Chapter 7

Assessing Study Quality

As well as defining and applying inclusion and exclusion criteria to select *relevant* studies from a set of candidates, it is also important for many types of review to define and apply criteria for assessing the *quality* of the selected primary studies. This stage of the process is highlighted in Figure 7.1.

In this chapter we discuss three aspects of quality assessment:

- *Why* (and when) it is important to assess quality
- Defining the *criteria* to use for quality assessment
- Establishing and applying *procedures* for performing quality assessment

We note also, that although quality assessment is quite distinct from data extraction, which is covered in Chapter 8, these two stages can be performed sequentially or together (performing data extraction and quality assessment on a study-by-study basis).

Examples of quality assessment criteria and of procedures for applying them are described to illustrate some of the approaches taken by systematic reviewers in software engineering.

7.1 Why assess quality?

Quality assessment is about determining the extent to which the results of an empirical study are valid and free from bias. For systematic reviews

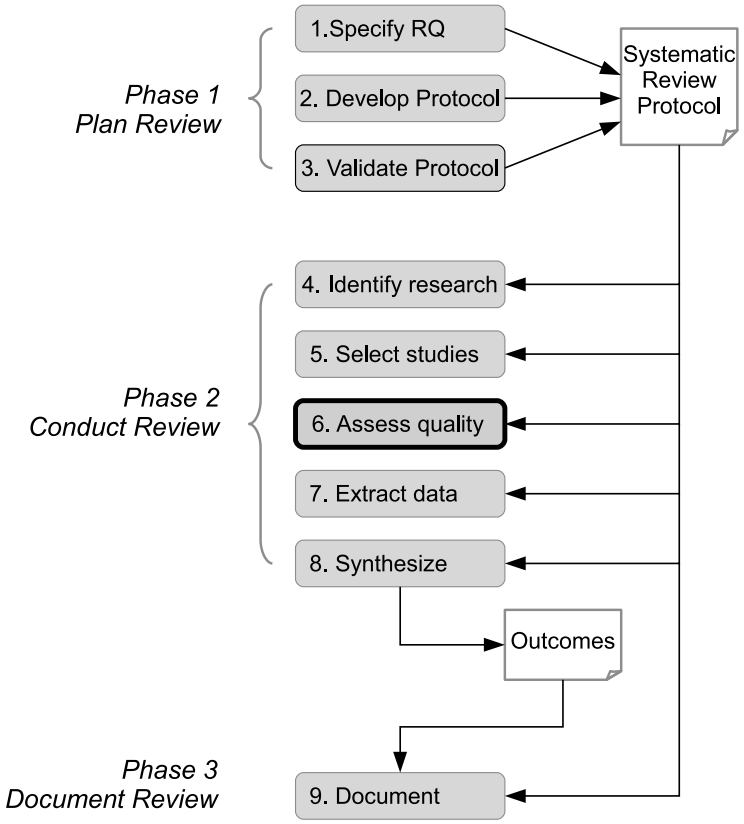

FIGURE 7.1: Quality assessment stage of the systematic review process.

and for some types of mapping study, evaluating the quality of the primary studies contributing to a review can enhance its value in a number of ways. For example:

- Differences in the quality of primary studies may explain differences in the results of those studies
- Quality scores can be used to weight the importance of individual studies when determining the overall outcomes of a systematic review or mapping study
- Quality scores can guide the interpretation of the findings of a review

For quantitative systematic reviews in particular, it is essential to assess the quality of the primary studies included in the review because if their results are invalid (or there is doubt about their validity) or if they are biased in some way, then this should be taken into account during the synthesis process. Reviewers might (simply) choose to exclude low quality primary studies from the synthesis process or they may choose to check whether their exclusion has a significant effect on the overall outcomes of a review. There have been a number of reports in the medical domain that have shown that if low-quality studies are omitted from the synthesis process of a systematic review (or from

a meta-analysis) then the results of the review (or analysis) change. One example is a systematic review of homoeopathy which suggested that it performs well if low-quality studies are included, whereas high-quality studies found no significant effect (Shang, Huwiler-Müntener, Nartey, Jüni, Dörig, Sterne, Pewsner & Egger 2005). If reviewers intend to exclude low quality studies then effort can be saved by assessing quality in advance of data extraction.

Quality assessment can be of less importance when undertaking a mapping study since the focus for these is usually on classifying information or knowledge about a topic. However, it can be important for some mapping studies, especially tertiary studies, if for example their research questions relate to changes in quality over time.

Assessing the quality of a primary study is a particularly challenging task as there is no agreed, standard definition of study 'quality'. Many of the guidelines and criteria for assessing study quality, some of which are described later in this chapter, indicate that quality relates to the extent to which the design and execution of a study minimises bias and maximises validity. These concepts are summarised in Table 7.1 (see also Section 2.5.2). Further discussions and pointers to relevant literature about quality, bias and validity, can be found in Dybå & Dingsøyr (2008*b*).

TABLE 7.1: Quality Concepts

Term	**Synonyms**	**Definition**
Bias	Systematic error	A tendency to produce results that depart systematically from the 'true' results. Unbiased results are internally valid.
Internal validity	Validity	The extent to which the design and conduct of a study are likely to prevent systematic error. Internal validity is a prerequisite for external validity.
External validity	Generalisability, Applicability	The extent to which the effects observed in a study are applicable outside of the study.

As well as being an intrinsically difficult task to perform consistently, quality assessment is confounded by:

1. constraints imposed by the publication venue for papers reporting a primary study,

2. the range of primary study types included in a review.

The first of these factors, the constraints, usually relating to length, imposed by publishers can mean that researchers are not able to include all of the details of their study in a single paper. This is particularly problematic for conference proceedings where papers are often limited to 10 or 12 pages. This can result in the omission of important methodological (and other) information that would provide *evidence* of study quality. It can also lead to a study

being reported in more than one paper adding to the difficulties of mapping papers to studies (as discussed in section 6.3). One approach to alleviating publishing constraints is to provide supplementary material on an associated web site. This facility is supported by a number of publishers. Another way of publishing more detailed information about a study than can be included in a single conference paper is to report different aspects of a study across multiple conference papers (which of course adds difficulties for the reviewer who has to extract and combine these) or in journal papers where length is usually less restricted.

The second of these factors which relates to the types of primary study, can cause problems where the studies included in a review are of diverse types. For example reviews can include quantitative studies, such as experiments and quasi-experiments, and qualitative studies, such as case studies and ethnographic studies. Here a dilemma arises — whether to use a generic set of quality criteria across all of the studies included in a review, regardless of their type, or whether to use specific sets of criteria tailored to each of the types of study that occurs in the set of primary studies. Not surprisingly each of these options has some strengths and some limitations. We look more closely at this issue in Section 7.2.2.

Once the decision has been made to assess the quality of the primary studies included in a systematic review or mapping study (and to use that assessment during the synthesis stage of the review) then a reviewer has two key questions to address. These are:

1. Against what criteria will the primary studies be assessed?

2. How will the assessment be performed and who should do it?

We also note that some reviews include non-empirical papers such as those reporting lessons learned or discussing some aspect of the topic of the review. The quality criteria discussed in this chapter are not appropriate for these types of papers.

7.2 Quality assessment criteria

A large number of quality assessment criteria and checklists for different types of empirical studies are published in the medical and social sciences literature. In addition to those indicated in the following section, the Support Unit for Research Evidence (SURE)[1], in the UK, provides a range of relevant links and resources including a set of critical appraisal checklists for quantitative studies, qualitative studies and systematic reviews. Work in the

[1] http://www.cardiff.ac.uk/insrv/libraries/sure/

medical and social science fields has provided the basis for many of the quality checklists proposed, used and/or evaluated for empirical studies in the software engineering field. We summarise here the checklists most widely used in software engineering reviews and also briefly discuss the problems associated with quality assessment across multiple study types.

7.2.1 Study quality checklists

A number of checklists that are tailored to specific study types have been proposed. For case studies, Runeson et al. (2012) present a checklist for both readers of case studies and for researchers who are performing case studies. These checklists are synthesised from a range of sources including literature in the social sciences and information systems fields and adapted to software engineering. The checklists for readers (and hence for reviewers) of case studies can be used to assess the quality of case studies included in a review. Primary studies of this type are commonly found in qualitative systematic reviews and mapping studies. The readers' checklist is shown in Table 7.2. Further details of the case study methodology and its use in systematic reviews can be found in Chapter 18.

TABLE 7.2: A Case Study Quality Checklist (Taken from Runeson, P., Höst, M., Rainer, A. & Regnell, B. (2012)). Reproduced with permission.

	Criteria
1.	Are the objectives, research questions, and hypotheses (if applicable) clear and relevant?
2.	Are the case and its units of analysis well defined?
3.	Is the suitability of the case to address the research questions clearly motivated?
4.	Is the case study based on theory or linked to existing literature?
5.	Are the data collection procedures sufficient for the purpose of the case study (data sources, collection, validation)?
6.	Is sufficient raw data presented to provide understanding of the case and the analysis?
7.	Are the analysis procedures sufficient for the purpose of the case study (repeatable, transparent)?
8.	Is a clear chain of evidence established from observations to conclusions?
9.	Are threats to validity analyses conducted in a systematic way and are countermeasures taken to reduce threats?
10.	Is triangulation applied (multiple collection and analysis methods, multiple authors, multiple theories)?
11.	Are ethical issues properly addressed (personal intentions, integrity, confidentiality, consent, review board approval)?
12.	Are conclusions, implications for practice and future research, suitably reported for its audience?

A quality checklist constructed for technology-intensive testing experiments is described by Kitchenham, Burn & Li (2009). The checklist focuses specifically on studies relating to testing; however, reviewers addressing other technology-intensive topics, such as cost estimation and performance, might find the approach to checklist construction and validation of interest. An adaptation of this checklist with suggestions for scoring each of the questions is shown in Figure 22.8.

A further quality checklist was developed and used for a qualitative, technology-focused systematic review on Agile methods (Dybå & Dingsøyr 2008*b*, Dybå & Dingsøyr 2008*a*). The 11 criteria making up the checklist were based on those proposed for the Critical Appraisal Skills Programme[2] and by the principles of good practice for empirical research in software engineering described by Kitchenham, Pfleeger, Pickard, Jones, Hoaglin, El Emam & Rosenberg (2002). The criteria, shown in Figure 7.3, cover four main areas of empirical research:

- *Reporting* - criteria 1-3 relate to the quality of reporting an empirical study,
- *Rigour* - criteria 4-8 address the details of the research design,
- *Credibility* - criteria 9 and 10 focus on whether the findings of the study are valid and meaningful,
- *Relevance* - criteria 11 concerns the relevance of the study to practice.

In the systematic review on Agile methods, the reviewers applied the checklist to 33 empirical studies, 24 of which were case studies, four were surveys, three were experiments and two used a mix of research methods. This checklist has been quite widely used by reviewers in software engineering as a basis for quality assessment. See for example, the reviews by Alves, Niu, Alves & Valença (2010), Chen & Babar (2011) and Steinmacher, Chaves & Gerosa (2013).

As discussed in Section 3.2, a tertiary study is a mapping study where systematic reviews and mapping studies constitute the 'primary' studies under review. Many researchers who undertake tertiary studies carry out quality assessment in order to identify trends in the quality of systematic reviews and/or mapping studies. To date, criteria to assess the quality of systematic reviews and mapping studies have not been developed specifically for software engineering reviews. However, one of the sets of criteria used in the medical domain, the DARE[3] criteria[4], has been applied in a number of tertiary studies. The criteria were initially based on four questions, with a fifth being added later. The five questions are:

[2] www.casp-uk.net
[3] Database of Abstracts of Reviews of Effects
[4] http://www.crd.york.ac.uk/CRDWeb/AboutPage.asp

TABLE 7.3: A Quality Checklist That Can Be Used across Multiple Study Types (Taken from Dybå, T. & Dingsøyr, T. (2008a)). Reproduced with permission.

	Criteria
1.	Is the paper based on research (or is it merely a 'lessons learned' report based on expert opinion)?
2.	Is there a clear statement of the aims of the research?
3.	Is there an adequate description of the context in which the research was carried out?
4.	Was the research design appropriate to address the aims of the research?
5.	Was the recruitment strategy appropriate to the aims of the research?
6.	Was there a control group with which to compare treatments?
7.	Was the data collected in a way that addressed the research issue?
8.	Was the data analysis sufficiently rigorous?
9.	Has the relationship between researcher and participants been adequately considered?
10.	Is there a clear statement of findings?
11.	Is the study of value for research or practice?

1. Are the review's inclusion and exclusion criteria described and appropriate?

2. Is the literature search likely to have covered all relevant studies?

3. Did the reviewers assess the quality/validity of the included studies?

4. Were basic data/studies adequately described?

5. Were the included studies synthesised?

Examples of the use of the DARE criteria include the broad tertiary studies reported in Kitchenham, Brereton, Budgen, Turner, Bailey & Linkman (2009), Kitchenham, Pretorius, Budgen, Brereton, Turner, Niazi & Linkman (2010) and da Silva et al. (2011) as well as a tertiary study by Cruzes & Dybå (2011*b*) which focused on research synthesis.

A number of other approaches to assessing the quality of systematic reviews are used within the medical domain, some of which are discussed in Dybå & Dingsøyr (2008*b*). In addition, we highlight two initiatives related to systematic reviews and meta-analyses within the clinical medicine field. One of these is the PRISMA[5] Statement which aims to help authors improve the reporting of systematic reviews and meta-analyses (Liberati, Altman, Tetzlaff, Mulrow, Gøtzsche, Ioannidis, Clarke, Devereaux, Kleijnen & Moher 2009). It is suggested that 'PRISMA may also be useful for critical appraisal of published systematic reviews'. However Moher, Liberati, Tetzlaff & Group (2009) do note that the PRISMA checklist is not a quality assessment instrument.

[5]Preferred Reporting Items for Systematic reviews and Meta-Analyses, http://www.prisma-statement.org/

A project undertaken by the Cochrane Editorial Unit (CEU) aims to specify methodological expectations for Cochrane protocols, reviews and review updates. As a result of this work, the CEU have produced a report describing methodological standards for the conduct of new Cochrane Intervention Reviews[6]. The report describes a checklist of 80 attributes relating to the conduct of reviews, indicating in each case whether they are considered mandatory or highly desirable.

7.2.2 Dealing with multiple study types

Many systematic reviews and mapping studies in software engineering include primary studies that utilise a range of different empirical methods. These typically include those methods described in Part II of this book. Where the primary studies are of a single type (for example, they are all case studies or all experiments) then a quality checklist can be selected or tailored for that specific study type. However, where a review includes multiple study types, researchers have to decide whether to use a single checklist or a set of type-specific checklists.

When a single quality checklist is used for a systematic review or mapping study, researchers have to consider which of the criteria (that is, which checklist items) are applicable for each study type. Of course this means that it is necessary to extract (and validate) the study type for each primary study before carrying out a quality assessment. When scores for a particular study are aggregated across the checklist items against which the study is assessed, the number of applicable items needs to be taken into account through a normalising process (see the third example in Section 7.4 which illustrates this approach).

Where multiple quality checklists are used, the same requirement to determine the study type arises. In this case, the study type is used to select the most appropriate checklist.

One problem that arises when there are multiple study types is that aggregated scores cannot be compared in a meaningful way across the different types. So, it becomes quite challenging to interpret these when considering the findings of a review. See Part III for further discussion about using quality assessment results from different types of study.

7.3 Procedures for assessing quality

Here we consider three aspects of the process of assessing the quality of empirical studies. These are:

[6] http://www.editorial-unit.cochrane.org/mecir

- *Scoring* studies against the checklist(s) used
- *Validating* the scores
- *Using* quality assessment results

Scoring studies

If a single checklist is being used, then each study will be scored against each criterion that is appropriate for the study type. If multiple checklists are used, then reviewers have to select the appropriate checklist and score the study against the items in that checklist. A number of approaches to scoring have been taken by reviewers. Some use a simple yes(1)/no(0) score (see, for example, Dybå & Dingsøyr (2008*a*) and Cruzes & Dybå (2011*b*)) whilst others recognise partial conformance to a criterion. For example da Silva et al. (2011) use a 3-point scale (yes(1)/partly(0.5)/no(0)). Whatever scale is used, it is important to ensure consistency by documenting the specific characteristics of a study that map to specific points on the scale.

Validating scores

As we have seen in Section 6.2, validation is an important element in maintaining confidence in the procedures and hence the outcomes of a review. The same options as are discussed for study selection are possible for validating quality scores. If quality assessment is being carried out by a team of researchers, then two or more members of the team can score each of the studies followed by a process of resolution. The process by which researchers obtain a consensus about the quality of a paper given a quality checklist has been investigated through a series of studies (Kitchenham, Sjøberg, Dybå, Brereton, Budgen, Höst & Runeson 2013). These studies found that using two researchers with a period of discussion did not necessarily deliver high reliability (that is, consistency in using a checklist) and simple aggregation of scores appeared to be more efficient (that is, involved less effort) than incorporating periods of discussion without seriously degrading reliability. The authors of the studies suggest using three or more researchers, where this is feasible, and taking an average of the total score using the numerical values of the scores. In contrast, a study by Dieste, Griman, Juristo & Saxena (2011) recommends against using aggregate scores and recommends only using validated checklist items.

Where quality assessment is being performed by a single researcher, such as a PhD student, then a test-retest approach to quality score validation can be used. This involves the researcher redoing the assessment of selected studies after a time delay. Alternatively, PhD students can ask members of their supervisory team to assess a random sample of the primary studies. Whether the assessment has been carried out by independent researchers, or where

a lone researcher has taken a test-retest approach, the level of agreement between the scores can be checked (for example, using a Kappa analysis).

Using quality assessment results

As indicated in Section 7.1, results from the quality assessment process can be used within a systematic review in a number of ways. These include:

- specific quality criteria or the overall score can be used to exclude studies that are considered to be of low quality,
- analyses can be performed with and without low quality studies to determine the impact of such studies on the overall results,
- one of the research questions addressed by a review may focus on trends in the quality of primary studies relating to the topic of a review.

Whatever the role played by quality assessment, reviewers will need to consider the study type as well as the quality score for each of the primary studies that contribute to the findings of a review.

7.4 Examples of quality assessment criteria and procedures

Here we summarise three examples of quality assessment undertaken as part of software engineering systematic reviews. These cover each of the three types of systematic review: quantitative technology-focused reviews; qualitative technology-focused reviews and qualitative research-focused reviews (see Section 3.1 and Figure 3.1).

Quality assessment performed by researchers undertaking tertiary studies is also briefly highlighted.

The first example is a quantitative systematic review by Kitchenham et al. (2007) of studies which compare the use of cross-company and within-company cost estimation models. This review uses the checklist shown in Table 7.4 which is split into two parts (Part I and Part II). The criteria in Part I relate to the quality of the primary study and those in Part II are about the quality of reporting. The parts are weighted differently, with Part I having a weighting of 1.5 and Part II having a weighting of 1. The table indicates the possible scores for each of the criteria.

Quality assessment was carried out in parallel with data extraction in the following way:

1. For each paper, a reviewer was nominated randomly as data extractor/quality assessor, data checker or adjudicator,

TABLE 7.4: A Quality Checklist for a Quantitative Systematic Review (Taken from Kitchenham, B. A., Mendes E.& Travassos G. H. (2007)). Reproduced with permission.

	Criteria
Part I	
1.	Is the data analysis process appropriate?
1.1	Was the data investigated to identify outliers and to assess distributional properties before analysis? Yes(0.5)/No(0)
1.2	Was the result of the investigation used appropriately to transform the data and select appropriate data points? Yes(0.5)//No(0)
2.	Did studies carry out a sensitivity or residual analysis?
2.1	Were the resulting estimation models subject to sensitivity or residual analysis? Yes(0.5)/No(0)
2.2	Was the result of the sensitivity or residual analysis used to remove abnormal data points if necessary? Yes(0.5)/No(0)
3.	Were accurate statistics based on the raw data scale? Yes(1)/No(0)
4.	How good was the study comparison method?
4.1	Was the single company selected at random (not selected for convenience) from several different companies? Yes(0.5)/No(0)
4.2	Was the comparison based on an independent hold out sample (0.5), random subsets (0.33), leave-one-out (0.17) or no hold out (0)?
5.	Size of within-company dataset? fewer than 10 projects (score 0), 10-20 (0.33), 21-40 (0.67), more than 40 (1)
Part II	
1.	Is it clear what projects used to construct each model? Yes(1)/No(0)
2.	Is it clear how accuracy was measured? Yes(1)/No(0)
3.	Is a clear what cross-validation method was used? Yes(1)/No(0)
4.	Were all model construction methods fully-defined (tools and methods used)? Yes(1)/No(0)

2. The data extractor/quality assessor read the paper and completed a form,

3. The checker read the paper and checked the form,

4. If the extractor and checker could not resolve any differences that arose, the adjudicator read the paper and made the final decision after consulting the extractor and checker.

The assignment of roles was constrained so that no-one performed data extraction or quality assessment for a paper that they had authored and as far as possible the work load was shared equally.

In the second example, Dybå & Dingsøyr (2008*a*) report a qualitative systematic review of studies relating to Agile software development. The reviewers used the criteria shown in Figure 7.3 and formulated quite detailed descriptions of the issues to consider when scoring studies against each of the criteria. Studies were scored using a simple yes/no scale. The detailed descriptions of

issues used to guide the scoring process can be found in Appendix B of Dybå & Dingsøyr (2008*a*).

Dybå & Dingsøyr and another researcher used the first criterion ('Is the paper based on research (or is it merely a 'lessons learned' report based on expert opinion)?') as the basis for inclusion/exclusion and they calculated their level of agreement for this criterion (94.4%). Disagreements were resolved by discussion among the three researchers.

The third example is a systematic review that addresses a research process, specifically the systematic review process (Kitchenham & Brereton 2013). This review included primary studies of many different types such as case studies, surveys and secondary studies. It also included discussion and 'lessons learned' papers. The reviewers chose to base quality assessment on the generic checklist developed by Dybå & Dingsøyr (2008*a*) (see Figure 7.3) with the additional question:

> "What research method was used: Experiment, Quasi-Experiment, Lessons learnt, Case Study, Opinion Survey, Tertiary Study, Other (Specify)?"

The determination of study type was based on the reviewers' own assessments rather than on the type claimed by the authors of a paper. Checklist items 5–8 were also adapted to address the different study types. The revised checklist items were:

> Item 5. "Was the recruitment strategy (for human-based experiments and quasi-experiments) or experimental material or context (for lessons learnt) appropriate to the aims of the research?"
>
> Item 6. "For empirical studies (apart for lessons learnt) was there a control group or baseline with which to evaluate systematic review procedures?"
>
> Item 7. "For empirical studies (apart for lessons learnt) was the data collected in a way that addressed the research issue?"
>
> Item 8. "For empirical studies (apart for lessons learnt) was the data analysis sufficiently rigorous?"

In addition, an allowable 'score' of 'not applicable' was included for questions 4-8. For most of the criteria, the allowable scores for applicable items were Yes (1), Partly (0.5), No (0) with interpolation permitted. The exception was the first criteria (Is the paper based on research?), for which only the scores of Yes(1) and No(0) were allowable.

The two reviewers undertook quality assessment (and data extraction) independently. Disagreements were discussed until agreement was reached. The reviewers noted some problems with their approach:

- Although they identified broadly which questions were relevant for particular types of study, they found that for some studies the context meant that further decisions about appropriateness had to be made during the quality assessment. This point is discussed in some detail in the paper and resulted in the reviewers assessing independently whether a question was relevant for a particular study as well as determining scores for each relevant criteria.

- Their assessments of study type frequently differed from those of the authors of a study. For example, if a case study was based on an opinion survey they classified it as an 'Opinion Survey' rather than a 'Case Study', and if a study was a post-hoc analysis of a systematic review they classified it as an 'Example' rather than a 'Case Study'.

- They found that using the checklist sometimes resulted in small studies obtaining good scores even though by their nature they could provide only very limited evidence of the value of the technique or method being studied. For example, if a study was a preliminary feasibility study it could score well on all checklist items even though it could provide very limited evidence of real value of the method being studied. Additionally, some lessons learned and experience papers scored well because relatively few checklist questions were relevant.

The level of agreement achieved for quality assessment, using values for the number of questions considered to be appropriate and the average quality score for each paper, was measured using the Pearson correlation coefficient.

As indicated in Section 7.2.1, a number of tertiary reviews have used the DARE criteria for assessing study quality. See for example, Kitchenham, Brereton, Budgen, Turner, Bailey & Linkman (2009), Kitchenham, Pretorius, Budgen, Brereton, Turner, Niazi & Linkman (2010), da Silva et al. (2011), Cruzes & Dybå (2011*b*) and Verner, Brereton, Kitchenham, Turner & Niazi (2014). With the exception of Cruzes & Dybå, reviewers scored each primary study (that is, each systematic review or mapping study) against each of the criteria with possible scores being Yes(1.0), Party (0.5) and No(0). Cruzes & Dybå scored studies as either meeting a criterion (Yes) or not (No). A range of different approaches was taken to allocating independent quality assessors and to resolving disagreements.