QUESTIONS TO ASK ABOUT DATA COLLECTION INSTRUMENTS*

Questions	Comment
1 Has the instrument been validated at all?	The results of research which uses *ad hoc* methods or instruments which have not been trialled, tested, or validated must be treated with a lower level of confidence. The results may be correct, but readers will have less opportunity for making a considered judgement than where research involves validated instruments.
2 Validity (see Chapter 6, section 7) Is the instrument valid – does it measure what it is supposed to measure? Does the instrument merely (or even) have face validity, or has it been validated for content, criterion, or construct validity?	There are at least four notions of validity: **Face validity** – the items appear to be relevant to the topic at hand. **Content validity** – there is an adequate coverage of all relevant aspects of the phenomenon of interest. **Criterion validity** – scores on the instrument will correlate with scores on some other instrument allegedly measuring the same thing. **Construct validity** – the extent to which a score on an instrument measures what it is supposed to measure. Usually tested in terms of whether the instrument gives measures as predicted by theory which defines the entity in question.
3 Reliability (see Chapter 6, sections 5 and 6) Is the instrument reliable – does it produce consistent results?	**Test–retest reliability** The instrument will produce the same results when retested in the same population (assuming that what is being measured stays the same).

* If the instrument is a questionnaire used for survey purposes, see also 'Questions to Ask about Surveys'.

Has the instrument been validated for retest reliability and, if appropriate for inter- or intra-rater reliability?	**Inter- and intra-rater reliability** The instrument produces the same results when used by different people, or by the same person on a different occasion. Usually measured by the Kappa coefficient: the higher the κ, the greater the reliability (see Chapter 6, section 6). **Internal consistency reliability** Items measuring a single attribute within an instrument should demonstrate at least some inter-correlation – usually measured with Cronbach's alpha. The higher the alpha score, the greater the internal consistency.
4 Discriminant validity Is the instrument sensitive (responsive) to the differences it is supposed to measure: between people, or changes over time?	If small differences are being looked for the measuring instrument must be calibrated finely. Beware floor and ceiling effects, where change below a threshold or above a threshold is not detected (see Chapter 6, section 4).
5 Robustness Is the instrument insensitive to effects that are irrelevant to what is being measured?	For example, would the measurement be affected by who was doing the measurement or the place in which it was done?
6 Acceptability	Instruments need to be acceptable to both those about whom data are being collected and those who do the measuring: time taken, comprehensibility, perceived relevance, cultural appropriateness and costs. Non-response rates to particular questions are sometimes used as a measure of acceptability (or 'feasibility').
7 Appropriateness to context and targets	Instruments designed for hospital use may not be suitable for domiciliary use. Instruments designed for young people may not be appropriate for older people. Has the instrument been trialled, or better validated, *in the setting for the clients* it is now being used with?
8 Cross-cultural transferability	Is it cross-culturally transferable. Will it translate?
9 Data level Will the scores be sufficiently precise for the study? (see Chapter 6, sections 3 and 4)	If scores need to be continuous for statistical analysis, does the instrument provide these? If it only provides 'yes' and 'no' answers, is this adequate for the purposes of the study? Do the data produced take a normal distribution/can they be transformed into a normal distribution if it is necessary to use parametric statistics see Chapter 6, section 4, and Chapter 7, section 6; see notes on 'floor' and 'ceiling effects' in (**4** above)

QUESTIONS TO ASK ABOUT EXPERIMENTS

Questions and issues	Internal validity You judge whether the experiment seemed true in its own terms (see Chapter 5, section 11)	External validity/ generalisability You judge whether the same results could be achieved in practice and whether it would be worth trying to do so (see Chapter 5, section 12)
1 Was/were the question(s) to which the research was addressed clearly articulated? Can you see what this research is about?	Unless the study addresses clearly articulated questions it will be impossible to judge whether it has answered them satisfactorily.	Are the questions addressed relevant to decisions which might be made in practice?

2 Was this an experiment involving only one individual subject? If yes, this checklist is not appropriate. Read Chapter 5, section 13 instead.
If otherwise, go to **3**

3 From what population was the sample drawn? Where data collection involved a survey technique, see 'Questions to Ask about Surveys'. What were the principles and processes for selecting *all* subjects for the experiment? Is it clear what wider population these subjects represent?	These are questions of external validity – see right-hand column.	Only insofar as the sample was a representative one can the results be generalised to the population from which the subjects were drawn (on representative samples see Chapter 10). But the sample of subjects for the experiment is unlikely to be representative of the case mix in any practice. Do researchers provide enough information about subjects for practitioners to be able to extrapolate the results to their own practices?
4 What was the unit of sampling? Was it appropriate to the questions asked? Alternatively ask, 'Was what happened to each subject independent of what happened to each other?' (subjects might be individuals, ulcers, clinics etc.) (see Chapter 5, section 4)	In experiments on individually delivered interventions the individual client is the appropriate unit of sampling. But where interventions are delivered at group or community level (e.g., group therapy), the outcomes for one will be influenced by the behaviour of all. A therapy group of six should then be regarded as a sample of 1, and sample size calculated accordingly (see **5** below).	Where the appropriate sampling unit is the individual there are no particular problems of application. But the results of group treatment trials are particularly problematic for application, since in order to reproduce the research results in practice a practitioner may need to reproduce the group dynamics which developed during the research.
5 What was the size of the sample? Was it big enough to accommodate: • all the relevant diversity among subjects, both within and between arms of the experiment • any diversity of treatment within arms of the trial • the diversity of outcomes on the chosen outcome measures and to detect differences of the size of interest? (see Chapter 7, sections 7 and 8)	The more the relevant diversity among subjects, and between treatments within arms of a study, the more points of measurement on a measurement scale, the smaller the difference of interest and the rarer the event of interest the bigger the sample needs to be. Statistical significance is reduced in small samples, perhaps giving spurious results – usually spuriously non-significant results.	If the sample was chosen to be representative of a wider population, a larger sample is likely to be more representative, but see **3** above. A large sample has a better chance of including a wider range of the types of clients a practitioner might encounter in practice.

6 Was the sample divided into comparison groups selected to be rather similar but treated in different ways (an experimental/treatment group and a control/alternative treatment group)?

If yes, go to **8**
If no, go to **7**

7 Were there comparison groups selected to be different from each other according to some criteria, but otherwise similar, subjected to the same treatment to see whether it had different effects on different kinds of subjects?
[Note that this design may be nested within the structure described in **6** above: a factorial design (Box 5.1, Chapter 5). If so, go to **8**.]

Otherwise, if yes, go to **10**
If no, go to **11**

8 How were the comparison groups formed?

If by randomisation, go to **13**
If otherwise, go to **9**

9 This is probably a matched pairs, or a matched group/reference area design or groups have been formed by purposive sampling (see Box 5.1 in Chapter 5).

Is the way comparison groups were formed fully explained? Are you satisfied that the variables which should have been evened out between groups were actually evened out? Differences in outcomes for groups may be due to differences which were in the groups in the first place rather than due to differences of intervention.

Was a cross-check made to establish the equivalence of groups?

All other things being equal, experimental designs that do not create comparison groups by randomisation are a less secure basis for generalisation than those that do (Chapter 5, section 3). Properly conducted matched pairs designs are preferred to matched group/reference area designs.

Go to **14**

10 Was the purpose of the experiment to see how different kinds of people responded to *the same* intervention? If yes, check whether the sample seemed large enough to accommodate at least 10 and preferably more of each kind of person. The reason why the researcher didn't divide them into sub-groups in the first place will probably be because it was unknown at the outset as to which kinds of people would respond differently and the experiment was to find this out.

Go to **11**

11 Was the purpose of the experiment simply to see if there was a diversity of response to the same intervention, without being interested in who responded in what way? This is a common design for investigating practitioner decision-making, a group of practitioners being confronted with the same client/case notes/ diagnostic test results and asked to make a judgement about them. The results are usually analysed in terms of statistical correlations (see Chapter 6, section 6).

If yes, go to **16**
If no, go to **12**

12 This is probably a pre–post trial without controls (see Box 5.1 in Chapter 5).

A pre–post trial without controls, where what happened to one subject was not independent of what happened to another, is probably not worth reading (see **4** above).

Subjects (groups, communities etc.) are being compared with themselves before and after an intervention. Does the author make a convincing case that measurements post-trial were caused by the intervention and were not due to being involved in a research project (an experiment effect) or due simply to the passage of time?

Unless the pre–post differences are very large and/or the same differences have been demonstrated in other studies, it is speculative to generalise from the results of pre–post trials (without controls).

Go to **17**

13 Randomisation
Was the sample divided into groups/ between arms by a truly random procedure? (see Chapter 5, section 3)

(Randomisation may be improved by stratification – see Chapter 5, section 3)

Randomisation is used to ensure that each group is a representative sample of *all subjects in the trial* (not of the population from which the sample was drawn). Deficiencies in randomisation may produce treatment and control groups which are different in important ways from each other, the differences filtering through to create spurious differences on an outcome measure (see Figure 5.1, Chapter 5).

In routine practice practitioners do not select clients for particular treatments at random. Their selection criteria may produce results which are different from those reported in a randomised study even if they 'do the same'. However, since there are no controls in practice, this may not be apparent.

Go to **14**

14 Were the comparison groups checked for their similarity at the outset?

Were the characteristics of the subjects clearly described?

Did one of the groups have more extreme baseline scores than the other(s)?

If the comparison groups are dissimilar, then these dissimilarities are likely to feed through to outcome differences creating spurious results. Groups with extreme scores at the outset set up the conditions for regression to the mean. If one group has more extreme scores than another this may feed through to outcome measurements

Unless practitioners are told about the subjects, practitioners will not know how like or unlike they were compared to the clients they deal with.

	to give spurious differences in 'improvement' or 'deterioration' (see Chapter 5, section 6).	
Go to **15**		
15 Did the design include planned 'cross-over'? Were the same subjects subjected to a sequence of treatments? (see Box 5.1, Chapter 5 and Chapter 5, section 13)	With cross-over were the subjects and the experimenter blind to the treatments being given? Was it certain that carry-over effects were not confounding the results? With random and blind assignment of treatments and proper management of carry-over effects this is a very powerful research design.	For n-of-1 experiments and single case evaluations (Chapter 5, section 13) generalisation is not the main purpose, though these designs do demonstrate the diversity of response to the same treatment. For group trials generalisation will depend on the similarity of the practice case mix with that of the subjects researched.
Go to **16**		
16 Drop out and unplanned cross-overs Were the characteristics of any unplanned cross-overs, and any subjects lost to the experiment recorded? Were the effects of lost subjects on the results estimated?	Did the loss of subjects from any group make the groups no longer similar?	If differences in outcome were actually produced by differential drop out then the results cannot be generalised.
Are the results expressed in terms of 'intention to treat'? (Chapter 5, section 8)	[Intention to treat] In the final analysis were results for those who were switched from one arm of the trial to another, and for those withdrawn counted as results for the arm to which they were originally allocated?	Results in terms of intention to treat may be closer to what happens in practice.
Go to **17**		
17 Subject reactivity (see Chapter 5, section 5) What were the subjects' understandings of the experiment: how might these have influenced the results?	Where subject interpretations and understandings influence the results of only one arm of the trial this may create spurious outcome differences.	If, as usually happens, all subjects know they are involved in an experiment this may affect outcomes for *all* arms of the trial. Where such effects implicate all subjects this may

Were the subjects blinded as to the intervention they received? Were they successfully blinded? Were Hawthorne/experiment effects likely – changes due to the fact of being researched, rather than to the intervention itself? (Chapter 5, section 12)	Reactivity is a particular problem where the judgements of subjects are used as the basis for outcome measures.	invalidate the generalisability of the results. In routine practice clients should know what treatment they are being given. In this respect they differ from a blinded group. Clients in the know may be expected to respond differently from blinded subjects.
Go to **18**		
18 Subject compliance Did the experiment depend on subjects doing what they were supposed to do? If so, what means were taken to judge whether subjects had been compliant?	Hidden non-compliance may give misleading results.	In routine practice non-compliance is common. Trials that enforce compliance successfully may give misleading results for practice where compliance cannot be enforced.
Go to **19**		
19 Practitioner/researcher judgements (see Chapter 5, section 5) Were practitioners/researchers blinded as to the interventions received by subjects? If so, were they successfully blinded? If not, how might their knowledge of group membership have altered their behaviour towards subjects, perhaps confounding the results? If practitioner judgements were used in making baseline or outcome measures, what	Blinding practitioners/researchers is sometimes impossible and sometimes ineffective. Is there any reason to suppose that the practitioners in the research treated members of different groups differently over and above the planned difference in intervention? Is there any reason to suppose that researchers did things to produce results they preferred? Is there any evidence of 'conflict of interest'? Different practitioners may not make judge-	In routine practice, practitioners know what treatments they are giving to clients. They will rightly do all that they think is necessary to produce benign outcomes, including all kinds of things not featured in the research. Hence the results in practice may differ from the results of an experiment where practitioners were blinded and tied to following the research protocols.

devices were used to ensure that their judgements were reliable? (see Chapter 6, section 6)

Go to **20**

20 Specification of intervention (s)

Was what was done to each group of subjects clearly specified?

Go to **21**

21 Standardisation of intervention

Was what was done to/for each subject within an arm of the experiment similar to what was done for each other in the same arm?

Go to **22**

22 Observations and measures

What baseline, interim and outcome measures were made? Was the same kind of data collected for each subject? Is there missing data?

ments in the same way. The effect of this may filter through as spurious outcome differences if different practitioners furnished data for different groups.

Unless what was done to/for subjects in *all* groups is clearly specified it will be impossible for readers to judge the results of the experiment as being the results of different treatments being given to otherwise similar groups, or as the result of the same treatment being given to groups with different characteristics. It will also be impossible to replicate the experiment, or to compare its results with those of others.

Where there is considerable heterogeneity of treatment within an arm of a trial, this is not really one arm but many arms and results should really be analysed in as many categories as there were differences in treatment. This requires a larger sample size (see Chapter 5, section 9 and Chapter 7, section 7).

Unless practitioners know precisely what was done in the research they cannot successfully replicate this in their own practice. And would it be possible for practitioners to deliver the same intervention in practice: issues of resources, expertise and so on?

Most routine practice entails customising interventions to the particularities of each client. The more this is true, the more difficult it is to investigate the effectiveness of such treatments, since investigations of effectiveness require reasonable sized groups of similar people treated much the same as each other.

Would the observations and measures used in the research be feasible for someone in routine practice – if not, it will be difficult for practitioners to know whether they can do, or could achieve the same results in practice.

Were the instruments used appropriate for the research question?

Had the instruments used been validated for use in a similar way?

Did the instruments used generate data in a form appropriate for the statistical analysis conducted?

Insofar as practitioner judgements were involved were they reliable judges? (see **19** above).

If the study was aimed at measuring effectiveness, what criteria of effectiveness were used? Might other definitions of effectiveness have been more appropriate?
Go to **23**.

See 'Questions to Ask about Data Collection Instruments'.

Are the researchers' ideas about effectiveness the same as those of practitioners and/or of their clients?

23 Duration

Was the period of the study long enough to ensure that all important differences and similarities of outcome would show themselves?

Go to **24**.

If the duration of the study is too short then:

- some important 'side-effects' might not show
- short-term, but temporary advantages may be mistaken for long-term and permanent gains
- some long-term benefits of intervention (or non-intervention) might not show.

For what duration of care is the practitioner responsible? Can practitioners track their clients for longer, or shorter than the period of the experiment?

24 Statistical analysis

Were all the subjects accounted for in the analysis?

If a large proportion of subjects are missing from the analysis then it is unlikely that the comparison groups are similar any more (although there are ways of estimating the effects of missing data).

Were the tests used appropriate for the data produced and the questions posed?	This is a question for which reference to a statistics textbook may be necessary, but see Chapter 6, sections 3 and 4; Chapter 7, sections 6, 7 and 8. If tests of association/correlation were used, see Chapter 6, section 6, and Chapter 10, section 11.
Were the differences statistically significant?	See Chapter 7, sections 1 to 4.
Were confidence intervals calculated/was there enough information given for readers to calculate CIs?	See Chapter 7, sections 4, 5 and 9.

Go to **25**

25 Effect sizes

An effect size which is statistically insignificant should be ignored. But a statistically significant effect may still be too small to be important in practical terms (see Chapter 7, section 10 *passim*).	Is any statistically significant difference shown also of practical significance? For example, is it big enough to convince a practitioner that it would be worth the cost and effort of changing practice?

Go to **26**

26 What works for whom?

Does the study describe what happened to sub-groups within each arm of the experiment? For example, those in a control group whose condition improved as much as those in a treatment group, or those in a treatment group whose condition deteriorated more than those in the control group. This is desirable, but sample size which is adequate for showing statistically significant differences between two arms of a trial may well be too small to show statistically significant differences between sub-groups within an arm of a trial. Beware of small trials making claims about sub-groups.	Does the study identify sub-groups affected differently by the intervention / non-intervention in a way that could be used to guide practitioners' decisions as to whom to treat by what means: indications and contraindications, or 'risk factors'?

Go to **27**

27 Conclusions

Were the conclusions answers to the questions asked initially?	Beware of research that asks one question and concludes with an answer to another. The research was probably not designed to answer the second question!	Were conclusions stated in practice-relevant terms? For example, in terms of 'numbers needed to treat' (Chapter 7, section 10.5).
Did the conclusions follow logically from the evidence presented and the analysis conducted?		Did the conclusions drawn about the applicability of the research take account of the realities of practice and the diversity of practice circumstances?
Is the conclusion restricted to what the evidence will support?		

Go to **28**

28 Other research

Did the authors place their research in the context of other research? (see also Questions to Ask about Systematic Reviews and Meta-analyses)	Does the study confirm findings from other studies, or run against them? Better evidence is needed to make a convincing case which runs against the grain.	All other things being equal, similar results from several studies are a safer basis for practice than results from a single study.

Go to **29**

29 Ethical considerations

What evidence is there that the study was conducted in an ethical way: look for the way subjects were screened for entry to the experiment, the gaining of informed consent, measures to prevent harm and ensure confidentiality. Is there reference to approval by an ethics committee?	Unethical research may none the less be internally valid.	Would it be ethical to apply the results of this research in practice?

Go to **30**

30 Was the experiment the basis for a cost-effectiveness analysis? If so, see Questions to Ask about Cost-effectiveness Studies.

QUESTIONS TO ASK ABOUT SYSTEMATIC REVIEWS AND META-ANALYSES

Questions about the relevance of the review	**Comments**
1 Does the review relate *at all* to the questions for which you want an answer?	Read the abstract or the introduction to the review to avoid wasting time on a review which won't answer your questions.
2 Does it provide evidence in relation to the outcomes you are interested in?	Do a quick check as to what outcomes feature in the review. Are they of the kind which you could relate to your own practice?
3 Does it provide evidence that would be relevant to the practices/kinds of practitioner/ kinds of agency in which you are interested?	Is this a topic where local socio-economic conditions, institutional arrangements or practice arrangements are likely to affect outcomes and if so how similar are your circumstances to the circumstances of the studies reviewed?
4 Does it provide evidence relevant to the kinds of clients you are interested in?	Are the clients in the studies reviewed sufficiently like the clients you are interested in?
5 If you are interested in cost-effectiveness, does the review provide the kinds of cost data which you can relate to cost data from your own circumstances?	See Chapter 3 of this volume and 'Questions to Ask about Cost-effectiveness Studies'
Questions about the quality of the review	**Comments**
6 Is this a systematic review emanating from an authoritative source, and itself peer-reviewed?	Most systematic reviews are published in peer reviewed journals or via one of the agencies specialising in systematic reviews and will have been vetted for their quality.
7 What is the date of the review, and of the studies it reviews?	Systematic reviews are always more out of date than the studies they review. Some topics date worse than others.
8 Is there any evidence that the reviewers are disposed towards or against particular kinds of practices, practitioners or client groups?	Some reviews are openly propagandist for particular approaches, others may suffer from the unwitting bias of the reviewers. It is worth looking for evidence of the reviewers' affiliations. Was the review funded by a drug company? Was a review with a conclusion favouring a social work practice written by a social worker committed to this practice?
9 Does the review address a clearly focused issue, or a set of clearly focused issues? Ideally the review should concentrate on studies asking similar questions about similar interventions, measured according to similar outcomes for similar kinds of people.	The possibility for reviewers to focus a review depends on what research is available for review. For example, the review by Roberts et al. in this volume (Chapter 4) deals with the effects of home visiting according to a wide variety of different kinds of injury occurring or not under a wide range of circumstances. That reflected the state of knowledge derived from RCTs at the time but it made for considerable difficulties in drawing general conclusions.
10 How widely and where has the reviewer looked for relevant studies to review?	Many systematic reviews are restricted to a single language, perhaps missing important papers in other languages. However, it is important to consider whether topics are culturally or institutionally bounded (see **3** above). Outcome measures such as 're-admitted to hospital' will not equate between societies where the structure of health services is different. There is a *publication bias* towards publishing studies which reach definite and positive conclusions. Good systematic reviews look for unpublished studies as well.
11 Did the reviewer ask authors to provide information missing from published accounts of their studies? Was this successful?	There are acute publishing constraints on how much information can be published. Sometimes it is not until a systematic review is conducted that it is clear what information is relevant.
12 What criteria did the reviewer use to decide which studies to include in the review. Are they appropriate criteria?	There will be two bases for such critieria. One is that they produce a group of studies where like can be compared with like. The other basis is that the criteria exclude poorly conducted studies. The latter criteria are similar to satisfactory answers to 'Questions to Ask about Experiments'.

13 Did the inclusion criteria exclude some studies which might have been interesting to you?	Many self-styled systematic reviews exclude all non-experimental research, and many all but RCTs. But other kinds of research can be relevant to practice (see Parts 2 and 3 of this volume).
14 Did the reviewer evaluate each study included so that less convincing studies count less towards the overall conclusion than more convincing ones?	Simply treating each study as equally convincing runs the risk of poorly conducted studies crowding out the better ones. The evaluation questions which should be asked by the reviewer will probably be similar to 'Questions to Ask about Experiments'. In Chapter 4 Prendiville's criteria are explained and used to weight studies.
15 Were the reviewer's judgements on the use of criteria evaluated for their reliability?	Asking several judges to apply the same criteria independently is the standard way of testing this: an inter-rater reliability test (see Chapter 6, section 6).
16 If the results of studies are combined in a meta-analysis, was it reasonable to do this?	Meta-analysis may be problematic if the numerical data have been produced using different instruments in different studies and because there is a risk that a large and poorly conducted trial will overwhelm the results of smaller but better ones (see Table 1 in Chapter 4). Data transformations are often necessary (Chapter 6, section 4; Chapter 7, section 6). In addition, meta-analysis presumes a high degree of similarity in all the studies combined in terms of people involved, context of practice, clientele, outcomes measured and so on. (For interpreting a meta-analysis, see Chapter 7, section 5.)
17 Does the author suggest implications for practice? If so do these follow logically from the review?	This includes considering whether the review addressed all possible outcomes, benign and adverse which might arise from adopting a particular intervention. Numbers needed to treat (NNT) can be a useful statistic for practitioners but needs to be regarded with caution (see Chapter 7, section 10.5).
18 Does the review clearly identify gaps in knowledge which would benefit from further research?	One of the main purposes of systematic reviews is to review the state of knowledge and to identify priorities for future research.

QUESTIONS TO ASK ABOUT COST-EFFECTIVENESS STUDIES

Questions	**Issues of internal validity** You judge whether the findings were true for the research location	**Issues of external validity or generalisability** You judge whether the findings would be true/ could be/should be, made true elsewhere
1 Evidence of effects How was this derived and is it valid evidence?	The most convincing evidence about effectiveness comes from experimental research (Chapter 5). See 'Questions to Ask about Experiments'	Is achieving these effects a high priority for you? Note: that the smaller the difference shown in research in effectiveness and/or popularity between two interventions the less likely it is that the same differences will show elsewhere.
2 The range of costs and benefits considered: What costs and benefits, whose costs and whose benefits? Costs and benefits over what period of time? Monetary and non-monetary costs?	No study could hope to include *all* conceivable costs and benefits to everyone, but there should be a clear statement about what costs and benefits are being considered. For internal validity the study should be judged on how well the researcher has accounted for the costs and benefits s/he chose to focus on.	Is this the range of costs and benefits of interest to you? For example, does a study of hospital at home care fail to include social services costs? Is aiming for these benefits incompatible with aiming for something else desirable?
3 Costing	Is the methodology of costing fully described?	Would it be possible to find the same costing

	Are the data on which costing is based: • accurate? • timely? Was costing even-handed as between two alternative interventions?	data locally and substitute them for those in the study? If not, how similar or dissimilar are the cost bases in the research location and the practice location? The more similar, the more applicable the study.
4 Sensitivity	Do the authors conduct sensitivity analyses (SA) to estimate the effect of varying important factors? If so, are the topics chosen for SA the factors which are both most likely to vary and most likely to affect the cost–benefit ratio?	Are there any reasons to assume that the cost basis is very different in the practice location? How out-of-date are the costings? Don't bother about matters that can be updated by a simple inflation factor: consider costs which might have changed radically. Do the authors' sensitivity analyses help to fit their study to your situation? Do they provide enough data to allow you to conduct a sensitivity analysis of your own?

QUESTIONS TO ASK ABOUT SURVEYS, CASE FINDING (OR 'CLINICAL EPIDEMIOLOGICAL') STUDIES AND CASE CONTROL STUDIES

1 What kind of study is this?

- Is it a survey, with a sample taken to represent a wider population? If so start at **2** (count 'cohort studies' as surveys; see Chapter 10, Box 10.6).
- Is it a case finding (or 'clinical epidemiological') study with cases of something being looked for, and their frequency being expressed as a proportion/rate of the population in which they were looked for? If so start at **5** (for an example see the study by King et al., in Chapter 11, sections 1 and 2).
- Is it a case control study, with cases of something being looked for, and then matched with controls selected in a different way? If so start at **4** (for examples, see Chapter 10, section 9).

Questions [Surveys only]	**Appraisal issues**
2 The sample Of what population was the sample supposed to be representative?	Usually this will be a population of individuals, but sometimes it might be a population of agencies, areas, or time periods. Researchers should clearly state the population they are trying to represent since this is the starting point for selecting a sample and the basis for judging whether the sample was adequately representative.
At what level of detail was the sample supposed to be representative?	Samples may be of an adequate size to represent a population at one level of detail, but not of an adequate size to represent the same population at another level of detail (see chapter 10, sections 1, 6 and 7). Check that the researcher does not make claims about matters for which the sample was too small.

Was a sampling of individuals preceded by clustering to group respondents conveniently? See Chapter 10, Section 4	Many national surveys begin by sampling areas, and then sample individuals within them. Clustering may undermine the representativeness of the sample (see Chapter 10, section 4).
Was the sample stratified? See Chapter 10, section 3	Where comparisons between sub-groups of different sizes are of interest separate samples may be chosen from each to avoid the sample for the smaller sub-groups being too small (see Chapter 10, section 3). If this is not done, check to see whether the researchers do not make comments about sub-groups for which the sample is too small (see Chapter 10, section 7).
What sampling frame (if any) was used? Was it a complete listing: were omissions likely to slant the sample away from representativeness?	Sampling frames may under-represent some groups of interest. Do the researchers' conclusions take this into consideration? (see Chapter 10: Sections 2 and 8)
What technique was used to select the sample and was this adequate for the purpose?	Various techniques of sampling are given in Box 10.1 of Chapter 10. Never accept any generalisation about frequencies in a population which are based on convenience samples, or self-recruited samples.
How large was the sample and was it large enough to represent at the level of detail aimed for?	Eyeball the data given in any tables. If *each* cell of the table is filled with a small number this is an indication that the sample was too small (see also Chapter 10, sections 6 and 7).

3 Non-response

What was the level of non-response?	
Did the researchers have accurate data as to the characteristics of the population in order to judge representativeness?	See Chapter 11, section 2.
Were the characteristics of non-respondents checked against known characteristics of the population?	The size of the non-response is less important than the characteristics of non-respondents (see Chapter 10, section 8)
Were estimates made of the extent to which non-response might have skewed the sample away from representativeness?	See Chapter 10, section 8.
Were the results weighted to redress greater non-response by some section of the population? Was the process of weighting adequately explained?	See Chapter 10, section 8 and Chapter 11, section 1.
Now go to **6**	

4 [Case control studies only]

How were the controls recruited, and matched to the cases found? (If there was an attempt to make the controls representative of a particular population, go back to **2** and start there, but consider the questions only in relation to selecting controls – and don't forget the question about matching.)	Faulty matching will lead to confounding (see Chapter 10, section 9)

5 [Case control and case-finding (clinical epidemiological) studies]

How were the cases found and what means were taken to ensure:

• that the cases found were all of the same kind? • that all the cases were found, or some estimate was made of the percentage not found?	Look out for unreliable categorisation (Chapter 6, section 5) and for ascertainment bias (Chapter 10, section 9)

6 Administration

How was the data collection instrument administered? (Postal questionnaire, telephone interview, face-to-face interview, data collected by clinical examination or assessment)	Is there any reason to believe that the way the research was done affected the results, overall or for sub-groups of respondents? Was it administered similarly for all respondents? Would the same situation have had different effects for different respondents?
What was the context in which the survey instrument was administered?	Is there any reason to believe that the context in which the survey was administered or the characteristics of the people administering the survey would have differential effects on different kinds of respondents: for example white interviewers with black respondents?
By whom was the survey instrument administered?	
Were different interviewers trained to give a standardised performance/ follow a common protocol?	Were the results of the survey analysed interviewer by interviewer to investigate interviewer effects? For all above see Chapter 10, section 14.

In research that involved diagnosis or assessment by an expert were these judgements reached following a common protocol?	If expert judgement was involved was inter-rater reliability established? (see Chapter 6, section 6)

Instead of the questions in cells **7** to **11** you might prefer to use the questions in cell **22** of Questions to Ask about Experiments

7 Validation and piloting	
Had the data collection instrument and procedures been piloted and amended in the light of this?	The re-use of tried and tested instruments has the dual advantage that the main problems will have been ironed out, and that the results of several pieces of research using the same instrument can be directly compared with each other (see Chapter 6, section 1).
8 Questionnaire or other instrument (for example, a self-completed diary or a diagnostic algorithm). Does the author give sufficient detail about the research instrument used?	Research that is reported without disclosing what instrument was used is virtually impossible to appraise; but for reasons of space details may be published elsewhere.
9 Questions (for each question)	
Did the question ask people things they could reasonably be expected to know?	People are often willing to give answers to questions even if they don't know the answers, or to invent opinions on matters they have never considered before.
Was the question about something which happened some time in the past?	Answers to retrospective questions are often inaccurate about what really happened, though they may be accurate about what someone thinks happened.
Was the question asked in a concrete way: for example, 'How many times have you visited the doctor in the last seven days?' and not 'How often do you usually visit the doctor?'	People vary greatly in the way they generalise. One person's 'usually' may be another's 'rarely'.
Was the question 'leading' in any way – suggestive that one answer might be favoured by the interviewer rather than another?	What is a leading question for one person, may not be a leading question for another.
Was the question offensive, over-intrusive or upsetting in some way?	People vary in what they find offensive or intrusive.

Was the question phrased in a straightforward and unambiguous way, for the people intended to answer it?	Including whether questions were asked in a language some people didn't understand, or people with limited reading skills were asked to read a questionnaire.
Did any question have a particularly low response rate?	This is usually an indication that this was not an effective question and the results should be interpreted accordingly.
10 Responses	
Were responses forced?	For forced choice questions were all the possible responses provided? For rating scales was there a 'neutral' middle position, or were respondents forced to opt for a positive or a negative answer?
Were responses open?	After the survey, responses to open-ended questions have to be coded into categories. How was this done?
Were respondents allowed to make more than one response to each question?	This leads to great difficulty in analysis, since loquacious respondents contribute more to the survey results than the reticent.
11 Analysis	
Were the results presented with confidence intervals/sampling errors?	See Chapter 10, sections 6 and 7
Were differences tested for statistical significance?	See Chapter 7, sections 1 and 2
Were correlations expressed as correlation co-efficients and were these tested for statistical significance?	See Chapter 10, section 10
Were the conclusions based on figures that were large enough to bear them?	This can be a difficult question to answer, but see Chapter 10, sections 6 and 7.
Were the statistical tests used appropriate for the kind of data collected?	See Chapter 7, section 6
Was the commentary on the statistics commensurate with the statistical analysis of the data?	For example, does the author conveniently forget the non-respondents? Does the author make much of differences/correlations which are not statistically significant?
If frequencies were expressed as population rates, did the researcher have accurate data about the population?	Under-estimating the number of people 'at risk' will over-estimate the rate (see Chapter 11, section 2)

Did analysis involve age standardisation?	See Chapter 11, sections 1 and 2
Did analysis involve the use of deprivation indices?	See Chapter 11, section 4
How successful was the analysis in terms of creating categories for comparison similar in all respects except for the variable of interest?	See Chapter 10, section 11 (see also Chapter 5, sections 1 and 3)
12 Presentation of results	
Did displays of data include all the data relevant to the conclusions drawn (for example 'don't knows' as well as those who gave answers)?	
If the survey involved non-proportional sampling (Box 10.1 in Chapter 10) is it made clear whether the results presented have been re-weighted to proportionality?	See Chapter 10, section 8 and Chapter 11, section 1
If results are presented in terms of a reference or standard population, is the standardisation fully explained?	See Chapter 11, section 1
13 Only if authors draw causal conclusions	
Was this a contemporaneous/snap shot study or a longitudinal/prospective/cohort study?	Causal statements based on correlations found in contemporaneous surveys may well pose direction of effect problems which cannot be resolved from the survey or case control data (see Chapter 10, sections 11 and 12).
14 Other research	
Does the author place conclusions in the context of wider literature? Does this strengthen or weaken the conclusions?	Usually more and better evidence is required to support findings that run counter to the weight of published research.
15 Conclusions	
Do the authors' conclusions follow logically from the analysis of the evidence presented?	The authors' conclusions may simply be a generalisation from the sample to the population from which it was drawn. But an author may draw more speculative conclusions about other populations, about causal linkages (**13** above) or about the success or failure of policies and programmes. Generalisations about frequencies in populations should not be drawn from case control studies (see Chapter 10, section 9).
16 Practice usefulness	
Are the results presented in ways that allow for extrapolation of the results to a practice area?	A local practice area is most unlikely to show the same patterns as a national survey or a local survey elsewhere. Age standardisation, the use of deprivation index scores or presentation of results in terms of a reference population make it easier to extrapolate from a survey to practice (see Chapter 11)
Insofar as you find the conclusions convincing, is there any practical use to which you might put them?	For example: • Providing a model for planning local surveys. • Providing a tested questionnaire which you might re-use (see Chapter 6). • Evaluating the impact/effectiveness of a policy/programme elsewhere in deciding whether to implement it locally. • Bench-marking local performance on performance elsewhere (see Chapter 11). • Identifying targets and 'at-risk' groups for intervention/prioritisation.

QUESTIONS TO ASK ABOUT QUALITATIVE RESEARCH

Questions	Appraisal issues
1 Is there a clear statement as to the aims of the research?	Qualitative research may not have a 'Research Question' as such, except 'What's going on around here and why?' This is a legitimate research question.
2 What kind of phenomena does the researcher think s/he is studying?	It makes a great deal of difference as to whether, for example, researchers think they are studying utterances as evidence of how people experience things/what they mean to them (see Chapters 12 and 13), or utterances as ways people do things with language at the time they speak (see Chapter 14). Assumptions about the phenomena being studied determine the appropriate way of collecting data and analysing them (see Chapter 16, sections 5 and 6, though the examples there by no means exhaust all the possibilities).
3 Does the author announce or imply a value position?	Qualitative research is sometimes 'value-led' with researchers wanting to draw attention to injustices, or to 'give a voice' to voiceless people. Given the great leeway there is for qualitative researchers to shape their data, out of sight of their readers, readers should beware of findings that exactly support the value position of the researcher. None of this is to say that allegedly 'value-free' research does not suffer from researcher bias (see Chapter 16, section 7).
4 Was this collaborative/participatory research (the researcher teaming with the subjects of the study to produce research together)?	What kinds of commitments might there have been between researcher and co-participants which might have shaped the way the findings were made public? (see Chapter 16, section 7)

5 Was this research an observation of events happening under natural circumstances, for example, a participant observation study of a clinic? If yes, go to **6**
Or
Was this research an interview/focus group study? If yes, go to **11**

6 The research location Was the reason for the choice of research location explained? Was the location suitable for the research undertaken? How typical is the research location?	In principle there is no reason why qualitative research should not be conducted anywhere at any time, in relation to any activities, or any people. But the questions opposite relate to the generalisability of the findings and the researcher's aspirations about making generalisations. For example, if generalisations are made from a study of a particular clinic, it is important that the clinic is representative of others in some ways, though not necessarily in the statistical sense of the term 'representative': note the difference between statistical and theoretical sampling in Chapter 16, section 4.
7 Time sampling Does the author state how long was spent in the research location(s)? Was this long enough? For example, are claims made about the outcomes of what was observed which happened after the period of observation? Which time periods were sampled, and which were not?	General claims about what happens in the research location covering times not observed should be treated with suspicion, as should general claims about activities when those observed may be atypical, and similarly general claims about types of people when only some of the type present in the research location have been observed.
8 Activity/event sampling Were there activities/events important to the topic of the research which were not observed? Were the activities/events observed typical of this class of activities? Were observations of atypical activities/events used to illuminate typical ones? (deviant case analysis: Chapter 16, section 4)	
9 Personnel sampling Who were the people observed? Were those observed a representative selection of the people of interest in the research location? Representative might mean representative in the statistical sense, or representing important theoretical categories (see 'theoretical sampling' in Chapter 16, section 4).	

Were the activities of idiosyncratic people used to illuminate those of people behaving more usually in this context? (deviant case analysis: Chapter 16, section 4)

10 Researcher's role and identity in the setting

How did other people understand what the researcher was doing?
How might this have affected what was said or done by them in the researcher's presence?
How might the role adopted have disbarred the researcher from certain settings, affected the way s/he was communicated with and so on?
Now go to **13**

The danger, of course, is that what the researcher sees is what happens when researchers are around, rather than what happens when they are not. In addition other people are often most anxious to make researchers see things in particular ways and not in others.

11 Sampling

How were the interviewees/focus group members selected?
Was this an appropriate way of selecting respondents for the topic of the research?
How many in total were there?

The importance of answers here depends on the researcher's aspirations about making generalisations. If generalisations are made on the basis of those who were respondents, applying to people who were not, then respondents need to be representative of the latter in some way. Representativeness does not have to mean statistically representative. It can mean representative of a range of theoretically interesting categories irrespective of their commonality or rarity. See the distinction between statistical and theoretical sampling in Chapter 16, section 4.

12 The interview/focus group situation

How were respondents briefed about the purpose of the research?
Does the author provide information as to what happened during the interviews/focus group which might have shaped the data produced?
Does the author provide information about the communicative behaviour of the researcher as well as about that of the respondents?

The ideal information comes from full transcriptions of interchanges between researchers and researched, but it is rarely practicable for authors to provide all that is necessary. None the less, there should be some data provided about the context in which the data were elicited so that readers can judge whether the 'findings' are an artefact of the context in which the data were gathered and/or of the communicative style of the interviewer.

Where interview respondents reported on what they did or said elsewhere, was there any means of verifying this, and were these means used?
(You might also like to look at cell **6** in 'Questions to Ask about Surveys')

There may be little relationship between what people themselves say they do, and what they can be observed to do.

13 Data recording

How (and when) did the researcher record the data? How did the researcher decide what to record?

The act of recording data can be intrusive and disruptive in observation research, but relying on memory and writing it up afterwards may suffer from selective remembering. Whether in observation or in interviews, handwritten notes about what people said are much less convincing than transcripts of tape/video recordings. Where records are made by audio or video recording they are already structured by decisions as to where to site the equipment and when to switch it on, but these are matters which are easy to know about. By contrast, where a researcher writes field notes in retrospect these will be shaped by implicit and difficult-to-know-about analysis in the form of selective attention, selective forgetting, and editing in terms of ideas of importance and relevance. Strategies such as a grounded theory approach (Chapter 16, section 4) make more of this explicit and accountable.

14 The analysis of the data

Does the author explain and *demonstrate* how the data analysis was done: for example, is a specimen transcript provided with examples of codings? Are decision-rules for coding given?

Was the coding of the data checked for inter-rater reliability? (see Chapter 6, section 6)

Are frequency statements made about the commonality of particular responses/behaviours?

Qualitative researchers suffer from the weightiness of their data and it is not always possible for them to provide readers with all that is needed to make an informed judgement about the adequacy of the analysis. However, it is not reasonable to expect readers to take matters entirely on trust.

What data were excluded is an important question. It is all too easy to carve a story out of a large amount of

Is an explanation given for what data were included and what excluded from the analysis?	data simply by ignoring most of it. The question of frequency relates to this. 'Most' only has a meaning in relation to what is counted as 'all'.
15 Fallibility testing Did the researcher submit the findings to the people researched for their opinion about their accuracy?	Fallibility testing is popular with some researchers, but it is certainly not an acid test of the accuracy of findings. People may accept findings because they like them, reject them because they don't and findings which are acceptable to some people may be objectionable to others.
16 Triangulation Were data produced by various methods compared (for example from observations and interviews, or from interviews and clinical measurements)? Did all the data support the same conclusions?	Conclusions supported by data collected by more than one method are more convincing than those based on a single data-collecting strategy. However, sometimes triangulation by method will lead to incompatible conclusions.

17 Overall, do the data support the conclusions the author derives from them?

18 Other research How does this research relate to other research in the same field?	All other things being equal, it requires more convincing evidence to undermine existing findings.
19 Generalisability Is it possible to use/does the author use data from elsewhere to locate this study in a wider context?	For example, for Blaxter (Chapter 12) the Health and Lifestyle Survey findings provided a kind of map in terms of which those she interviewed could be located. They may not have been statistically representative, but at least it was clear where they fitted in the wider picture.
Does the author generate any useful analytic concepts?	Qualitative research has generated many useful general purpose concepts, such as stigma, career, total institution, deviancy amplification and so on.
Is there anything you have learned from this study which helps you to understand other people, or social or psychological or organisational processes better?	It is sometimes claimed that the main function of qualitative research is to produce vicarious experience, providing insight into the worlds of other people (see Chapter 16, section 8).

QUESTIONS TO ASK ABOUT ACTION RESEARCH

1 Was the research design 'experimental' in the sense that there was some attempt to compare the effects of an intervention with the effects of no intervention, or of some different intervention; was there some degree of control built into the research? (see Chapter 5, section 3)
If yes, use 'Questions to Ask about Experiments' and Chapters 5, 6 and 7
If no, go to **2**

2 What strategies were used to produce evidence?
If this involved telling the inside story, see also 'Questions to Ask about Qualitative Research', cells **6** to **10** and **13** to **16**, and Chapter 16
If this involved loosely structured interviews, see also 'Questions to Ask about Qualitative Research', cells **11** to **16**, and Chapter 16
If this involved questionnaire research, see also 'Questions to Ask about Surveys', cell **6**, and Chapter 10
If this involved other kinds of data collecting instruments, see also 'Questions to Ask about Data Collection Instruments' and Chapter 6

Questions	Issues
3 Does the study provide convincing evidence that the 'action' actually led to the effects claimed for it?	For example, if it is claimed that the research was successful in improving the knowledge and competence of those involved, what measures were there of this prior to and at the end of the research? Or if it is claimed that at the end of the research there was a more cohesive community – what measures were used to indicate more and less cohesiveness? Or if the claims are about improving the quality of service, what criteria were used to judge service quality and how were these turned into measures of quality?
How was the situation prior to the research investigated and characterised?	
How was the situation at the end of the research investigated and characterised?	If testimonial evidence is used, is it reasonable to assume that those who bear witness are:

	• Reliable witnesses – know what they are talking about? • Have not been specially selected by the author to give a clean bill of health to the project? • Are representative of all those affected by the action research?
Were effects for all people involved reported?	Are real or possible adverse effects reported as well as 'success' stories?
4 Replicability Is sufficient evidence provided for readers to be able to carry out the same intervention elsewhere? This includes information about: • What was done. • The characteristics of people who benefited, and of those who did not. • About the practitioners involved; their skills and other relevant characteristics. • About the resourcing of the project. • About the organisational structure of practice and/or political context.	Most action research features 'complex' interventions (Chapter 5, sections 9 and 12), involving many activities, often customised to individuals and circumstances and carried out over a longish time period. They may be particularly difficult to specify, thus leading to difficulties of knowing which aspects of the intervention had which effects, if any at all, and what someone would have to do to produce the same results elsewhere. Better studies attempt to specify the conditions necessary for the same actions to have the same effects elsewhere.
5 'Experiment effects' (see Chapter 5, section 5) Is it clear how much of what happened was because of the actions taken, and how much of what happened was due to the commitment and enthusiasm of those involved?	Effects produced by mobilising the enthusiasm of participants may be real enough, but may not be replicable by someone else adopting the same strategy elsewhere.
6 Biased reporting How likely is it that the research report was written to give a favourable picture of the practitioners and clients involved, an agency, or a favoured way of working?	Publication bias ensures that most published action research consists of 'success' stories, and most is written by researchers with strong value commitments, or commitments to particular forms of practice. Much is collaborative or participatory, such that the final report will reflect a consensus among those involved. Considerations of image and reputation may influence what is published, and some action research is specifically designed to improve the public image of the participants.
7 What insights and understandings does the study provide which might improve the way another practitioner practices?	The answer to this depends partly on the answers to the earlier questions but this is also a question about 'naturalistic generalisation' as discussed in Chapter 16, section 8.