A questionnaire study by Quine(67) of bullying of junior doctors shows how common such behaviours still are. Of 594 doctors, 37% identified themselves as having been bullied in the past year, and 84% had experienced one or more bullying behaviours. Black and Asian doctors fared worse than white doctors, and women fared worse than men. The most common bullying strategies were attempts to belittle and undermine work, unjustified criticism, humiliation in front of colleagues, and intimidatory use of discipline and competence procedures.

Despite this, the evaluations from the trainees at senior house officer level report good to excellent scores for induction, supervision and clinical experience in these posts. This is using the post-evaluation tool, developed and used every six months on all senior house officer posts within the West Midlands since 1997.(68) Also, evidence from our Deanery Quality Assurance visits to intensive care units has repeatedly shown that senior house officers value their time there very much, feel well supported and rate the jobs highly.

So can the education environment be measured using a practical, valid and reliable tool? The Dundee Ready Education Environment Measure was developed in Dundee by Roff *et al*.(69) It is a valid and reliable measure of the perceived education environment. It has been widely used in many countries throughout the world, including the Gulf States, Nepal, Nigeria, the West Indies, Thailand, China, Canada and the UK. It has been used for medical students, nurses, dentists, chiropractors and other professions allied to medicine.

From this original environmental measure, further postgraduate measurement scales have been developed. For the postgraduate medical education environment, the Dundee group has also developed a 40-item inventory for the various aspects of junior doctor training in the UK and Ireland, the Postgraduate Hospital Education Environment Measure.(70) The Anaesthetic Theatre Educational Environment Measure(71) has been developed and validated for teaching in anaesthesia in the operating theatre, and for junior surgeons in the operating theatre; the Surgical Theatre Environmental Educational Measure(72) has been developed and validated.

In anaesthetic trainees, Holt and Roff(71) showed that there was a better educational climate perceived by first-year senior house officers than by specialist registrars, and these results were statistically significant. In the operating theatre, Nagraj *et al*.(73) showed that the educational climate was reasonably good for medical students and slightly less so for senior house officers. Only at specialist registrar level did the climate improve significantly. So, educational climate can be measured using a variety of well-researched, valid and reliable tools that are available free of charge in the public domain. One of the great assets of using these tools is that comparisons may be made of your own evaluations with others throughout the UK and indeed other parts of the international medical education community.

References

1 Mohanna K, Wall D and Chambers R (2004) *Teaching Made Easy – a manual for health professionals* (2e). Radcliffe Medical Press, Oxford.

2 Goldie J (2006) AMEE guide no. 29: evaluating educational programmes. *Medical Teacher.* **28**: 210–24.
3 Norcini JJ, Blank LL, Duffy FD and Fortna GS (2003) The miniCEX: a method of assessing clinical skills. *Annals of Internal Medicine.* **138**: 476–83.
4 SCOPME (1996) *Appraising Doctors and Dentists in Training*. Standing Committee on Postgraduate Medical and Dental Education, London.
5 General Medical Council (2001) *Good Medical Practice*. GMC, London.
6 Musick D (2006) A conceptual model for program evaluation in graduate medical education. *Academic Medicine*. **81**: 759–65.
7 Wilkes M and Bligh J (1999) Evaluating educational interventions. *British Medical Journal.* **318**: 1269–72.
8 Kirkpatrick DI (1967) Evaluation of training. In: Craig R and Mittel I (eds) *Training and Development Handbook*, pp. 87–112. McGraw Hill, New York.
9 Belfield CR, Thomas HR, Bullock AD *et al*. (2001) Measuring effectiveness for Best Evidence Medical Education – a discussion. *Medical Teacher.* **23**: 164–70.
10 Morrison J (2003) Evaluation. *British Medical Journal.* **326**: 385–7.
11 Field A (2004) *Discovering Statistics Using SPSS for Windows* (2e). Sage Publications, London.
12 Brennan RL (2001) *Generalizability Theory*. Springer, New York.
13 Bedward J, Davison I, Field S and Thomas H (2005) Audit, educational development and research: what counts for ethics and research governance? *Medical Teacher.* **27**: 99–101.
14 Worthen BL, Sanders JR and Fitzpatrick JL (1997) *Program Evaluation: alternative approaches and practical guidelines* (2e). Longman, New York.
15 Berk RA (2006) *Thirteen Strategies to Measure College Teaching*. Stylus Publishing LLC, Sterling, VA.
16 Arreola RA (*2000*) *Developing a Comprehensive Faculty Evaluation System: a handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system* (2e). Anker, Bolton, MA.
17 Murray HG (1983) Low inference classroom teaching behaviours and student ratings of college teaching effectiveness. *Journal of Educational Psychology.* **71**: 856–65.
18 Falchikov N and Boud D (1989) Student self-assessment in higher education: a meta-analysis. *Review of Educational Research.* **59**: 395–430.
19 Barber LW (1990) Self assessment. In: Millman J and Darling-Hammond L (eds) *The New Handbook of Teacher Evaluation*, pp. 216–28. Sage, Newbury Park, CA.
20 Braskamp LA and Ory JC (1994) *Assessing Faculty Work*. Jossey-Bass, San Francisco, CA.
21 Wall D, Bolshaw A and Carolan J (2006) Is undergraduate medical education fitting for purpose to be a pre-registration house officer? *Medical Teacher.* **28**: 435–9.
22 Overall JU and Marsh HW (1980) Students' evaluation of instruction: a longitudinal study of their stability. *Journal of Educational Psychology.* **72**: 321–5.
23 Harden RM (1986) Approaches to curriculum planning. *Medical Education.* **20**: 458–66.
24 Posluns E, Sharon Safir M, Keystone JS *et al*. (1990) Rewarding medical teaching excellence in a major Canadian teaching hospital. *Medical Teacher.* **12**: 13–22.
25 Harden RM, Crosby JR and Davis MH (1999) An introduction to outcomes based learning. *Medical Teacher.* **21**: 7–14.
26 Feldman KA (1989) The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*. **30**: 583–645.
27 Oppenheim AN (1992) *Questionnaire Design, Interviewing and Attitude Measurement*. Continuum, London.
28 Field A (2000) *Discovering Statistics Using SPSS for Windows*. Sage Publications, London.
29 Howitt D and Cramer D (2003) *A Guide to Computing Statistics with SPSS 11 for Windows*. Pearson Education, Harlow.
30 Cohen L, Manion L and Morrison K (2000) *Research Methods in Education* (5e). RoutledgeFalmer, London.
31 Bowling A (1997) *Research Methods in Health*. Open University Press, Buckingham.
32 Cartwright A (1988) Interviews or postal questionnaires? Comparisons of data about women's experiences with maternity services. *The Milbank Quarterly.* **66**: 172–89.
33 Kitzinger J (1996) *Introducing Focus Groups*. In: Mays N and Pope C (eds) *Qualitative Research in Health Care*, pp. 36–45. BMJ Publishing Group, London.
34 Pereira Gray DJ (1982) *Training for General Practice*. MacDonald and Evans, Plymouth.
35 Myford CM (2002) Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education.* **15**: 187–215.
36 Bothell TW and Henderson T (2003) Do online ratings of instruction make sense? In: Sorensen DL and Johnson TD (eds) *New Directions for Teaching and Learning*, no. **96**, pp. 69–80. Jossey-Bass, San Francisco, CA.
37 Johnson TD (2003) Online student ratings: will students respond? In: Sorensen DL and Johnson TD (eds) *New Directions for Teaching and Learning*, no. **96**, pp. 49–60. Jossey-Bass, San Francisco, CA.
38 Edwards P, Roberts I, Clarke M *et al*. (2002) Increasing response rates to postal questionnaires: systematic review. *British Medical Journal.* **324**: 1183–5.
39 Edwards A and Talbot R (1999) *The Hard-Pressed Researcher* (2e). Pearson Education, Harlow.
40 Handy C (1989) *The Age of Unreason*. Business Books, London.
41 Iles V and Sutherland K (1997) *The Literature on Change Management. Review for Healthcare Managers, Researchers and Professionals*. National Co-ordinating Centre for NHS Service Delivery and Organisation, London.
42 Donaldson L (2002) *Unfinished Business: proposals for the reform of the SHO grade. A Report by Sir Liam Donaldson, CMO for England*. Department of Health, London.
43 Department of Health (2004) *Modernising Medical Careers: the next steps*. Department of Health, London.
44 Mintzberg H (1987) Crafting strategy. *Harvard Business Review.* **65**(4): 66–75.
45 Walzman M, Allen M and Wall D (2008) General practice and the Foundation Programme: the views of FY2 doctors from the Coventry and Warwickshire Foundation School. *Education for Primary Care*. **19**(2): 151–9.
46 Weiss C (1999) The interface between evaluation and public policy. *Evaluation.* **5**: 468–86.
47 Swanwick T (2007) Introducing large-scale educational reform in a complex environment: the role of piloting and evaluation in modernizing medical careers. *Evaluation.* **13**: 363–73.
48 Harland J (1996) Evaluation as Realpolitik. In: Scott D and Usher R (eds) *Understanding Educational Research*. Routledge, London.
49 Harden RM (1986) Ten questions to ask when planning a course or curriculum. *Medical Education.* **20**: 356–65.
50 Whitehouse A (1997) Warwickshire consultants' 'training the trainers' course. *Postgraduate Medical Journal*. **73**: 35–8.

51 Yolsal N, Bulut A, Karabey S *et al*. (2003) Development of training of trainers' programmes and evaluation of their effectiveness in Istanbul, Turkey. *Medical Teacher.* **25**: 319–24.
52 Steinert Y, Nasmith L, McLeod PJ and Conochie L (2003) A teaching scholars program to develop leaders in medical education. *Academic Medicine*. **78**: 142–9.
53 Godfrey J, Dennick G and Welsh C (2004) Training the trainers: do teaching courses develop teaching skills? *Medical Education.* **38**: 844–7.
54 Morrison EH, Rucker L, Boker JR *et al*. (2004) The effect of a 13 hour curriculum to improve residents' teaching skills – a randomised trial. *Annals of Internal Medicine.* **141**: 257–63.
55 Panderchuck K, Harley D and Cook D (2004) Effectiveness of a brief workshop designed to improve teaching performance at the University of Alberta. *Academic Medicine.* **79**: 798–804.
56 Steinert Y, Mann K, Centeno A *et al*. (2006) A systematic review of faculty development initiatives designed to improve teaching effectiveness in medical education: BEME guide no. 8. *Medical Teacher.* **28**: 497–526.
57 SCOPME (1992) *Teaching Hospital Doctors and Dentists to Teach: its role in creating a better learning environment. Proposals for Consultation – full report*. Standing Committee for Postgraduate Medical Education, London.
58 SCOPME (1994) *Creating a Better Learning Environment in Hospitals 1. Teaching Hospital Doctors and Dentists to Teach*. Standing Committee on Postgraduate Medical Education, London.
59 Chambers R and Wall D (2000) *Teaching Made Easy: a manual for health professionals*. Radcliffe Medical Press, Oxford.
60 Spencer N (2003) The clinical teaching context: a cause for concern. *Medical Education.* **37**: 182–3.
61 Lowry S (1993) Teaching the teachers. *British Medical Journal.* **306**: 127–30.
62 Lowry S (1992) What's wrong with medical education in Britain? *British Medical Journal.* **305**: 1277–80.
63 Wolf TM, Randall HM, Von Almen K and Tynes LL (1991) Perceived mistreatment and attitude change by graduating medical students: a retrospective study. *Medical Education.* **25**: 182–90.
64 Guthrie EA, Black D, Shaw CM *et al*. (1995) Embarking on a medical career: psychological morbidity in first year medical students. *Medical Education.* **29**: 337–41.
65 Metcalfe DH and Matharu M (1995) Students' perceptions of good and bad teaching: report of a critical incident study. *Medical Education.* **29**: 193–7.
66 Anon (2001) Personal view. Bullying in medicine. *British Medical Journal.* **323**: 1314.
67 Quine L (2002) Workplace bullying in junior doctors: questionnaire survey. *British Medical Journal.* **324**: 878–9.
68 Wall DW, Woodward D, Whitehouse A *et al*. (2001) The development and uses of a computerised evaluation tool for SHO posts in the West Midlands Region. *Medical Teacher.* **23**: 24–8.
69 Roff S, McAleer S, Harden RM *et al*. (1997) Development and validation of the Dundee Ready Educational environment Measure (DREEM). *Medical Teacher.* **19**: 295–9.
70 Roff S, McAleer S and Skinner A (2005) Development and validation of an instrument to measure the postgraduate clinical learning environment for hospital based junior doctors in the UK. *Medical Teacher.* **27**: 326–31.
71 Holt M and Roff S (2004) Development and validation of the Anaesthetic Theatre Educational Environment Measure (ATEEM). *Medical Teacher.* **26**: 553–8.
72 Cassar K (2004) Development of an instrument to measure the surgical operating theatre learning environment as perceived by basic surgical trainees. *Medical Teacher.* **26**: 260–4.
73 Nagraj S, Wall D and Jones E (2006) Can STEEM be used to measure the educational environment within the operating theatre for undergraduate medical students? *Medical Teacher.* **28**: 642–7.

Further Reading

Here are some texts I have found useful, both as reference guides and for further information on a day-to-day basis.

BERA (2004) *Revised Ethical Guidelines for Educational Research (2004)*. British Educational Research Association, Southwell. (http://www.bera.ac.uk/files/guidelines/ethica1.pdf; accessed 1 December 2006.)
Bowling A (1997) *Research Methods in Health*. Open University Press, Buckingham.
Brennan RL (2001) *Generalizability Theory*. Springer, New York.
Cohen L, Manion L and Morrison K (2000) *Research Methods in Education* (5e). RoutledgeFalmer, London.
Cramer D (2003) *Advanced Quantitative Data Analysis*. Open University Press, Maidenhead.
Dent JA and Harden RM (2005) *A Practical Guide for Medical Teachers* (2e). Elsevier Churchill Livingstone, Edinburgh.
Edwards A and Talbot R (1999) *The Hard-Pressed Researcher* (2e). Pearson Education, Harlow.
Field A (2004) *Discovering Statistics Using SPSS for Windows* (2e). Sage Publications, London.
George J and Cowan J (1999) *A Handbook of Techniques for Formative Evaluation*. Kogan Page, London.
Haig A and Dozier M (2003) BEME guide no. 3: systematic searching for evidence in medical education. Part 1: sources of information. *Medical Teacher.* **25**: 352–63.
Haig A and Dozier M (2003) BEME guide no. 3: systematic searching for evidence in medical education. Part 2: constructing searches. *Medical Teacher.* **25**: 463–84.
Howitt D and Cramer D (2003) *A Guide to Computing Statistics with SPSS 11 for Windows*. Pearson Education, Harlow.
Oppenheim AN (1992) *Questionnaire Design, Interviewing and Attitude Measurement*. Continuum, London.
Pell G (2005) Use and abuse of Likert scales. *Medical Education.* **39**: 970.
Pereira Gray DJ (1982) *Training for General Practice*. MacDonald and Evans, Plymouth.
Shavelson RJ and Webb NM (1991) *Generalizability Theory – a primer*. Sage Publications, London.