Chapter 5

How to appraise the studies: An introduction to assessing study quality

5.1 INTRODUCTION

If synthesizing evidence is like assembling a jigsaw, then most of the preceding chapters have been about finding the pieces – in this case, the individual studies. In the next stage, each individual piece is examined critically to determine whether it really is part of the picture, and where to place it (if anywhere) – and to find out how it fits with other pieces (if at all). In the context of a systematic review, this involves determining whether any of the individual studies are affected by significant bias, as this may affect the weight to place on them when it comes to putting the whole picture together. This involves assessing whether the study is representative of the wider population, whether the numbers add up (for a quantitative study), and whether the study was affected by problems or other events that might affect your interpretation of its results.

The process of assessing the methods and results of each study is often referred to as critical appraisal, and sometimes as “assessing study quality.” In a systematic review, this exercise aims to determine whether the study is adequate for answering the question. Research evidence can be produced by a wide range of methods and approaches, and some of it is produced by studies, which for reasons of cost, practicality, or accident will be subject to some degree of systematic error (bias). For example, in the case of trials, some participants enrolled at the start of a study later refuse to take part, and their outcomes may remain unknown (though research tells us that their outcomes are likely to be worse than among those who stayed in the study). Similarly in surveys, a large percentage of people will not respond. In each
of these cases the non-responders will often differ systematically from the responders. Some of the difficulty with interpreting social (and other) research is due to such biases, but it can also be due to non-reporting of essential information by the study authors. This is the case not only in large-scale quantitative research, but also in studies aimed at collecting qualitative data, where too little information on the context, methods, analysis, and background to the study may be presented by authors to allow the data to be interpreted meaningfully.

**The effect of bias**

It is clear that particular methodological biases may affect the results of the study. They have been shown in some cases to be a common source of inflated estimates of effect size, for instance. This finding is the basis of Rossi’s Iron Law of Evaluation,¹ which implies that simply including all studies in a systematic review without taking this into account will lead to a biased review. This is a strong argument against “vote counting” reviews – those which count the number of studies with positive findings, and those with negative findings (see chapter 6).² If biased studies do indeed overestimate effects, then this increases the risk that the review will falsely reject the null hypothesis. The review will then conclude that an intervention is effective, when it is not.

The list of potential research biases in quantitative studies is long. Sackett identified over 30 biases to which case-control studies alone were susceptible.³ It has been suggested that there are three important biases to which randomized controlled trials are particularly subject: lack of blinding, attrition bias, and inadequate randomization.⁴ (See: http://www.bmjpg.com/rct/chapter4.html for a readable summary of the issue of bias in RCTs.) Although other criteria should be considered when appraising an RCT (for example, the extent to which the study is representative of the users of the intervention), these are the three criteria that have been explored most extensively (see section 5.12). As described later in the chapter, critical appraisal questions for qualitative studies have also been suggested.⁵

### 5.2 NON-REPORTING BY AUTHORS: WHEN NO NEWS ISN’T ALWAYS GOOD NEWS

Critically appraising a study is often made difficult because too little information is presented. In such situations it is common practice for reviewers to contact the original authors for additional information, and occasionally to
request access to the original data (for example, in order to perform
meta-analysis on the individual-level data).\textsuperscript{6,7} Contacting the author for
verification of study details avoids confusing inadequate reporting with a
poor-quality study.

Huwiler-Müntener et al. examined this issue in a review of 60 RCTs, in
which they analyzed the association between a quantitative measure of
reporting quality,\textsuperscript{8,9} and various indicators of methodological quality.
They concluded that reporting quality \textit{is} associated with methodological
quality, but that similar quality of reporting may hide important differences
in methodological quality, and that even well-conducted trials may be
reported badly. This led them to suggest that a distinction be made between
the quality of reporting and the methodological quality of trials.\textsuperscript{10}

5.3 WHAT IS STUDY QUALITY?

Study quality means different things to different people working in different
disciplines. For those doing systematic reviews, assessing study quality is often
used as a shorthand to mean “internal validity” – that is, the extent to which a
study is free from the main methodological biases (such as \textit{selection bias},
response bias, attrition bias, and observer bias). Jadad,\textsuperscript{11} for example, suggests
that the following items may be relevant to assessing the quality of trials:

- the relevance of the research question;
- the \textbf{internal validity} of the trial (the degree to which the trial design,
conduct, analysis, and presentation have minimized or avoided biased
comparisons of the interventions under evaluation);
- the \textbf{external validity} (the precision and extent to which it is possible
to generalize the results of the trial to other settings);
- the appropriateness of data analysis and presentation; and
- the ethical implications of the intervention they evaluate.

By comparison, a recent report on types and quality of knowledge in social
care includes a wider set of sources of knowledge, and criteria on which a
study may be judged. Here “quality assessment” is used in a wider sense to
include issues of transparency (clarity about how the knowledge was gen-
erated), accuracy, purposivity (the extent to which the methods used were
appropriate or “fit for purpose”), utility (“fit for use”), and propriety (which
includes legal and ethical considerations).\textsuperscript{12}

Quality assessment is often used in a more restricted fashion than this in
systematic reviews of quantitative studies, focusing primarily on identifying
methodological problems. This information is used to “weight” each study summarized in the review, and may help with making appropriate methodological recommendations regarding future research. This weighting is usually done narratively, by differentiating clearly between higher and lower quality studies. The impact of study quality can also be investigated by means of a sensitivity analysis, which involves testing how sensitive the review findings are to the inclusion and exclusion of studies of different quality.

5.4 CRITICAL APPRAISAL: TOO MUCH CRITICISM, NOT ENOUGH APPRAISAL?

Critical appraisal aims to ensure that the reader directs her attention to all the key aspects of the study – its design, methods, participants, setting, and any key measures or variables. This is rather different from simply reading a study closely to see if it is “good enough.” Instead, it often involves using a checklist or scale to formalize the process of appraising the study. This ensures that the main methodological issues are examined systematically, using the same approach for each study, and this makes it less likely that problems or biases will be overlooked. Of course, a reader often implicitly assesses the quality of any paper they read as a matter of course; however this may not be done in an unbiased or transparent manner.

While critical appraisal is essential to sound systematic reviewing, being “critical” is not an end in itself. No study is perfect, and it is easy to find flaws in every piece of research. Steven Woolf, of the US Preventive Services Task Force, which has done much to promote the judicious use of critical appraisal techniques, warns against taking critical appraisal to extremes:

critical appraisal can do harm if valid evidence is rejected. . . . At one extreme of the spectrum, where data are accepted on face value (no appraisal), the risk of a type I error (accepting evidence of efficacy when the intervention does not work or causes harm) is high, and that of a type II error ( discarding evidence when the intervention actually works) is low. At the other extreme (excessive scrutiny) the risk of a type II error is great.13

The skill in critical appraisal lies not in identifying problems, but in identifying errors that are large enough to affect how the result of the study should be interpreted. For reviewers assessing social science research studies, care should be taken when using “off-the-shelf” critical appraisal tools, because they may have originally been designed for use with clinical interventions, and may encourage inappropriate criticism of evaluations of social interventions,
where, for example, double-blinding will often be impossible. Similarly, what are considered “appropriate” methods or standards varies between disciplines, so intelligent adjustment of checklists for use in other disciplines is sometimes needed – though of course this may affect the validity of the scale.

Assessing the susceptibility to bias of each study is the key to critical appraisal. Study methods do significantly affect study outcomes. For example, Heckman et al. have shown how drop-outs from a trial (people who drop out of the intervention group, and do not participate further in the study), can result in inaccurate findings in a trial of a classroom training program. Contamination (where participants in a trial receive an intervention which they were not intended by the experimenters to receive, perhaps because they sought it out for themselves from another source) can have the same effect. Bias due to drop-outs is not confined to trials of course; participants who drop out of studies are likely to differ significantly from those who remain in the study, and so the study findings may be based on a biased sample. In one US-based longitudinal study of smoking, the drop-outs tended to be of lower academic achievement, have lower knowledge about tobacco and health, and were more likely to be smokers. Educational status and ethnicity have also been shown to be related to attrition in other research studies.

Just because a critical appraisal checklist or tool has been widely used in other systematic reviews, this does not mean that it produces the “right” answer about a study’s validity. One should be wary in particular of adding up the individual items in a checklist to get a summary score. Jüni et al. for example warn that very different results may obtained by scoring the same studies with different checklists. They used 25 different scales to score 17 RCTs and examined the relationship between summary scores and pooled treatment effects in a series of meta-analyses of medical RCTs. The results showed that, depending on the scale used, very different (even opposite) results could be obtained. They concluded that it is more important to examine the methodological aspects of studies individually, rather than in summary form.

**5.5 THE HIERARCHY OF EVIDENCE AGAIN**

Many reviews have used a “hierarchy of evidence” to appraise studies. However, this is not the same as critical appraisal. The hierarchy of evidence ranks studies according to the degree to which they are affected by bias, but it is more useful in reviews as an aid to determining study appropriateness,
than for exploring issues of study “quality.” Moreover, an assessment of study quality alone is unlikely to be informative; one at least needs also to know whether the study results are likely to be generalizable. For example, one can imagine a situation in which a systematic review includes an RCT and an observational study both evaluating the same intervention. Although the RCT may have higher internal validity (that is, it will be affected less by selection and possibly other biases), it may include a highly selected sample of participants. It is known that women, the elderly, and people from ethnic minority groups are more likely to be excluded from trials. Moreover, those participating in trials of treatment tend to be poorer, less educated, and more severely ill than those who do not, while the opposite appears to be the case for trials of prevention (such as health promotion interventions). The types of intervention evaluated in RCTs and observational studies are likely to differ. RCTs are more common in evaluations of interventions delivered to individuals, while interventions delivered at the population level have more often been evaluated using non-randomized study designs. Appraising a study involves making a series of judgments (and trade-offs) about the population, intervention, and methodological biases together.

5.6 IN SEARCH OF THE METHODOLOGICAL “HOLY GRAIL”

The apparent obsession of systematic reviewers with methodological adequacy and bias can seem obsessive and hypercritical, even to other social scientists. One argument goes that most research is good enough for its purpose, and that scientific nit picking contributes little to the understanding of social phenomena. A second argument suggests that formalizing the process of critical appraisal is unnecessary, as any intelligent reader is likely to critically examine the studies that he or she reviews.

However, a century or more of psychological research on human information processing suggests otherwise. We often cannot tell when the numbers do not add up, we recognize patterns in the data that are not there, we miss those that are there, and we use rules of thumb (heuristics) which often work but sometimes don’t, to help us manage large amounts of information (such as those in a review). It is well known from the work of Kahneman, Tversky and others that we select, evaluate, and remember information in a way that supports our individual preferences, we fail to look for evidence that disconfirms our pet hypotheses, and we cannot spot errors in our own reasoning: for example, we examine evidence that contradicts our own views more critically than when it supports them.
Critical appraisal is not carried out in pursuit of some “holy grail” of perfection. If reviewers were only interested in perfect studies they would have nothing to review. Moreover, like physical perfection, our views of methodological perfection are subject to change over time. Critical appraisal, therefore, involves assessing the degree to which a study is affected by bias and whether the degree of bias is large enough to render the study unusable. It is not simply a matter of “sorting the wheat from the chaff,” but an assessment of whether the research is “fit for purpose.” We would probably be both surprised and pleased if we found that a study was methodologically perfect, but we would be more interested in whether it is evidentially adequate.

5.7 CRITICAL APPRAISAL: TOOLS, SCALES AND CHECKLISTS

Wortman describes quality assessment tools as deriving from two main frameworks.25 The first derives from work in the social sciences by Campbell and colleagues and involves identifying the main threats to validity in mainly non-randomized (quasi-experimental) studies.25, 26 This approach led to the development of a 33-item checklist which can be used as a guide to the overall quality of the study, by assessing whether the study is affected by any of Cook and Campbell’s threats to validity.27 The second approach, used initially for assessing the quality of RCTs, was developed by Thomas Chalmers and colleagues and focuses on the objective methodological characteristics of the study.28 The reviewer codes each of these in turn, using this to generate an overall measure of study quality. This widely used tool also allows the coding of many study details in addition to methodological items.28, 29

Many of these “tools” take the form of checklists, some of which have been designed along strict psychometric principles, so that their reliability and validity is known. Others have not gone through this process, but are based on lists of the most common or well-documented biases to which certain types of study are susceptible. The simplest checklists highlight the main sources of bias for each type of study, and the reader uses this information to guide their overall assessment of the study’s soundness. This information is then used in one of several ways:

- To “weight” the studies qualitatively, when summarizing the results; for example, on the basis of the results of the critical appraisal, studies may be categorized as high, intermediate, or low quality, and this
may be taken into account when deriving conclusions from the review.

- To weight the studies quantitatively; a score may be derived for each study, and this may be used to “weight” the study in a meta-analysis – with low-scoring, more biased studies contributing less to the final summary effect size estimate. This is not often done as weights derived in this fashion are not usually empirically based.30

Examples of checklists appear below. Some checklists have been extensively validated (for example, Downs and Black),31 while others have been developed solely for one-off use in a particular review.

**Randomized controlled trials**

One of the most widely used scales for assessing RCTs with respect to these criteria is the Jadad Scale4,11 (see Box 5.1).

Jadad suggests that the total score obtained using the scale can be used to help decide which trials to include in the review, as well as being an overall guide to the methodological soundness of the included studies. The validity of the scale has been demonstrated with the finding that studies that score low on the scale produce treatment effects that are 35 percent larger than those produced by trials with 3 or more points – another demonstration of the Iron Law of Evaluation.1

For behavioral or population-level interventions, it may be impossible to achieve a maximum score on the Jadad scale – for example, one cannot imagine double-blinding recipients of a juvenile delinquency intervention. Participants and providers alike would be pretty sure to notice that they were in a prison or in a boot camp, rather than on a safari trip (though blinded outcome assessment is possible). However the principles of assessing

---

**Box 5.1 Assessing the validity of RCTs using the Jadad Scale**4,11

1. Was the study described as randomized?
2. Was the study described as double-blind?
3. Was there a description of withdrawals and drop outs?
and dealing with dropouts (attrition, and intention to treat analysis) and quality of randomization remain relevant.

Kristjansson et al. have used the Jadad scale in a Campbell Collaboration review of school feeding programs for reducing inequalities in health.32 The scale was supplemented with four other items (relating to allocation concealment, baseline measurement, reliable primary outcome measures, and protection against contamination), and as the review included other study designs, a range of other checklists from the EPOC group (see Box 5.5) was used to assess methodological quality.

Using the Jadad scale does not preclude the assessment of other quality issues relating to trials. One could use a further set of questions to assess the extent to which the results are transferable, and other issues could be assessed. A framework for assessing the ethics of trials in systematic reviews has also been proposed.33

**Assessing the quality of cluster RCTs**

Cluster RCTs, in which allocation to intervention or control arms takes place at the cluster level (such as clusters of individuals, or organizations such as schools, workplaces, or prisons, or geographical areas) are subject to particular biases.34 Puffer et al. have produced a summary of the main sources of biases in cluster RCTs.35 They point to the need to assess bias at the cluster level, and at the individual level. Allocation to clusters, as with individuals in RCTs, should be done in such a way that it cannot be subverted, and clusters should be retained in the study to avoid attrition bias at a cluster level. Identifying and obtaining consent from participants before randomization should also reduce the risk of selection biases, which are introduced if individuals are recruited differentially into the intervention and control arms of the trial, and inclusion/exclusion biases, in which participants may be included or excluded from the different arms of the trial at differential rates after randomization has taken place (see Box 5.2).

**Observational studies (including prospective controlled studies)**

The phrase “observational studies” covers a wide range of study designs and purposes. It includes controlled and uncontrolled studies used to evaluate the effects of an intervention, such as studies with a control or comparison
group, but where no random allocation has taken place. It also includes prospective studies where a population is followed over time to determine the effects of an intervention (sometimes called before-and-after studies; or, if a control group is involved, controlled before-and-after, or CBA studies). Studies used to assess etiology also fall into this category, such as (prospective) cohort studies, and case-control studies (retrospective). There are many other variants of these basic study types (such as retroprospective studies, nested case-control studies, crossover trials, and limited-prospective, or quasi-prospective studies). Interrupted Time Series studies are a special case of observational study and are discussed separately below.

There is no shortage of tools for assessing the validity of non-randomized studies. Deeks et al. found 194 of them in their systematic review, of which six appeared to be of use to systematic reviewers, on the grounds that they allow an overall assessment of quality to be made and compared between studies, and they covered all major threats to validity (see Table 5.1). Deeks and Cooper concur that tools which follow a mixed-criteria approach may be most useful for systematic reviews; this approach involves assessing objective aspects of the study methods, followed by a judgment as to quality. For example, one might record how confounding was dealt with, followed by a judgment as to whether this was adequate. The other important issue to take into consideration when choosing a tool for critical appraisal is useability. Long, complex questionnaires can be slow to use and difficult to interpret, and while this is not a problem in reviews with a small number of primary studies, it is unwieldy in much larger reviews. A general rule of thumb is to try to avoid unvalidated tools, and tools that have not

### Box 5.2 Assessing sources of bias in cluster RCTs

1. Did cluster allocation seem secure?
2. Cluster allocation stratified?
3. Evidence of cluster imbalance?
4. How many clusters lost after randomization?
5. Patients identified before randomization?
6. Could selection have been biased?
7. Evidence of risk of bias?

*(Based on Puffer et al., 2003)*
previously been used (and useful) in similar reviews to your own, though this is not always possible. The use of untested tools may be difficult to defend on scientific grounds when it comes to publishing the completed review in a journal. Box 5.3 gives details of the quality criteria for critical appraisal of observational studies, adapted from the CRD handbook.

The Maryland Scientific Methods Scale (SMS) has been applied to studies in Campbell systematic reviews of crime reduction strategies.\textsuperscript{43, 44} The SMS uses a five-point scale which describes five study designs, ranked in order of internal validity, somewhat similar to a “hierarchy of evidence.” It has been employed along with assessments of statistical and other aspects of internal validity such as response rate, and attrition, and validity of outcome assessment, in a major set of systematic reviews of crime prevention interventions. These reviews used the Maryland SMS to help categorize the interventions into one of four categories: “What works,” “What does not work,” “What is promising,” and “What is unknown.” For example, in the case of school-based crime prevention, strategies that focus on changing the environment were more effective than those focusing only on changing an individual’s attitudes, behaviors, or beliefs (except for effects on truancy and dropout).\textsuperscript{45}

\begin{table}
\centering
\caption{Six tools suitable for use in systematic reviews of quantitative studies}
\begin{tabular}{ll}
\hline
Author(s) & Purpose \\
\hline
Cowley\textsuperscript{37} & 13 items used for assessing comparative studies \\
Downs and Black\textsuperscript{31} & 27 questions for the assessment of randomized and non-randomized studies \\
Newcastle-Ottawa tool (Wells et al.)\textsuperscript{38} & 8 items relating to cohort studies examining causation; (further changes needed to apply it to cohort studies assessing effectiveness). See: <http://www.ohri.ca/programs/clinical_epidemiology/nosgen.doc> \\
Thomas: Quality assessment tool for quantitative studies\textsuperscript{39} & Any study design (randomized or non-randomized) 21 items; items on non-randomized allocation methods \\
Reisch et al.\textsuperscript{40} & 57 items, applicable to any study design \\
Zaza et al.\textsuperscript{41} Tool used for the US Community Preventive Services Reviews & Any study design; 22 items; may require detailed understanding of validity issues \\
\hline
\end{tabular}
\end{table}

\textit{Source:} Deeks et al. (2003)\textsuperscript{29}
Box 5.3 Quality criteria for critical appraisal of observational studies

1. Are the study participants adequately described? For example, look for adequate descriptive data on age, sex, baseline health status, and other relevant variables.

2. If there is a comparison or control group, are they similar to the intervention group, in terms of variables that may affect the outcome of the intervention (including demographic and other socio-demographic characteristics). This may be achieved by matching or other means – it may be taken into account in the statistical analysis – for example, by means of ANCOVA or regression techniques.

3. If the study involves an assessment of an intervention, is the intervention clearly described, with details of who exactly received it?

4. If the study is an etiological study (e.g., does maternal stress cause behavior problems in children?) were the independent and dependent variables adequately measured (that is, was the measurement likely to be valid and reliable)? This may include valid reliable measures, such as well-validated questionnaires if appropriate.

5. Are the measures used in the study the most relevant ones for answering the research question?

6. If the study involves following participants up over time, what proportion of people who were enrolled in the study at the beginning, dropped out? Have these “drop-outs” introduced bias?

7. Is the study long enough, and large enough to allow changes in the outcome of interest to be identified?

8. If two groups are being compared, are the two groups similar, and were they treated similarly within the study? If not, was any attempt made to control for these differences, either statistically, or by matching? Was it successful?

9. Was outcome assessment blind to exposure status? (That is, is it possible that those measuring the outcome introduced bias?)

(Adapted from the CRD handbook)
Case control studies

Case control studies compare groups of individuals with, and without a particular disease or condition, and examine their past history to attempt to determine the cause. There are fewer validated tools for assessing case-control studies, but the main sources of bias in these studies are well known. A useful appraisal checklist is shown in Box 5.4. One of the well-known biases is recall bias, in which cases may be more likely to recall potential risk factors for the outcome in question – for example, in case-control studies of stress and breast cancer, women with breast cancer and healthy controls are often asked about stressful life events in the past year, or several years. Breast cancer patients are indeed more likely to recall more such events, but this is less likely to be due to the events “causing” the cancer, than to women seeking harder to recall stressful events which might help to explain a subsequent diagnosis. This is understandable as it is a common but probably mistaken belief that stress causes cancer. In general, recall bias may introduce spurious associations between apparent risk factors and outcomes. One way to avoid this is to ensure that information is sought

Box 5.4  Appraising case-control studies

1. Are the study participants adequately described (with descriptive data on age, sex, baseline health status and other variables as appropriate to the research question)?
2. If the study is an assessment of an intervention, is the intervention clearly described, with details of who exactly received it?
3. If it is an etiological study (e.g., do food additives cause behavior problems?) were the independent and dependent variables adequately measured (that is, was the measurement likely to be valid and reliable)? Were they measured in the same way in both cases and controls?
4. Are the outcome measures used in the study the most relevant ones for answering the research question?
5. Are the two groups being compared similar, from the same population, and were they treated similarly within the study? If not, was any attempt made to control for these differences, either statistically, or by matching? Was it successful?

(Adapted from NHS CRD Report 4)
in exactly the same way from cases and controls – in the example given above, information should ideally be elicited by means of a **standardized** questionnaire, which should be administered by an interviewer who is blind to whether the interviewee is a case or a control.

**Interrupted time series (ITS) studies**

For some social interventions, few RCTs will have been carried out and the main evaluation methods will have been longitudinal studies, such as interrupted time series studies (ITS) (which are usually seen as part of the family of quasi-experimental studies). This is the case, for example, with evaluations of many fiscal interventions. For example the association between taxation levels and tobacco consumption can be examined by collecting data at several time points before and after the policy of interest was introduced. The Cochrane Effective Practice and Organization of Care (EPOC) group, which publishes systematic reviews of educational, behavioral, financial, organizational, and regulatory interventions as part of the Cochrane Collaboration (and which contributes to the work of the Campbell Collaboration), has developed a guide to the quality assessment of ITS studies (Box 5.5). Some of the items in this and other checklists will not apply to all types of social intervention, but nevertheless provide a valuable framework. The EPOC criteria were used as a guide to assessing the quality of ITS studies in a systematic review of new road building (Box 5.6). Longitudinal studies are already commonly included in systematic reviews that aim to assess the impact of policies, where RCTs or controlled studies have not been carried out. Examples include the Cochrane Tobacco Group’s systematic reviews, which have examined the impact of interventions delivered to communities to reduce smoking among young people and adults, and interventions that included the use of the mass media (such as advertising and television campaigns).

**Cross-sectional surveys**

Many of the study designs discussed above have a temporal dimension; that is, they involve following study participants over time, and recording events that happen to those people. However, many research studies have simple cross-sectional designs, where data are collected at just one point in time. Surveys are the most common example, and are commonly used to collect information on the use and acceptability of services. This type of process information is increasingly included in reviews of effectiveness – not
Box 5.5 Seven quality criteria for ITS designs from the EPOC group

A. Protection against secular changes

1. The intervention is independent of other changes:
   DONE if the intervention occurred independently of other changes over time
   NOT CLEAR if not specified, and will be treated as NOT DONE if information cannot be obtained from the authors
   NOT DONE if it is reported that the intervention was not independent of other changes in time

2. Sufficient data points to enable reliable statistical inference
   DONE if at least 20 points are recorded before the intervention
   AND the authors have done a traditional time series analysis (ARIMA model)
   OR at least 3 points are recorded pre- and post-interventions
   AND the authors have done a repeated measures analysis
   OR at least 3 points are recorded pre- and post-intervention
   AND the authors have used ANOVA or multiple t-tests
   AND there are at least 30 observations per data point
   NOT CLEAR if not specified in the paper (e.g., the number of discrete data points is not mentioned in the text or tables; treat as NOT DONE if the information cannot be obtained from the authors)
   NOT DONE if any of the above conditions are unmet

3. Formal test for trend (complete this section if the authors have used ANOVA modeling)
   DONE if formal test for change in trend using an appropriate method is reported
   NOT CLEAR if not specified in the paper (treat as NOT DONE if information cannot be obtained from the authors)
   NOT DONE if formal test for trend has not been done

B. Protection against detection bias

4. Intervention unlikely to affect data collection
   DONE if the authors report that the intervention itself was unlikely to affect data collection (for example, if sources and

(Continued)
Box 5.5  (Cont’d)

methods of data collection were the same before and after the intervention)

*NOT CLEAR* if not reported (treat as *NOT DONE* if information cannot be obtained from the authors)

*NOT DONE* if the intervention was likely to affect data collection (for example, any change in source or method of data collection)

5. Blinded assessment of primary outcomes

*DONE* if the authors state explicitly that the primary outcome variables (that is, the main outcomes relating to the author’s main hypothesis or question) were assessed blindly OR the outcome variables are objective (e.g., length of hospital stay, or other objective, independently recorded measure of effect, such as length of prison sentence). If some primary outcomes were assessed blindly, and others not, score each separately.

*NOT CLEAR* if not specified (treat as *NOT DONE* if information cannot be obtained from the authors)

*NOT DONE* if the outcomes were not assessed blindly

6. Completeness of data set

*DONE* if the data set covers 80–100 percent of the total number of participants or episodes of care in the study

*NOT CLEAR* if not specified (treat as *NOT DONE* if information cannot be obtained from the authors)

*NOT DONE* if the data set covers less than 80 percent

7. Reliable primary outcome measures

*DONE* if the study used two or more raters with at least 90 percent agreement, or with Kappa $\geq 0.8$ OR the outcome is obtained from some automated system

*NOT CLEAR* if reliability is not reported for outcome measures that are obtained by extracting information from medical or other charts, or are collected by an individual researcher (treat as *NOT DONE* if information cannot be obtained from the authors)

*NOT DONE* if agreement is $< 90$ percent, or Kappa is $< 0.8$

(If some outcome variables were assessed reliably and others not, score each separately)

*(See the EPOC group website http://www.epoc.uottawa.ca/resources.htm)*
least because it helps us understand not just whether something “works,”
but whether it is likely to be used. Some questions that should be borne in
mind when reading a report of a survey are outlined in Box 5.7.

As a final word on surveys, it is worth emphasizing, as Roger Jowell has
pointed out in relation to the British Social Attitudes Survey, that no finding
from a survey is definitive: “Every finding is an approximation, part of a body
of evidence which needs to be examined in the context of other evidence.”

There are many useful guides to critical appraisal of quantitative studies;
Crombie provides one readable overview and the NHS CRD Guide to
Systematic Reviews provides detailed guidance on this topic. Examples of
critical appraisal in education can also be found in the reviews of the EPPI-
Centre <http://eppi.ioe.ac.uk/EPPIWeb/home.aspx>, in the Research
Evidence in Education Library; these reviews show how the quality of each
study can be used to assess the overall “weight of evidence” in a review.

---

**Box 5.6 Quality assessment criteria used in a systematic review of the health and social impact of new road building**

Nine quality criteria were developed, based on previous epidemi-
ological criteria for assessing longitudinal studies, and on additional
criteria relevant to new road building:

1. Whether the researchers controlled for general trends;
2. Whether the data appeared to be a reliable/representative sample;
3. Whether sufficient data were presented to validate results;
4. Whether the authors controlled for regression to the mean (an
   approach commonly used in transport research when control
groups are absent);
5. Assessment of whether data are available for at least 3 years before
   and after the intervention was implemented;
6. Compares more than one new road;
7. Injury severity considered;
8. Number of individual casualties included;
9. Whether accident migration across wider road network was
   considered.

*(Egan et al., 2003)*

---
Box 5.7 Framework for appraising a survey

**General orientation questions**

- What question(s) is the study aiming to answer?
- Was the survey specifically designed with this question in mind?
- Do the survey measures used allow this question to be answered clearly?
- Is the population surveyed clearly described?
- How was the survey carried out?

**Selection of the sample**

Those who respond to surveys differ systematically from those who do not. For example, responders to household surveys are more likely to be at home and willing to answer the door; children who spend a lot of time out of the house or who live in residential care or a boarding school are likely to be under-represented in surveys of children of adult respondents to surveys. The size and nature of the biases is likely to vary from topic area to topic area (and from culture to culture).

- What is the response rate? If it is not reported, calculate it yourself, if enough data are presented. Is the response rate high enough to ensure that response bias is not a problem, or has response bias been analyzed and shown not to significantly affect the study?
- Is the denominator reported? (What information is given about the size and type of population from which the survey sample is drawn?)
- Is the sample surveyed representative? (i.e., representative of the population to whom the results will be generalized)

**Measurement issues**

- Are the measures reported objective and reliable? (e.g., in a study of diet, is a standardized, valid measure of diet used?)
Many of the study designs described above could probably be included in a review of effectiveness, though they would be used to answer different questions. Surveys for example might be included to answer questions on the likely take-up and acceptability of an intervention, while some of the

Survey methods

- How was the survey carried out? (e.g., postal survey, interview, web-based survey)
- Is the survey method likely to have introduced significant bias? (e.g., sampling method could include random versus quota sampling; web-based surveys, and so on)

Data and statistical issues

- Is the study large enough? (e.g., sample size justification, or discussion of statistical power)
- Is there an adequate description of the data? (including tables and summary statistics describing the sample, and adequate information on the results of any analyses)
- Is there evidence of multiple statistical testing, or large numbers of post hoc analyses?
- Are the statistical analyses appropriate?
- Is there evidence of any other biases? (e.g., funding bias)

(Adapted from Crombie, 1996)
other study designs would be more likely to be used to assess its effectiveness. However, evaluating interventions requires knowing more about the intervention than whether it works, and systematic reviews increasingly seek to include detailed information on the content and implementation of the intervention. This information is often overlooked in systematic reviews in favor of extracting information on methodological biases. However, without knowing what actually was done, information on whether something “worked” or not is unhelpful. We now discuss some of the process data that may be included in a systematic review. Not all of this information is likely to be reported in the original study, and contact with study authors may be necessary to obtain it.

5.9 TREATMENT INTEGRITY, IMPLEMENTATION, AND CONTEXT

Treatment integrity (sometimes referred to as program adherence) refers to the degree to which an intervention was implemented as prescribed by the study protocol. The rationale for assessing this in systematic reviews is that it is assumed that effectiveness is directly related to the fidelity with which the intervention is implemented.\textsuperscript{53, 54} One approach to assessing treatment integrity is to code information from each study relating to aspects of the content and delivery of the intervention, and then to examine the association between these indicators and outcomes (such as effect sizes, as was done by Devine and Cook in their meta-analysis of the effects of psycho-educational interventions on lengths of stay in hospital).\textsuperscript{25, 55} For many reviews, however, and particularly for reviews of complex social interventions, more is required than simply coding methodological aspects of the study. One also needs to assess both the fidelity and the “intensity” of the intervention as it was delivered.

DuBois et al.’s meta-analysis of 55 studies of youth mentoring programs examined these issues, in each case assessing 125 aspects of the program.\textsuperscript{56} They rated each study on 11 features relating to the quality of implementation of the intervention. These features had been included in previous recommendations for establishing effective mentoring programs, and the reviewers used these to produce a theory-based “index of best practice” (Box 5.8). This in turn was used to explore the association between best practice and effect size. Larger effect sizes were found in the programs that had engaged in a majority of the 11 “best practices.” This gives a demonstration of how implementation can be assessed and used to explore
moderator effects in a systematic review (in this case a meta-analysis was carried out).

Lumley et al. describe how their review of smoking cessation interventions aimed at pregnant women sought to include a range of non-trial data.57 This review included reports on who developed the interventions, how they were developed, their theoretical basis, and how they were delivered. They describe how reviewing these process indicators helped explain some of the heterogeneity in outcomes, as there was heterogeneity in how the interventions were delivered (some poorly, some not), and some were delivered more intensively than others. The reviewers found that

---

**Box 5.8 Assessing implementation in a meta-analysis of the effectiveness of youth mentoring programs**

DuBois et al. began with an extensive coding of the characteristics of each study. This included not just methodological issues, but also the geographical location, program goals, characteristics of the mentors and participating youths, characteristics of the mentor–mentee relationships, among many others. They also created a theory-based index of best practice to provide an assessment of the program and its implementation. Eleven program features were assessed, and a score was calculated based on the number of best practices each program adhered to. The program features included:

- Monitoring of program implementation
- Screening of prospective mentors
- Matching of mentors and young people
- Mentor pre-match training
- Ongoing training of mentors
- Supervision of mentors
- Existence of support group for mentors
- Structured activities for mentors and young people
- Parent support or involvement component
- Expectations for both frequency of contact
- Expectations for length of relationships

*(DuBois et al., 2002)*56
those interventions delivered more intensively and those that were theoretically based appeared to have the greatest impact.

The following list of issues to consider when describing interventions in a systematic review may be helpful. This list draws on existing criteria for appraising process evaluations:

- Details of intervention activities (what the intervention consists of, where, how, how often, and by whom it is delivered, and for how long);
- Resources: this includes: time, money, people, information, technology, and other assets needed to conduct the intervention;
- Staffing, and where appropriate to the intervention, the level of skills of staff;
- Context (as described above, this may include the political, community, organizational, and other contexts);
- The stage of development of the intervention (whether it is a mature, well-developed intervention, which perhaps has been implemented elsewhere);
- Sustainability (including any measures taken to ensure retention of people in the study);
- Whether any adverse or negative effects were observed or reported by researchers or participants; and
- The logic model: how the intervention is theorized to bring about change in the desired outcome – that is, by what steps, or by what theorized causal pathway.

This last point may be difficult to identify or may not be reported in the research paper or report, but it will help the user of the review to understand the purpose of the intervention, and why anyone ever thought it could work. As illustration, this information was extracted by Secker-Walker et al. in their review of community level smoking cessation interventions:

In 23 studies (72%) there was a description of the theoretical background for the interventions. Fourteen (44%) drew explicitly on social cognitive theory, 11 (34%) on communication theory and 10 (31%) on diffusion of innovation theory. All three theories were specified in seven studies (22%). In 18 (56%) studies, other theories were used including community participation in seven (22%), stages of change in five (16%), social marketing in three (9%), and planned behaviour or reasoned action in three (9%). The PRECEDE model of planning and evaluation for health education and policy interventions was used in six (19%) studies.
Anderson et al. provide another example of a logic model, in which they clearly outline the theoretical framework underlying their systematic review of the effectiveness of early childhood development programs. In this, they describe the pathways linking the components of childhood development programs, which run from intermediate social, health, and other outcomes (such as improvement in social skills), through to final outcomes such as educational attainment, and changes in health-related outcomes (such as risk behaviors like drug and alcohol use). Their approach uses the “Guide to community preventive services,” which has been used to produce a wide range of reviews of healthcare and social interventions. For example, another review in this series summarizes evidence on the effectiveness and economic efficiency of interventions to promote healthy social environments, which covers housing, educational, and other programs (see <http://www.thecommunityguide.org/>).

The EPPI Centre has also produced a 12-question checklist to help with assessing process evaluations in health promotion. The full version of this checklist can be found at: <http://eppi.ioe.ac.uk/EPPIWeb/home.aspx?page=/hp/reports/phase/phase_process.htm>.

**Incorporating information on context**

The context within which a study is carried out also affects the effectiveness of the intervention; this will vary between studies within a single review. Rychetnik et al. suggest that important contextual characteristics could include factors in the political or organizational environment, as well as socioeconomic or demographic characteristics of the population in the study, though some of these contextual factors may not be reported consistently (Box 5.9).

If enough information is reported (or can be gleaned from the authors) context can be taken into account in interpreting the review’s results. For example, “welfare to work” programs have been implemented in many places including the US, Scandinavia, and the UK, but they vary in their content, and their impact is likely to depend on the political and other settings within which they are delivered. A systematic review of such studies would need to interpret the data they collect in the light of this variation. Alternatively the reviewers could reduce variability by confining the review to evaluations carried out in countries with similar welfare systems, or could confine the review to one country. One recent review took the latter approach. It examined the evidence on the impact of the UK’s welfare to work programs on the employment rates of people
with a disability or chronic illness, and included 17 studies in the final synthesis. Its conclusion was that “welfare to work” has a positive effect on employment rates.62

### 5.10 ASSESSING EXTERNAL VALIDITY

Even if a study has high internal validity, it may not be generalizable (high external validity). There is often a trade-off between internal and external validity. One starting point in determining generalizability is to explore whether the study population appears to be representative of the population to which you wish to apply the results. However, even if the populations are

---

**Box 5.9 Finding evidence on implementation: A case study**

In a study funded by the Health Development Agency (now incorporated into NICE), Arai et al. started to map the evidence on the implementation of interventions aiming to reduce accidental injuries, to explore ways in which this evidence can be quality appraised and synthesized in a manner accessible to policymakers and practitioners. Their review found that:

- many papers contained little information on implementation processes;
- a majority of the papers containing material on implementation processes were judged to have too little information on study design to allow study quality to be assessed; and
- in many cases, it was unclear whether “implementation relevant” statements were based on research evidence.

Some journal editors have already sought to improve the quality, not only of research reporting, but also of research conduct, through guidelines on presentation and content. There is scope for further improvements in reporting on implementation issues.

*(Arai et al., 2005)*61
similar, they may differ in other ways; as above, differences in the settings, and in cultural or other contextual factors, should also be considered.

If the study involves an evaluation of an intervention, then its transferability needs consideration; what is acceptable to one population at one point in time, may not be acceptable to another. Ethical appropriateness may also vary between populations; an intervention which aims to prevent harm may be ethical in a high-risk population, but not in the general population. This may be the case with screening programs for example. Conversely, categories of interventions that appear effective in some populations may have the opposite effect in others; Pawson gives the example of “naming and shaming” (public disclosure of misdemeanors). This penalty may work when aimed at car manufacturers who make cars that are easily broken into and stolen – but may have the opposite effect in other situations.

Reviewing the results of a number of studies of course itself provides a test of generalizability; if the results have been replicated in several settings with different populations, then this gives an indication of whether the results are transferable. If the number of studies is large enough, it can suggest the range of effect sizes to be expected in different settings. Generalizability is not often assessed separately in systematic reviews, though consideration of the issue is included in some critical appraisal checklists. Rychetnik et al. suggest that assessing this requires obtaining information on the intervention itself, on the context within which it is delivered and evaluated, and on interactions between the intervention and its context – sometimes referred to as effect modification (see also chapters 6 and 7).

As some of the examples above show, theory has an essential role to play in systematic reviews; a theoretical model of how the intervention works, and for whom, is important when deciding on the review question, and what types of studies to review. It will help in interpreting the review’s findings, and will be valuable in assessing how widely applicable those findings may be. In turn systematic reviews can contribute to developing and testing the limits of theories, by examining how contextual or temporal variables moderate outcomes. Theories themselves can also be the subject of systematic reviews.

5.11 QUALITY ASSESSMENT AND QUALITATIVE RESEARCH

Information on process and implementation issues will often, though not exclusively, be derived from qualitative research, and there are systematic reviews that have explicitly set out to review only qualitative studies.
For example, systematic reviews may use qualitative research to explore barriers or facilitators of the implementation of interventions, or to synthesize qualitative evidence on people’s experiences. An example of the latter is given in Box 5.10.

**Box 5.10  Resilient young mothering: A systematic review of qualitative research on the experiences of teenage mothers**

The problem: Teenage pregnancy and motherhood is a central focus of the UK government’s policies to tackle social exclusion and health inequalities.

The review: It is widely acknowledged that an understanding of the needs and experiences of young mothers is essential for the development of effective services, an understanding provided by studies that record their experiences of their lives. The authors therefore searched for reviews of qualitative studies of teenage mothers and mothers-to-be. To ensure the review’s relevance to the UK policy context, it focused on UK studies of teenage pregnancy and motherhood published from 1990 to 2003, seeking out primary studies that recorded and analyzed the accounts of young women. As far as possible, the review used systematic review methodology developed for quantitative studies, but drew on templates developed by qualitative researchers who have tried to adapt this methodology for the critical appraisal and synthesis stages of the review.

The critical appraisal: The authors found that methods for critically appraising qualitative research and for synthesizing their findings are still under development, with as yet no consensus about, and therefore no accepted guidelines for, these two key stages of a systematic review. If there is any agreement, it is that a single set of methods is unlikely to be appropriate for qualitative reviews: it is the purpose and type of the review that should drive the choice of appraisal criteria and synthesis strategy.
There is ongoing debate about the appropriateness of quality assessment methods to qualitative research, and about what the “goals” of such methods of appraisal should be. However, it is clear that one needs to be able to distinguish “good quality” from “poor quality” qualitative research, just as one does for quantitative studies. In some cases, good quality studies may not necessarily be fit for the purpose of a particular systematic review, since, unlike quantitative studies, they will less frequently directly answer the review question. If you are doing a review on maternal views of infant growth for instance, it will be a bonus to find a study on precisely this subject. However, other studies may well present relevant information, and in all cases, it will be appropriate to ask just how well the study was carried out. Nonetheless, while there is now a large number of critical appraisal tools for qualitative research, there are as yet no widely accepted criteria as to “the best” method for qualitative study appraisal. Work exploring this issue is currently in progress.

Other frameworks and guidance have been developed to aid in appraising qualitative research, and these will be of value in systematic reviews which include qualitative studies. A recent example was produced by Spencer et al. at the National Centre for Survey Research in the UK, working on behalf of the UK Government’s Cabinet Office. Their framework was
developed with particular reference to evaluations concerned with the
development and implementation of social policy, programs, and practice.
It drew on a systematic review of existing literature on approaches to
judging the quality of qualitative research, and was based around 18 open-
ended appraisal questions, which are intended to be applied mainly to four
methods used in qualitative research: in-depth interviews, focus groups,
observation, and documentary analysis (Box 5.11). The authors emphasize
that the framework is explicitly designed to aid informed judgments, rather

Box 5.11 Eighteen appraisal questions for qualitative research

1. How credible are the findings?
2. How has knowledge or understanding been extended by the research?
3. How well does the evaluation address its original aims and purpose?
4. How well is the scope for drawing wider inference explained?
5. How clear is the basis of evaluative appraisal?
6. How defensible is the research design?
7. How well defended are the sample design/target selection of cases/documents?
8. How well is the eventual sample composition and coverage described?
9. How well was the data collection carried out?
10. How well has the approach to, and formulation of, analysis been conveyed?
11. How well are the contexts of data sources retained and portrayed?
12. How well has diversity of perspective and content been explored?
13. How well have detail, depth, and complexity (i.e. richness) of the data been conveyed?
14. How clear are the links between data, interpretation and conclusions – i.e., how well can the route to any conclusions be seen?
15. How clear and coherent is the reporting?
16. How clear are the assumptions/theoretical perspectives/values that have shaped the form and output of the evaluation?
17. What evidence is there of attention to ethical issues?
18. How adequately has the research process been documented?

(Chenery et al., 2003)

**Generalizability and qualitative research**

Qualitative research that puts a metric on findings should ring alarm bells. To be told that “most” children don’t like greens, or “few” women enjoy an episiotomy on the basis of a qualitative study are unlikely to be the kind of data that would add much to knowledge. Qualitative work depends not on numerical but conceptual analysis and presentation. As Fitzpatrick and Boulton point out, if qualitative data are reported mainly in terms of frequencies and proportions of respondents with a particular view, that is a quantitative study. This is very far from saying that “anything goes” in qualitative research, but in trying to ram qualitative research into a glass slipper of a different research paradigm, the kinds of understandings that sound qualitative research can bring are likely to be lost.

It is more likely that high quality qualitative research can help generate new hypotheses on why quantitative research shows what it does, what the appropriate and acceptable interventions might be, and whether the “right” kinds of questions are being asked. It should of course be presented in a context that describes the background and characteristics of those who are part of the study, and the strengths and limitations of these data in terms of lessons that might be drawn from them. As Green puts it: “Data extracts taken out of context tell us little about the situated nature of beliefs and behaviour, and inferences that are not rooted in a theoretical understanding are unlikely to be generalisable to other settings.”

**Quality in qualitative research: Are checklists sensible?**

Checklists for qualitative research have been criticized for their potential as “tail wagging the dog” exercises, and the possibility that researchers may contort their work to meet checklist criteria, while paying less attention to some of the more basic requirements of scientific rigor. Checklists are not a shortcut to ensuring quality in qualitative research, but when using a research method such as systematic reviewing, one of whose claims to legitimacy is transparency, checklists provide one means of ensuring that included studies contain sufficient information about the sample, the question, the data analysis, and so on. This enables those reading a review to check that the basics are reported in each study.
5.12 DOES QUALITY ASSESSMENT MATTER?

It is sometimes assumed that biases in primary studies are “self-canceling;” that is, that errors occur randomly across the set of studies in the review and so in aggregate (if there are enough studies in the review) they will cancel each other out.\textsuperscript{25} If this was the case, then the assessment of internal validity would not be crucial. However, there is now ample evidence that study biases in general operate systematically, and often result in inflated effect sizes. For example, non-random assignment in RCTs has been shown to over-estimate effects by at least one-third – though systematic underestimates of effects have also been observed.\textsuperscript{25}

Schulz et al.’s study of 250 trials provides a clear empirical demonstration of the effects of study biases on outcomes.\textsuperscript{79} Trials in which concealment was either inadequate or unclear yielded larger effect size estimates than those that had adequately concealed the treatment allocation; odds ratios were exaggerated by 41 percent for inadequately concealed trials (adjusted for other aspects of quality). Trials that were not double-blinded also yielded larger estimates of effects, with odds ratios being exaggerated by 17 percent.

The case for critical appraisal is also supported by an RCT carried out among general practitioners, which compared critical appraisal (using a checklist) versus “free” appraisal, in which readers were asked to score papers on the basis of “their importance to me in everyday work.” This showed clear differences. Those using the critical appraisal tool gave a consistently lower overall score, and applied a more appropriate appraisal to the methodology of the studies.\textsuperscript{80}

5.13 DATA EXTRACTION

Critical appraisal is often carried out as an integral part of the data extraction process. Data extraction is the process of extracting the relevant information from each study, either by copy- ing it onto printed pro-forma templates, or by directly entering it into a database or table. An example of a data extraction sheet for quantitative studies appears in Appendix 4. Other example data extraction sheets are available from the EPOC website (<http://www.epoc.uottawa.ca/resources.htm>) and from the CRD guide to systematic reviews (available on the CRD website <http://www.york.ac.uk/inst/crd/>).
5.14 CHECKING THE RELIABILITY OF CRITICAL APPRAISAL AND DATA EXTRACTION

Wortman and others refer to “data extraction bias” in which the reviewer introduces bias into the review – perhaps by differentially extracting information from studies that are in accord with their own views, or by applying differential judgments of quality or methodological adequacy to different studies. Data-extraction bias may also (in theory, at least) be introduced by the reviewer’s awareness of the study authors or the journal, or their disciplinary background, or by awareness of other aspects of the study being reviewed. Some of these issues can be dealt with by blinding the reviewers (for example, by obscuring study details with tape or a marker), by using multiple reviewers, and by validating the data extraction. In the latter case, this can be done by ensuring that all data extraction is checked by another person, (a common practice), or by double data extraction by two people working independently (much less common, but likely to be more reliable).

5.15 EXTRACTING DATA ON DIFFERENTIAL EFFECTS OF INTERVENTIONS

Research studies often report their results at a population-level; that is, the study findings are reported as if the participants were an undifferentiated group. However, for many interventions there are likely to be interactions that modify their effects – for example, interactions between characteristics of the participants and their outcomes. Interactions with level of education are common in many health promotion interventions, as are interactions with age: for example a systematic review by Shepherd et al. found some suggestion that the effectiveness of healthy eating interventions in school-children varied with age, and found a distinct interaction with gender (knowledge and consumption of healthy foods were more likely to improve in young women than in young men). Where data on impacts in sub-groups are reported this should be extracted by the reviewer. This is discussed further in chapter 7.
5.16 TRANSLATING NON-ENGLISH PAPERS

The cost of translating non–English language papers can be significant and difficult to estimate. Full-text translations in particular can be expensive. One way of reducing cost without compromising the quality of the information extracted is to translate only those selected sections of the paper that are essential – this often excludes the introduction, and discussion – and to translate key sections of the methods and results, and headings for tables and figures. One problem with this approach is that important contextual information may be excluded through this process.

An English language abstract is often available, though this is sometimes hidden at the back of the paper. Instead of employing a translator, it may be more efficient to send them a copy of the paper to read, and then question them to decide whether it meets the inclusion criteria. If you are close to a language school or university language department, this can be a useful source of help. Cheaper still is to do it yourself. Even with limited linguistic skills, a knowledge of the main methodological terms in that language can be enough to allow a paper’s key details to be extracted. Web-based translators (such as Alta Vista’s Babelfish (<http://world.altavista.com/>)) can be useful as they are often accurate enough to allow the reviewer to determine whether the study meets the review’s inclusion criteria (see Box 5.12).

Box 5.12 How to avoid getting “Lost in translation”

The decision about which studies to translate can be made on grounds of relevance. For example, a recent systematic review of infant growth investigated interventions that might lead to optimal growth for children.82 One of the researchers, Patricia Lucas, started by scanning relevant world literature, including literature from the countries of origin of the UK’s main immigrant groups. For this, she used national data on country of origin. Given that most immigrants gravitate towards the larger cities, and in particular capital cities, she then obtained data on the main languages spoken in London. This information helped her develop inclusion and exclusion criteria to help decide (on the basis of the abstract) which articles needed to be translated.
Despite the difficulties in applying critical appraisal, it is an essential part of a systematic review, though unfortunately it still too often remains undone. Ignoring critical appraisal on grounds that studies are probably good enough, or that the biases cancel each other out, is simply misleading and makes for a potentially misleading review. In some cases it may be harmful to take studies at face value, and so systematic reviews that have made no attempt at critical appraisal should not be considered reliable.

Key learning points from this chapter

- Biases in individual studies can affect their conclusions; ignoring this fact can in turn bias the conclusions of a systematic review.
- Critical appraisal is, therefore, an essential step in any systematic review; without this the review may be unreliable.
- Critical appraisal tools are available for all major study designs, and guides to the appraisal of qualitative research have been produced.
- This is generally done with a view to systematizing the assessment of the reliability and utility of the information presented by the studies, and not as an exercise in criticism for its own sake.
- The appraisal of the primary studies is sometimes carried out in tandem with the data extraction; in extracting data, information on subgroup effects, and on moderator variables (that is, variables which appear to moderate the effects of other important independent variables) should also be extracted where available.

REFERENCES


8. CONSORT Guidelines, online at: http://www.consort-statement.org/


73. Popay, J., Rogers, A., and Williams, G. Rationale and standards for the systematic review of qualitative literature in health services research. Qualitative Health Research 1998, 8: 341–51.


