Chapter 6

Study Selection

Once candidate papers have been identified through the search process, these need to be checked for relevance to the research questions being addressed by a review. The focus of this chapter is on this selection process which forms the second step of the conduct phase of the systematic review process, as highlighted in Figure 6.1.

Study selection is a multi-stage process which can overlap to some extent with the searching process. It is multi-stage because, ideally, many candidates that are clearly irrelevant can be quickly excluded, at an early stage, without the overheads of reading more than their titles and abstracts. In later stages candidate papers have to be read 'in full'. Study selection can overlap with the searching process when searching involves backwards snowballing or contacting authors of relevant papers (or studies). In this situation, relevance needs to be established before these searching methods are used.

In this chapter we discuss three aspects of study selection:

- The selection criteria,
- The selection process,
- The relationship between papers and studies.

The chapter concludes with some examples of the selection criteria used and the procedures followed by software engineering reviewers.

6.1 Selection criteria

The criteria for selecting studies to include in a review are formulated in order to identify those studies that are able to provide evidence that is of

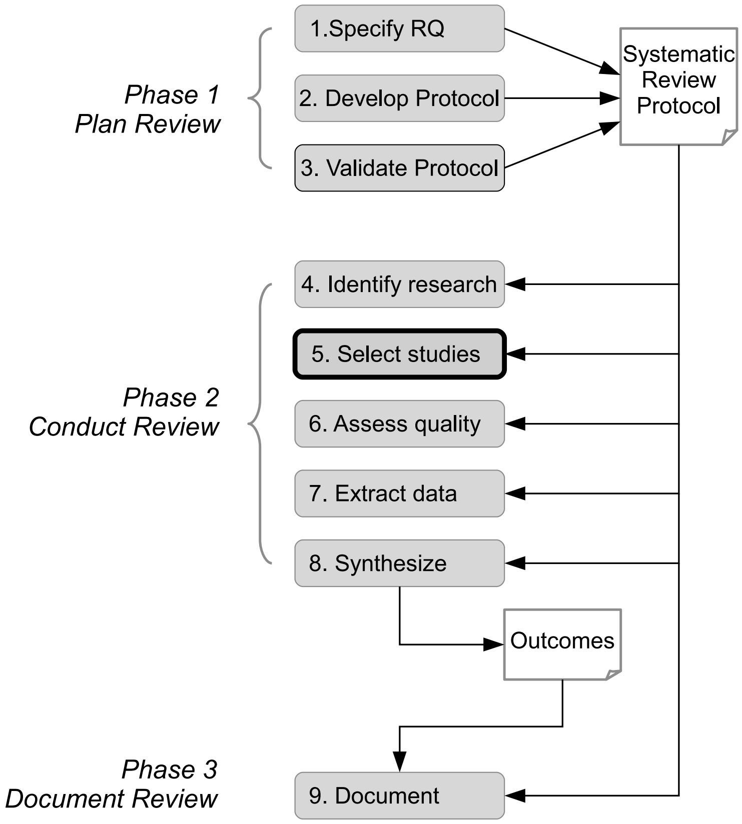

FIGURE 6.1: Study selection stage of the systematic review process.

relevance to the research questions. The criteria are generally (although not universally) expressed as two sets: one for the inclusion criteria and one for the exclusion criteria.

Some selection criteria are quite generic and fairly easily interpreted. For example, criteria relating to publication date are reasonably straightforward, although even this can be complicated by:

- The practice followed by some publishers of providing online access to draft papers before they are incorporated into a specific issue of a journal,

- Some studies being reported in more than one paper, particularly if not all of the papers fall within the scope (especially the time period) of a review.

It is also often the case that studies are included only if they are published in English and in ‘full’ peer-reviewed papers (as opposed, for example, to being reported in extended abstracts or in ‘grey literature’ such as technical reports and PhD theses).

For the more technical elements of a review, scoping the literature can be quite challenging and sometimes the criteria need to be revised as reviewers become more familiar with the topic and its boundaries. A point to note here is that it is important to be explicit about the scope of a review in any resulting publications, by fully reporting:

- The criteria used in the selection process,
- Details of the papers that are included in the review,
- The rationale for excluding marginal or 'near-miss' papers.

When planning a review, the study selection criteria can be piloted to ensure that they can be sensibly and consistently interpreted by members of a review team and that for the known papers they lead to the desired outcome. Even then, some refinement of the criteria (and hence of the protocol) may be needed as a review progresses.

6.2 Selection process

Study selection is usually carried out in a number of stages. Initially, once a set of candidate papers has been identified, those that are clearly irrelevant can be excluded on the basis of their title or their title and abstract. After this early screening, papers have to be looked at in more detail. For example, reviewers might decide to exclude a paper after reading some of the sections (such as the introduction, a methods section or the conclusions), however, the likelihood is that many papers will have to be read 'in full' before the decision to exclude (or not) can be made. Sometimes, the decision to include a study is overturned later in the review process. This may arise, for example, if the required data cannot be extracted or if a study fails to reach a quality threshold. In the end, there may well be marginal studies and the best that reviewers can do is report and explain their decisions in such cases.

Reviewers may find that they have a very large number of candidate papers (and what constitutes 'very large' will depend to some extent on the size of the review team). Possible strategies for dealing with this problem are:

- Refining the search strings to improve recall and precision,
- Reducing the scope of the review (through refinement of research questions),
- Use of a text mining tool to support the selection process,
- Increasing the size of the review team.

Also, if the selection process results in a large number of papers being included in a review then reviewers may choose to complete the process using only a sample of the papers.

Where study selection is being performed by a team of reviewers, there is the opportunity to validate the outcomes of the selection process by two (or more) members of the team independently applying the inclusion/exclusion criteria and checking their level of agreement (for example, by performing a *kappa* analysis (Cohen 1960)). Although tools are available to calculate the kappa coefficient, this is briefly explained below. As well as calculating the level of agreement, a mechanism is needed for resolving any differences that arise. Common approaches are to do this through discussion or by using a third reviewer to act as mediator.

A kappa (κ) coefficient is calculated using the following equation[1]:

$$\kappa = \frac{\text{actual agreement } - \text{ agreement expected by chance}}{\text{scope for doing better than by chance}} \tag{6.1}$$

TABLE 6.1: Example Data for Study Selection by Two Reviewers

		Reviewer B	Reviewer B	
		Included	Excluded	Total
Reviewer A	Included	10	3	13
Reviewer A	Excluded	4	25	29
	Total	14	28	42

Consider the data shown in Table 6.1. Reviewer B has classified 14 of 42 studies as ‘included’ while Reviewer A has included 13 of the 42 studies. The number of studies for which there is actual agreement is 10 plus 25 giving a total of 35 out of 42 which equals 0.8333 (83.33%) of the studies. By chance alone, the probability of an ‘include’ from Reviewer A is 13/42 = 0.3095 and for Reviewer B is 14/42 = 0.3333. The chances of agreement by chance are these two probabilities multiplied together, that is, 0.3095 x 0.3333 = 0.1032. Using a similar calculation, the chances of agreement to exclude by chance is 0.4604. Adding together these two probabilities of agreement by chance gives a total of 0.5636. That is, 56.36% agreement would be expected by chance. This gives a kappa score as shown below:

$$\kappa = (0.8333 - 0.5636)/(1 - 0.5635) = 0.618 \tag{6.2}$$

Kappa scores are generally interpreted as shown in Table 6.2. We see, therefore, in our example, that agreement between the two reviewers is *Good/Substantial*.

One approach to sharing the workload associated with study selection is for the lead reviewer to perform the early screening stage(s), excluding papers on the basis of titles or on titles and abstracts, which is usually quite straightforward, with the later, more difficult, stages being performed independently by two members of the review team. See, for example, the process followed

[1] Further details can be found at: http://www.ganfyd.org/index.php?title=Statistical tests for agreement

TABLE 6.2: Interpretation of Kappa

Value of kappa	Strength of agreement
0 - 0.29	Poor
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Good/Substantial
0.81 - 1.00	Very good/Almost perfect

by Marshall & Brereton (2013), described in the next section, which adopted this approach.

For PhD students, it is not always possible for selection to be performed independently by two reviewers. Where this is the case there are a number of ways that confidence in the decisions made can be enhanced. For example, a member of the supervisory team can check a random sample of papers (or those papers that are considered marginal or about which the student is uncertain). Alternatively, PhD students or other lone researchers can use a test-retest approach which entails repeating (after a suitable time delay) some or all of the study selection actions and comparing the outcomes. For each of these approaches to study selection validation, if agreement is good, then the review can proceed with some confidence in its reliability, if it is not, then the criteria and their interpretation need to be reconsidered.

Another means of checking the decisions made (whether by one or by multiple reviewers) is to carry out some form of text analysis (also referred to as text mining) to help determine whether papers that are 'similar' in some way have been either all included or all excluded during the study selection process. The general approach is to use a text mining tool to identify and count the frequency of important words or phrases in each paper. A visual display tool can then be used to show clustering with respect to these, highlighting where papers in the same cluster (that is, papers that seem to be 'similar') have been treated differently in the selection process. A number of small studies have demonstrated the feasibility of using text mining to support study selection (Felizardo, Andery, Paulovich, Minghim & Maldonado 2012). Some text mining and visualisation tools that have been used to support the systematic review process are listed in Chapter 13.

6.3 The relationship between papers and studies

The relationship between research papers (or other dissemination forms) and the studies that they report is important for systematic reviews. Researchers undertaking systematic reviews or mapping studies are usually (although not exclusively) looking for empirical studies that provide some sort of evidence about a topic of interest. They will find, however, that

- Papers can report more than one study,
- Studies can be reported in more than one paper.

Where a paper reports multiple studies, these can generally be considered as separate studies for the purposes of a systematic review. The study selection process may result in some of the studies being included in a review and some being excluded. Although this seems quite straightforward, this is not always the case. Sometimes, one or more studies are preliminary or pilot studies undertaken in advance of the 'main' study. Also, sometimes, several case studies are reported which could be treated separately or as a single multi-case study. These issues are discussed further in Part III, Section 22.6.4.

A study may be reported in more than one paper. This is not unusual in software engineering. A conference paper may be followed by a more detailed or enhanced journal paper. Also, a large study may be reported in many papers which focus on different aspects of the research. It is important that such multiple publications of a (single) study are identified, so that the results are not counted more than once, and, where the quality of the study is being assessed, all of the published information about the study can be used for making that assessment. It is not always straightforward to establish that multiple publications report a single study. Of course there may be some cross-citation and it may be that titles and author sets are similar across a set of papers. In the absence of these fairly obvious indicators, reviewers should pay particular attention to sets of papers where the same number of participants are recorded for 'similar' studies reported by similar sets of authors. Again this issue is discussed further in Part III, Section 22.6.4.

6.4 Examples of selection criteria and process

In this section we describe the criteria used and the processes followed for some of the published reviews. We can see in these examples that quite a wide range of approaches to applying the criteria is taken. Sometimes, however, only limited information about the process and specifically about the roles taken by members of a review team is available in published papers.

Examples of study selection for quantitative systematic reviews

We look at two reviews in this category. They compare:

- Two approaches to software effort estimation (MacDonell & Shepperd 2007),
- Two development life cycle models (Mitchell & Seaman 2009).

The review by MacDonell & Shepperd (2007) compares the effectiveness of software effort estimation models that use within-company (that is, local) data with models that use cross-company (that is, global) data. The reviewers only included studies where the experimental design met the following (inclusion) criteria:

- Data was from five or more projects per company and for at least two companies,
- There was a comparison between within-company and cross-company models,
- The projects covered were substantially software projects (that is, not hardware or co-design),
- Projects were commercial (that is, not student projects),
- Publications were demonstrably peer-reviewed, written in English and published between 1995 and 2005.

Abstracts of all papers retrieved by the search process were read by the reviewer who had performed the search to determine whether the paper should be included. If the decision could not be made, the reviewer read the whole paper and then applied the inclusion/exclusion criteria. The second reviewer provided comments on a small number of borderline papers.

Mitchell & Seaman (2009) performed a review of studies that compare the cost, development duration and quality of software produced using a traditional waterfall approach with those of software produced using iterative and incremental development (IID). Their search process found 30 candidate papers, nine pairs of which were found to be duplicates, leaving 21 unique papers. At this stage, the reviewers applied the following inclusion criteria, requiring that papers should:

- Be written in English,
- Be peer-reviewed,
- Report a primary study,
- Report empirical results,
- Compare waterfall and IID processes,
- Present results concerning development cost and/or development duration and/or resulting product quality.

This process reduced the number of candidates to 11. The subsequent identification of duplicate reports of the same study, the realisation that the waterfall process for one of the studies included iteration, plus the application of a quality threshold reduced the final count of studies to five. It is interesting to note here, that the first two of these additional 'criteria' (relating to duplicate reports and details about the processes being compared) are essentially exclusion criteria although in the paper they are not labelled in this way. There is no indication in the paper about whether both or only one of the authors performed the study selection.

Examples of study selection for qualitative systematic reviews

Here we summarise the criteria and the process for study selection in a management focused review relating to motivation in software engineering (Beecham et al. 2008) and in a research-oriented review of studies about the systematic review process (Kitchenham & Brereton 2013).

The study by Beecham et al. (2008) reviewed knowledge about what motivates developers, what de-motivates them and how existing models address motivation. Before the authors applied the inclusion and exclusion criteria, they checked for duplicate publications of individual studies and only included one of the reports (either the most comprehensive or the most recent). The reviewers stated that they included 'texts' that:

- Directly answer one or more of the research questions,
- Were published from 1980 to June 2006,
- Relate to any practitioner directly producing software,
- Focus on de-motivation as well as motivation,
- Use students to study motivation to develop software,
- Focus on culture (in terms of different countries and different software environments),
- Focus on 'satisfaction' in software engineering.

They excluded texts:

- In the form of books or presentations,
- Relating to cognitive behaviour,
- Not relating to software engineering,
- Focusing on company structures and hierarchies unless expressly linked to motivations,

- In the form of opinion pieces, viewpoints or purely anecdotal,
- That focus on software managers (who do not develop software),on group dynamics or on gender differences.

Beecham et al. retrieved over 2000 references through the search process and eliminated approximately 1500 of these on the basis of titles and abstracts. This left 519 papers. These (except for 9 papers which could not be obtained) were looked at in full by 'a group of primary researchers' who accepted 95 papers. An independent researcher looked at 58 of the 519 papers, which were randomly selected by taking (approximately) every 10th paper from an alphabetic list, and re-applied the inclusion/exclusion criteria. The inter-rater reliability was 99.4% indicating a high level of agreement and giving confidence in the decisions made. A further validation exercise was carried out on the 95 included papers by an independent expert on motivation in software engineering who checked how each paper addressed the research questions. There was a high level of agreement (99.8%) and the three papers where the decision differed were considered by a third independent researcher. Once the disagreements were resolved, 92 papers remained in the set of included papers.

Kitchenham & Brereton (2013) performed a qualitative systematic review to identify and analysis research about using and improving the systematic review process. As well as reporting selection criteria, these researchers also explain the rationale for each criterion. For conciseness, the rationale is omitted from the following descriptions of the criteria used and the process followed. The inclusion criteria used were:

- the main objective of the paper is to discuss or investigate a methodological issue relating to systematic reviews.
- The paper addresses the construction and/or evaluation of quality instruments,
- There must be a software engineering context,
- The paper must be written in English,
- The paper may be a short paper.

Papers were excluded if:

- Their main objective was to report a systematic review or mapping study,
- They discussed evidence-based software engineering principles,
- They were methodological studies with a general (that is, a non-software engineering) focus,
- In form of PowerPoint presentations or extended abstracts,

- They produced guidelines for performing or reporting primary studies.

The search strategy for this review involved an initial informal search followed by a 3-stage process which included both a manual search and an automated search. Here we summarise the selection aspects of the search and selection process.

Stage 1 A manual search was performed by both authors who independently applied the inclusion and exclusion criteria, with an emphasis on inclusion unless a paper was clearly irrelevant. Disagreements were discussed and where agreement was not reached, the paper was included. Following an automated search, both reviewers applied the selection criteria, using the title and abstract of the papers found. Again the main emphasis was on including papers unless they were clearly irrelevant. Disagreements were discussed and where agreement could not be reached, the papers were provisionally included.

Stage 2 Papers included from the manual search, from the automated search and from the known set (determined through the informal search and using personal knowledge) were collated into a set of candidate papers. Where papers were treated differently across these inclusion sets, they were discussed and if no decision could be reached the paper remained a candidate. The final inclusion/exclusion decisions were made when the full papers were read during data extraction and quality assessment; again disagreements were discussed until agreement was reached.

Stage 3 At this stage, additional searching methods were used (snowballing and approaching individual researchers), and search and selection validation was carried out. Validation was based on the kappa agreement achieved between the authors for the decisions made during manual selection and for the selection from the candidates identified by the automated search.

Examples of study selection for mapping studies

The following examples report details of the processes followed as well as the criteria used. These mapping studies aimed to:

- Find out how extensively, and by what means, the Gang of Four (GoF) design patterns have been evaluated (Zhang & Budgen 2012),
- Identify and classify tools to support the systematic review process (Marshall & Brereton 2013).

Zhang & Budgen (2012) aimed to identify which Gang of Four (GoF) design patterns had been evaluated, what lessons had been learned from the evaluation studies and what further research might be needed to address ‘gaps’ in the evidence. The reviewers applied the following inclusion criteria:

- Papers describe software design patterns, although only empirical papers were used for the analysis,

- If several papers report the same study only the most comprehensive would be included,
- Where several studies were reported in a paper, each study would be treated independently.

Studies were excluded if they were:

- Reported in the form of abstracts or PowerPoint presentations,
- Documented in technical reports or papers submitted for publication.

The authors followed a 3-step process for excluding irrelevant papers or studies.

1. exclude on the basis of title,
2. exclude after reading the abstract,
3. exclude after reading the full paper.

The authors performed each of these steps independently, and then produced an agreed-upon list. They took a conservative approach for steps 1 and 2. A kappa score was calculated for an inter-rater agreement for each step.

The study by Marshall & Brereton (2013), which looked at the use of tools to support systematic reviews and mapping studies, included papers that:

- Report on a tool to support any stage of a systematic review or mapping study in software engineering,
- Report on a tool that is at any stage of development (e.g. proposal, prototype, functional, etc.)

Exclusion criteria were:

- Papers not written in English,
- Abstract or PowerPoint presentations.

The inclusion and exclusion criteria were applied in two stages. Initially, the first author checked the titles and abstracts of candidate papers and those that were clearly not relevant were excluded. After this stage, 21 papers were included. In the second stage, both authors checked the full texts, resulting in 16 papers remaining in the inclusion set. Subsequently two further papers were excluded during data extraction.