Chapter 22

Systematic Review and Mapping Study Procedures

22.1 Introduction

This rather long chapter is a revision of our previous guidelines for systematic reviews for software engineering research (Kitchenham & Charters 2007). There are deliberate overlaps with information provided in the preceding chapters of this book, so that this chapter can be used as a self-standing set of guidelines. However, we do cross-reference sections of the book where more detailed information about specific topics can be found.

Compared with the previous version of the guidelines, we have included more detailed advice for mapping studies and for the procedures needed by lone researchers including PhD students. We also include more guidelines based on our experiences of performing systematic reviews and the experiences of other software engineering researchers (Kitchenham & Brereton 2013).

Diagrams used in this chapter follow one of two standards:

1. When simple yes/no decisions are involved, we use a simple flowchart with the decisions shown as diamonds and actions shown in rectangles. Links (lines) between diamonds and rectangles have arrows which show the direction of flow, see Figure 22.1.

2. If decisions are more complex, as is often the case when planning a part of the systematic review process, we specify the decision using a rectangle with double lines. If decisions are inherently sequential, they are linked with lines using arrows to show the direction of flow. Simple rectangles are used to specify factors that affect a decision, or options available for a decision, or factors that influence the choice of options. Factors that influence a decision or an option are linked to the respective decision or option with a broken line with arrows at each end. Decisions are linked to options using a line with an arrow pointing at the option. An example of this form of diagram is shown in Figure 22.2.

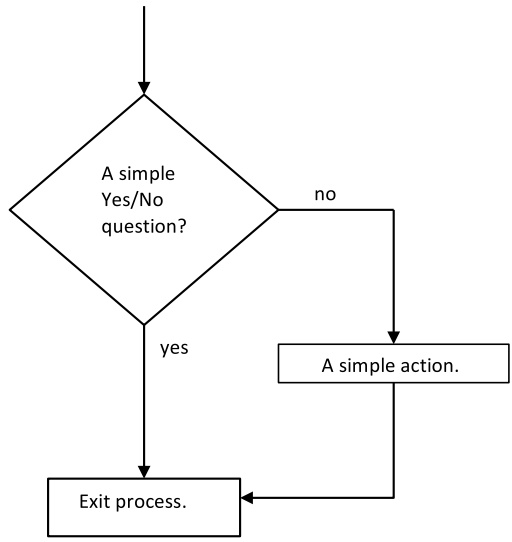

FIGURE 22.1: A simple flowchart.

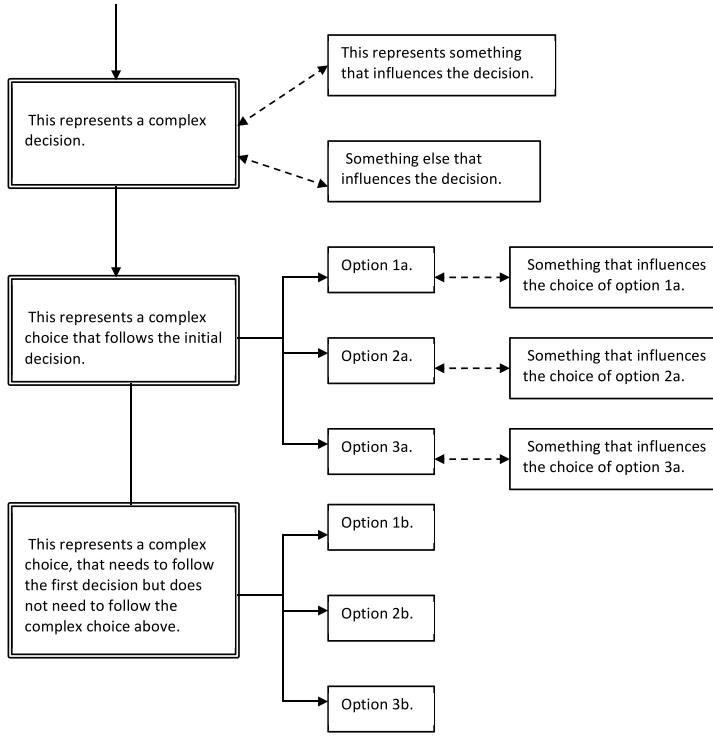

FIGURE 22.2: A complex planning process diagram.

22.2 Preliminaries

Before starting on a systematic review or mapping study, consider whether you have adequate background knowledge of the proposed topic area to be able to make decisions about the various choices involved. If not, you should begin by first reading about the topic (see Figure 22.3).

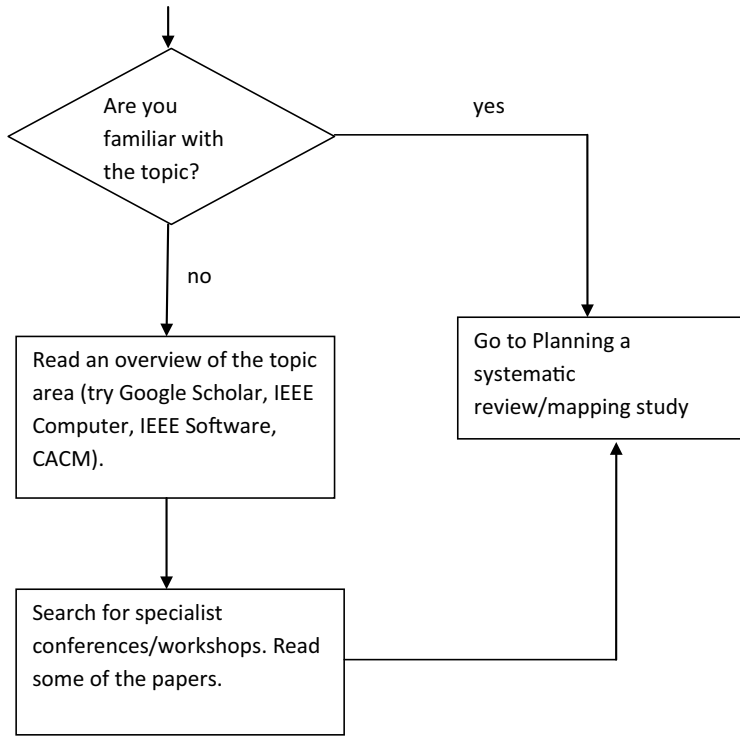

FIGURE 22.3: Initial considerations.

Point to remember

If you don't have any knowledge about the topic area, do not start planning your systematic review or mapping study yet. You need to read around the topic before you start.

22.3 Review management

A review is usually conducted by two or more researchers who comprise the review team. One researcher must act as the review manager or team leader in order to ensure all task activities are properly coordinated.

In the context of developing the protocol, the team leader is responsible for:

- Producing the protocol.
- Specifying the time scales for the review.
- Assigning the tasks specified in the protocol to named individuals.
- Obtaining any tools required to manage the review process and conduct individual tasks.
- Deciding how the protocol will be validated.
- Overseeing the protocol validation.
- Signing off the protocol and any subsequent changes to the protocol.

During the conduct of the review the team leader is responsible for monitoring the review progress, ensuring that team members complete their assigned tasks and managing any contingencies that arise during the review (such as disagreements about such aspects as primary study selection, quality evaluation, and data extraction).

Once the review is complete, the team leader is also responsible for signing off the final report.

Point to remember

The more researchers there are in a team, the more critical the role of the team leader becomes.

22.4 Planning a systematic review

Planning involves four main processes:

1. Justifying the need for a systematic review or mapping study.
2. Specifying the research questions.
3. Developing the protocol.
4. Validating the protocol

However, since reviews are usually done by a research team, planning also involves undertaking the basic project management actions such as task assignment, review coordination and monitoring, as discussed in the previous section.

22.4.1 The need for a systematic review or mapping study

An overview of the process of justifying the need for a systematic review is shown in Figure 22.4.

You should begin by checking whether any systematic reviews or mapping studies already exist in the topic area you want to study. If there are some, you may not need to do a review. Don't forget it is correct to use other researchers' work as the foundations for your own research. A major goal of systematic reviews in general, and mapping studies in particular, is to facilitate future research in a specific topic area.

However, if the existing reviews(s) do not cover the specific area you are interested in, or those that do exist are out of date, continue with planning the review. If there are already some existing secondary studies (that is, any literature reviews or state of the art surveys, whether systematic or not), you need to read these studies and decide whether you can use their results *as-is*, or you need to update one the review(s), or whether you need to undertake a new more focussed systematic review. In the case of an out of date systematic review, a sensible choice is to base your study on the protocol used by the initial study, while also amending the process if you can identify any limitations with the initial protocol.

If you decide to undertake a new more focussed review, the existing review(s) will have a significant impact in your research:

- You should read the papers describing the reviews and summarize their results. This will be the basis of the "related work" section in your final report. It should also allow you to specify the baseline of existing knowledge about the topic area and explain clearly how your results add to existing knowledge in the discussion part of your final report.

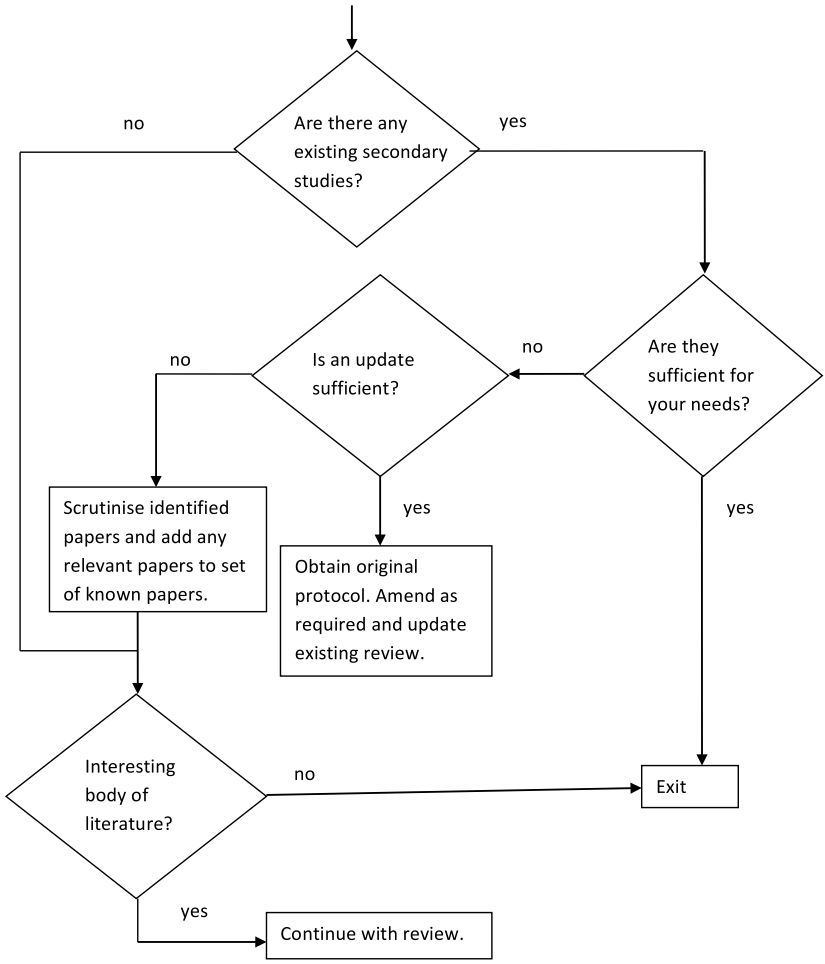

FIGURE 22.4: Justification for a systematic review.

- You should extract a list of any primary studies found by the previous reviews that are relevant to your topic area. These will be the basis for your set of known papers that can be used to construct and refine search strings for digital libraries and to help assess the completeness of your search (see Figure 22.5.1.3).

If there are no previous reviews, you need to be sure that there are likely to be sufficient relevant papers to make a systematic review or mapping study worthwhile. One way is to undertake a quick informal search using Google Scholar or a digital indexing system to look for relevant studies. In some cases a limited form of mapping study, called a scoping review, can be performed

to determine whether there are sufficient empirical primary studies to justify a systematic review.

Finally, justifying the need for a review is about making a case that the topic is of interest and that it is an appropriate time to engage in an activities aimed at organising the literature or answering questions raised by the literature. For example, the reasons for performing a systematic review include:

1. There is a new development or testing method and practitioners would like to know if it is better than existing methods.
2. There are disagreements among researchers about the efficacy of a new method and the current empirical evidence needs to be collated.
3. There are field reports about a new software development method, or an international standard, and practitioners would like to understand what is known about the method or standard in terms of benefits and risks.

Reasons for doing mapping studies include:

1. To help assess the extent of research available in a topic area in order to identify sub-areas suitable for systematic reviews and sub-areas that need more basic research.
2. To organise a large number of independent research papers into a structured body of knowledge.

Points to remember

1. You need a genuine reason for undertaking a review, such as the likely existence of a large number of independent studies that have not previously been organised.
2. The existence of a substantial body of literature is not by itself a justification for a review. The topic for the review needs to be important to researchers and/or practitioners and the review needs to be timely.
3. Previous literature reviews (systematic or not) are extremely valuable for identifying known primary studies and validating your search process.

22.4.2 Specifying research questions

Research questions (RQs) are related to the justification for doing the review and the type of review being proposed.

22.4.2.1 Research questions for systematic reviews

If you are concerned with evaluating a technology, you should be planning to do a systematic review and the research questions should specify the type of evaluation you propose. For systematic reviews, the research question defines much of the search process. It is important to make sure the research questions(s) are properly formulated and stable, since changes to the RQ(s) will propagate other changes throughout the systematic review protocol.

If you are comparing two alternative techniques your research question will be of the form: Is technique A better than technique B? This basic question may need to be refined to determine what is meant by *better*, for example, more cost effective. Furthermore, you may want to qualify any answer in terms of any limitations or constraints on the answer, for example, does the answer apply to students or professionals, or to particular types of tasks, leading to questions of the type:

> Under what conditions, if any, is technique A more cost effective than technique B?

In software engineering, many empirical studies consider the impact and effectiveness of paradigm, method or standard A in an industry setting. Systematic reviews of such studies have research questions of the type:

- What are the risks or benefits associated with adopting paradigm, method or standard A?
- What factors motivate or de-motivate adoption of paradigm, method or standard A?
- How best should an organisation plan the adoption of paradigm, method or standard A?

22.4.2.2 Research questions for mapping studies

For mapping studies, research questions are often quite high level. This is because the characteristics of interest in the specific topic area may be hard to specify in advance. Thus, there is more likelihood of research questions being amended as result of identifying interesting aspects of the topic during data extraction.

Mapping studies usually have the overall goal of categorising the research literature for a specific topic in some way. This leads to research goals of the type: What trends can be observed among research studies discussing topic B?. The problem with a mapping study is deciding what trends will be of interest.

In practice, most software engineering mapping studies consider issues such as:

- The number of publications per year over the time period of the review, which gives an indication of the interest in the topic.
- The number of papers reporting studies of different types, often using the requirements engineering classification developed by Wieringa et al. (2006), which indicates the type of research being undertaken.
- The main researchers and research groups which identifies groups that interested researchers or practitioners might want to keep up with.
- The sources which published papers on the topic which identifies sources interested researchers or practitioners might want to monitor for future research.

This may be sufficient for the purposes of a student mapping study but is unlikely to be sufficient for a conference or a journal publication.

Mapping studies are far more interesting and beneficial to other researchers (and more likely to be published) if they also identify interesting subsets of the literature for example, the main subtopics, the different approaches/methods reported in the topic area and the extent to which they have been evaluated empirically, any significant limitations in existing research, as well as any significant controversies.

Points to remember

- For systematic reviews, research questions need to be well-defined and agreed to before the protocol is developed.
- For mapping studies, research questions are usually fairly high level and may be refined as the mapping study progresses.
- Student mapping studies are not always suitable for publication.

22.4.3 Developing the protocol

The research protocol defines and justifies what technical processes will be used to conduct and report the review and identifies which individuals will be assigned to which tasks. A template for a systematic review protocol is shown in Figure 22.5. The main technical issues that have not already been considered (that is, points 3 to 9 of Figure 22.5) will be discussed in later sections.

Search strings, quality extraction, data extraction, and data synthesis procedures need to be trialled as the protocol is developed.

Points to remember

- Sections of the protocol need to be tried out to ensure that the process is feasible and understood by all.
- The team leader is responsible for developing the protocol although some aspects can be delegated to other team members.

22.4.4 Validating the protocol

The protocol is a critical element of any systematic review. Researchers must agree to a procedure for validating the protocol. Where possible, you should try to find an independent reviewer.

Research teams should walk through the protocol and ensure that each researcher understands exactly what tasks he or she is scheduled to perform and the process he or she needs to follow to perform their allocated tasks. PhD students should present their protocol to their supervisors for review and criticism.

Since a systematic review aims to address specific research questions, the protocol should explain how those questions will be answered. Thus, a reviewer of a protocol needs to confirm that:

- The search strings are appropriately derived from the research questions.
- The data to be extracted will properly address the research question(s).
- The data analysis procedure is appropriate to answer the research questions.

The systematic review team leader is responsible for coordinating all the changes to the draft protocol and the final decision that the protocol is sufficiently complete for the systematic review to formally get under way.

Template for a Systematic Review Protocol

1. Change Record

This should be a list or table summarizing the main updates and changes embodied in each version of the protocol and (where appropriate), the reasons for these.

2. Background

a) explain why there is a need for a study on this topic
b) specify the main research question being addressed by this study
c) specify any additional research questions that will be addressed
d) if extending previous research on the topic, explain why a new study is needed

3. Search Process

a) specify and justify basic strategy: manual search, automated search, or mixed
b) for automated searches, specify search terms and compounds of these and record results of any prototyping of the search strings
c) for automated searches, identify resources to be used (specifying the digital libraries and search engines)
d) for manual searches, identify the journals and conferences to be searched
e) specify the time period to be covered by the review and any reasons for your choice
f) identify any ancillary search procedures, for example, asking leading researchers or research groups, or accessing their web sites; or checking reference lists of primary studies
g) specify how the search process is to be evaluated (for example, against a known subset of papers; or against the results from a previous systematic review)

4. Primary Study Selection Process

a) identify the *inclusion* criteria for primary studies
b) identify the *exclusion* criteria
c) define how selection will be undertaken (roles of reviewers)
d) define how agreement among reviewers will be evaluated
e) define how any differences between reviewers will be resolved

5. Study Quality Assessment Process

a) specify the quality checklists to be used
b) specify how the checklist will be evaluated (if a new checklist has been developed)
c) define how agreement among data extractors will be evaluated
d) define how any differences between data extractors will be resolved
e) identify the procedures to use for applying the checklists, such as details inclusion/exclusion, partitioning the primary studies during aggregation or meta-analysis, and explaining the results of primary studies

6. Data Extraction Process

a) design data extraction form (and check via a dry run)
b) specify the strategy for extracting and recording the data (for example, paper form, on-line. Form or database)
c) identify how the data extraction process is to be undertaken and validated, particularly any data that require numerical calculations, or are subjective

7. Data Synthesis Process

a) specify the form of analysis/synthesis to be used (for example, narrative, tabulation, meta-analysis)
b) discuss how the synthesis will be validated

8. Study Limitations

a) assess the threats to validity (construct, internal, external), particularly constraints on the search process and deviations from standard practice
b) specify residual validity issues including potential conflicts of interest that are inherent in the context of the study, rather than arising from the plan

9. Reporting

a) identify target audience, relationship to other studies, planned publications, authors of the publications
b} agree in advance who will be included in the list of authors and whose assistance will reported in the acknowledgements section.

10. Schedule

Provide time estimates for all of the major steps.

FIGURE 22.5: Template for a systematic review protocol

Points to remember

- Protocols will change throughout the conduct of a study.
- The team leader should take responsibility for keeping the protocol up to date and the team notified of all changes.

22.5 The search process

Planning the search process begins by defining a search strategy. After deciding the basic scope of the search strategy, you will need to determine the specific sources that will be searched and the search strings that will be used for automated searches and the sources that will be searched manually.

The final element of the search process is to integrate the set of candidate primary study references, remove duplicate copies of the same paper found in different sources, and store the references in the agreed storage tool (which can be a reference manager system, a spreadsheet or a database).

22.5.1 The search strategy

The factors that influence your search strategy are shown in Figure 22.6. There are three main decisions:

1. To decide whether completeness is critical or not.
2. To decide how to validate your search process.
3. To decide on an appropriate mix of search methods.

22.5.1.1 Is completeness critical?

The first issue to be decided is whether completeness is critical or not. In the case of a systematic review comparing SE technologies, completeness is a critical issue. In the case of a mapping study looking at the high level research trends in a broad topic area, completeness might be less critical; however, having an unbiased search strategy remains crucial. Nonetheless, there are occasions where even a mapping study may have a requirement for completeness. In particular, the more detailed the topic area, the more likely it is that completeness will be important. If you are in doubt consider your research questions. Can they be answered adequately if some relevant papers are not found by your search process?

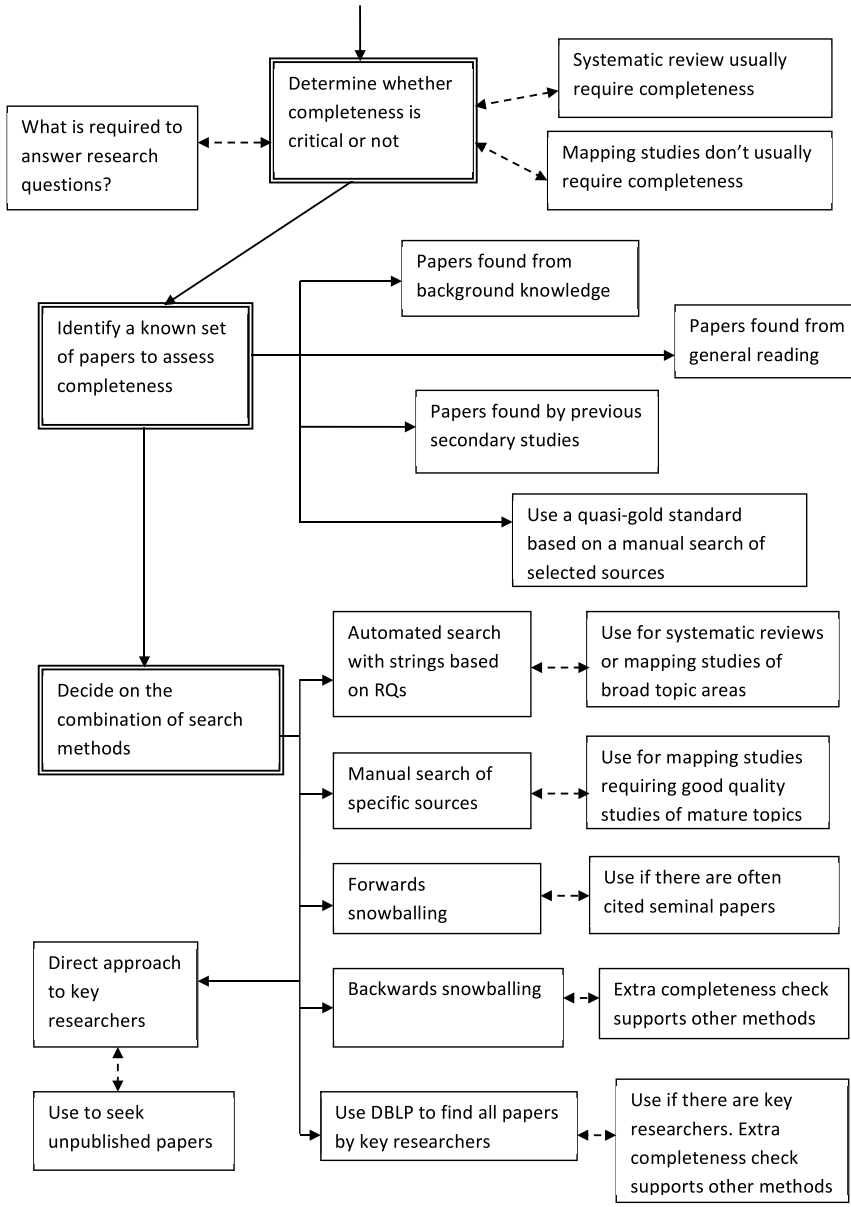

FIGURE 22.6: How to devise a search strategy.

22.5.1.2 Validating the search strategy

The next issue involves how you intend to refine and validate your search process. In the context of a systematic review or mapping study, validating

the search process means quantifying, in some sense, the level of completeness achieved. The best way of doing this is to compare the primary studies identified by your search process with a known set of studies. To obtain a known set of studies several processes are possible:

1. If you have done some preliminary reading, you should be able to identify a set of papers that ought to be included in your review.
2. If you (or one of your research team) is an expert in the topic area, you (or your expert colleague) should make a list of the papers that are already known to be relevant.
3. If there are other related literature reviews (systematic or not) in the same topic area, review the papers which were included by them, and identify those papers that should also be included in your review.
4. For mapping studies, if none of these options is possible or the number of known papers is insufficient, you will need to construct a quasi-gold standard as proposed by Zhang et al. (2011). A quasi-gold standard is explained in Chapter 5 and involves conducting a systematic manual search of several defined sources including important journals and specialist conferences to identify a set of primary studies that is treated as a set of known papers. Deciding how many known papers is sufficient is clearly a subjective assessment, and depends upon the expected number of primary studies (which is obviously unknown at the start of a review, but may be clarified as you try out some of your search procedures). We suggest that 10 papers or fewer are insufficient for an assessment of completeness for a mapping study, whereas 30 papers would be enough (since completeness is judged by the percentage of known papers found).

Option: If you have a large set of known papers, select half the papers at random to use for constructing and refining automated search strings, and put aside the other half of the papers to help measure completeness.

In the case of mapping studies you should set an acceptable level of completeness which should be larger than 80%.

In the case of a systematic review there are likely to be fewer available papers and you may only have one or two known papers, so a numerical measure of completeness may be inappropriate. In this case, your search process needs to be as stringent as possible (that is, covering all options, automated and manual), and completeness may only be assessed against all the elements of your process rather than a numerical figure attached to the outcome of the process.

If the search process does not reach the required level of completeness, you need to specify a contingency plan in the protocol. For example:

1. Adding other search methods such as backwards snowballing.
2. Refining your search strings until the required completeness level is obtained.

22.5.1.3 Deciding which search methods to use

Finally you need to specify the detailed processes you will adopt. Methods include:

- Automated searches of digital libraries using search strings derived from the research questions. An automated search is usually required if you are doing a systematic review (and require completeness) or performing a mapping study of a broad topic area.

- Manual search of a restricted set of sources (that is, specific journals and conference proceedings). A manual search process aimed at specific journals is likely to find good quality research papers on mature topics. A manual search of specialist conferences is usually needed as an auxiliary method for reviews of new Software Engineering topics.

- Backwards and forwards snowballing, that is, searches based on extracting information from reference lists (backwards snowballing) and citation information (forwards snowballing). Backwards snowballing is based on searching the citations in each candidate paper to look for additional candidate studies. It is an ancillary method which is mainly used to support string-based automated searches. Forwards snowballing is based on finding all the papers that have cited a specific paper and searching that list for candidate primary studies. It is particularly useful if there are one or two seminal papers that first introduced the topic and are therefore cited by most subsequent papers. Both types of snowballing are supported by general indexing systems such as Scopus and Web of Science.

- Direct approach to active researchers or searching DBLP Computer Science Bibliography[1] for papers published by a specific author. A direct approach to active researchers is usually an ancillary method and is used to find out whether there are any related studies that have not yet been published. If there are specific authors who are known to contribute to the topic area, you can use the DBLP database to list all papers by those authors. This can be used as a completeness check.

A good search strategy will use a combination of these methods, although in most cases, one option is selected as the main search method, and then supported by other method(s).

When checking the references found by primary studies (that is, doing backwards snowballing), there are two main approaches:

1. If you have relatively few primary studies (for example, < 10), you may decide to use a manual approach. Two researchers should read the "Introduction","Related Work" and "Discussion" sections of each paper

[1] http://dblp.uni-trier.de/db/

and identify candidate studies. The candidate studies are the union of the set of studies found for each paper by each researcher.

2. Another approach is to use a general indexing system such as Scopus. Find each selected research paper in turn and extract all the references for that paper. The set of candidate studies is the union of the references extracted from each research paper.

The manual approach results in fewer candidate studies, since some screening of references takes place when the papers are read. However, both approaches can have some difficulties identifying duplicate reports because authors do not report their references in the same format and some authors make mistakes in their citations (for example, putting in the wrong date or leaving out "The" or "A" in the title).

Points to remember

- Systematic reviews usually require completeness. Mapping studies usually don't.
- Have a set of known studies to help assess completeness.
- Specify an appropriate completeness level.
- Have a contingency plan if your search does not reach the required completeness level.
- You will almost certainly need to do an automated search either using search strings or using citation analysis.
- You will usually need to consider ancillary search processes to achieve required completeness levels.

22.5.2 Automated searches

There are two main decisions:

- Decide on the sources that will be searched.
- Specify the search strings that will be used (unless the search is to be based on snowballing of some sort).

22.5.2.1 Sources to search for an automated search

Appropriate sources include publisher specific sources and general indexing systems. A mix of the two types of sources is best. In particular, the IEEE Digital library and the ACM digital library together cover important international journals such as IEEE Transactions on Software Engineering and

most of the important computing-related conferences. These seem to be the best combination of publisher specific libraries. In addition, Springer publish a large number of conference proceedings and for new topics you may want to use SpringerLink as an additional source.

General indexing systems find many publisher specific sources (including ACM and IEEE papers) but may not index conference proceedings as quickly as ACM and IEEE. The Scopus, Web of Science and EI Compendex indexing systems are all possibilities and index papers published by Elsevier, Wiley and Springer which together with the IEEE publish most of the main internationally recognised software engineering journals that regularly publish empirical studies (which are of particular importance for systematic reviews). These include:

- *Empirical Software Engineering Journal* (Springer)
- *Journal of Systems and Software* (Elsevier)
- *Information and Software Technology* (Elsevier)
- *Software Quality Journal* (Springer).
- *Journal of Software Maintenance and Evolution: Research and Practice* (Wiley).

If completeness is critical, use several different indexing systems. The general indexing systems often provide mechanisms for extracting the references of papers they index and/or lists of papers that have cited a specific paper. These features are essential for efficient snowballing. It is possible to do backwards snowballing manually, although it is easier to extract references using facilities in an indexing system. It is not possible to perform citation analysis (forwards snowballing) without an automated system.

Standards in other domains emphasise the need to search for unpublished material such as Masters and PhD theses, technical reports, or industry "white papers". This is to ensure completeness of systematic reviews and minimise the possibility of publication bias (which occurs if negative results are less likely to be accepted by journals or conferences). It is unlikely to be necessary for mapping studies, but does need to be considered in the context of systematic reviews. Some researchers suggest using Google Scholar to search for unpublished searches but our experience of Google Scholar is that although it does identify unpublished material, it is often not possible to find a reliable source document that can be properly cited and guaranteed to remain publicly available. An alternative procedure is to approach key researchers directly and ask them if they have any relevant unpublished studies (including Masters or PhD theses or technical reports) that are publicly available.

22.5.2.2 Constructing search strings

PLEASE NOTE. *Previous versions of systematic review guidelines for software engineering researchers suggested using struc-*

tured questions to construct search strings. However, this approach has not proved to be very useful for software engineering reviews. Terminology in software engineering is neither well-defined nor stable, making it difficult to identify reliable keywords. Digital sources have limitations on the complexity of search strings and these are different for different libraries. Complex search strings are intended to identify a small number of highly relevant papers; however, in software engineering, they usually deliver large numbers of false positives.

Although in some rare cases a structured research question may help specify appropriate search strings, we recommend using fairly simple search strings based on the main topic of interest. Simple strings are more likely to work on a variety of different digital libraries without extensive refinement. To determine appropriate keywords:

- Review your research questions (RQs) and identify important concepts or terms used in the RQs.
- Review the terms used in the abstracts, keywords and title of your known set of papers. Match the frequently used terms to those found from your RQs.
- Try out your search strings on one of your selected digital indexing systems and identify the percentage of known papers you find. If the percentage of known papers found is low ($< 50\%$) (excluding, of course, any papers that could not have been found, such as papers not indexed by the specific digital indexing system or papers published in very recent conferences), review the papers that were not found. Refine your search strings by replacing existing keywords (for example, using more general terms) or adding new keywords (for example, adding qualifiers to make terms more specific).

If you are undertaking a systematic review aimed at aggregating comparative studies, you will need to specify some keywords to restrict your search to empirical studies, for example, "empirical" or "experiment".

Points to remember

- If automated searching is your main search strategy, you should search a variety of sources including IEEE, ACM and general indexing systems.
- General indexing systems have useful facilities for supporting the use of snowballing.
- Derive search strings from your research questions and terms used in known studies.

- Keep your search strings fairly simple.
- Try out your search strings on a general indexing system and refine them if they do not find the majority of known papers.

22.5.3 Selecting sources for a manual search

If you are using a manual search as the main strategy for a mapping study, there are several approaches you can take. If you are interested in high quality studies in a relatively mature topic (particularly if you are interested in empirical studies), the following sources are likely to be suitable for general software engineering topics:

- *IEEE Transactions on Software Engineering*
- *ACM Transactions on Software Engineering Methodology (TOSEM)*
- *Empirical Software Engineering Journal*
- *Journal of Systems and Software*
- *Information and Software Technology*
- *Proceedings of the International Conference on Software Engineering (ICSE)*
- *Empirical Software Engineering and Metrics Conference (ESEM).*

For a new topic area, you will need to review specialist conference and workshop proceedings. Whatever the circumstances, you should check the sources that published your known papers.

Points to remember

- Look for sources that are particularly likely to publish papers on your topic of interest.
- New topics are most likely to be reported in specialist workshops and conferences.
- If a topic is new, terminology may not be well-defined, complicating automated searches.
- Use your known papers to help with identifying possible sources.

22.5.4 Problems with the search process

A major search process problem is searching for topics that are unlikely to be the main research topic of research papers. For example, if you are interested in the use of some specific automated tools in a particular topic area, there may be many papers that report the use of the tools to support their validation or evaluation activities but do not mention the name of the tool in the title, abstract or keywords. Alternatively, if you are interested in the use of specific research practices, there will be difficulties because not only are specific experimental methods seldom identified in the title, abstract and keywords of primary studies, but it is also the case that software engineers are extremely poor at correctly specifying the empirical methods they used.

Searching for detailed aspects of a research process or a topic requires searches of the full research papers, which is not supported by indexing services nor by all the publishers' digital libraries. In such cases, you will probably need to do a relatively broad search and prepare to manage a large number of candidate primary studies including many false positives. In some cases, you might be able base your set of candidate papers on a randomly selected subset of the papers within the broad topic area.

Two other problems with the search process are finding too many (for example, many thousands) or too few candidate primary studies (for example, one or two).

You may find yourself with a very large number of primary studies if you are doing a mapping study of a topic with a very broad scope. In this case, you need to consider revising any automated search strings. However, before changing any search strings, you may need to reconsider your research questions. Are the research questions too broad in scope? Are any research questions unnecessary for your main research goals?

If you are doing a more focused systematic search you may find yourself with very few studies. In this case you have several options:

- Check whether your search parameters are too stringent. It might be possible that broadening the search would find additional relevant studies.
- If you have a well-designed search process, for example, your initial set of known papers was small and all those papers were found by the search, it may be that there is insufficient research for a systematic review. This can be recorded as an outcome of your systematic review, and you should, perhaps, plan to undertake your own primary study.

Points to remember

- In some cases, the information needed to answer your research question(s) won't be found in the title, abstract or keywords. This raises problems both for the search process and the selection process.

- If you have too many candidate papers about a topic likely to be mentioned in the title or abstract, are your research questions too broad?
- If you have too few candidate papers about a topic likely to be mentioned in the title or abstract, are your research questions too narrow, or are more primary studies needed?

22.6 Primary study selection process

Study selection is a multi-stage screening process by which irrelevant papers are removed from the set of candidate primary study papers. The selection process needs to be documented in the review protocol.

22.6.1 A team-based selection process

For a team-based systematic review or mapping study, the process is shown in Figure 22.7. At least two researchers should assess each candidate paper. The team leader is responsible for assigning researchers to individual papers and for collating the result of their evaluation.

Stage 1 selection is usually based on title and abstract. However, if the number of papers found by the search process is very large (for example, > 500), it may be appropriate to base a preliminary screening on title alone. Any paper that is considered irrelevant by all the researchers who assess it, based on the inclusion and exclusion criteria that can be evaluated from the title alone, is removed from the set of candidate papers.

The main Stage 1 selection activity is based on assessing the title, abstract and keywords of the remaining candidate papers. Again the process is to remove any paper that is considered irrelevant by all researchers that assess it, based on the inclusion and exclusion criteria that can be evaluated from the title, abstract and keywords alone.

At this point the team leader should review the agreement among the researchers and calculate an appropriate agreement statistic such as Cohen's kappa (Cohen 1960) or Krippendorff's alpha (Krippendorff 1978). If the agreement is poor, it is possible that some members of the research team are unclear about the interpretation of the research questions or the inclusion and exclusion criteria. The research team should meet to discuss possible reasons for poor agreement (which in our experience of software engineering papers, is sometimes due to poor quality abstracts).

After Stage 1 selection, the Stage 2 selection activity is based on the full text of the paper and all the defined inclusion and exclusion criteria. The goal of this screening activity is to positively include relevant papers as well as to exclude irrelevant papers.

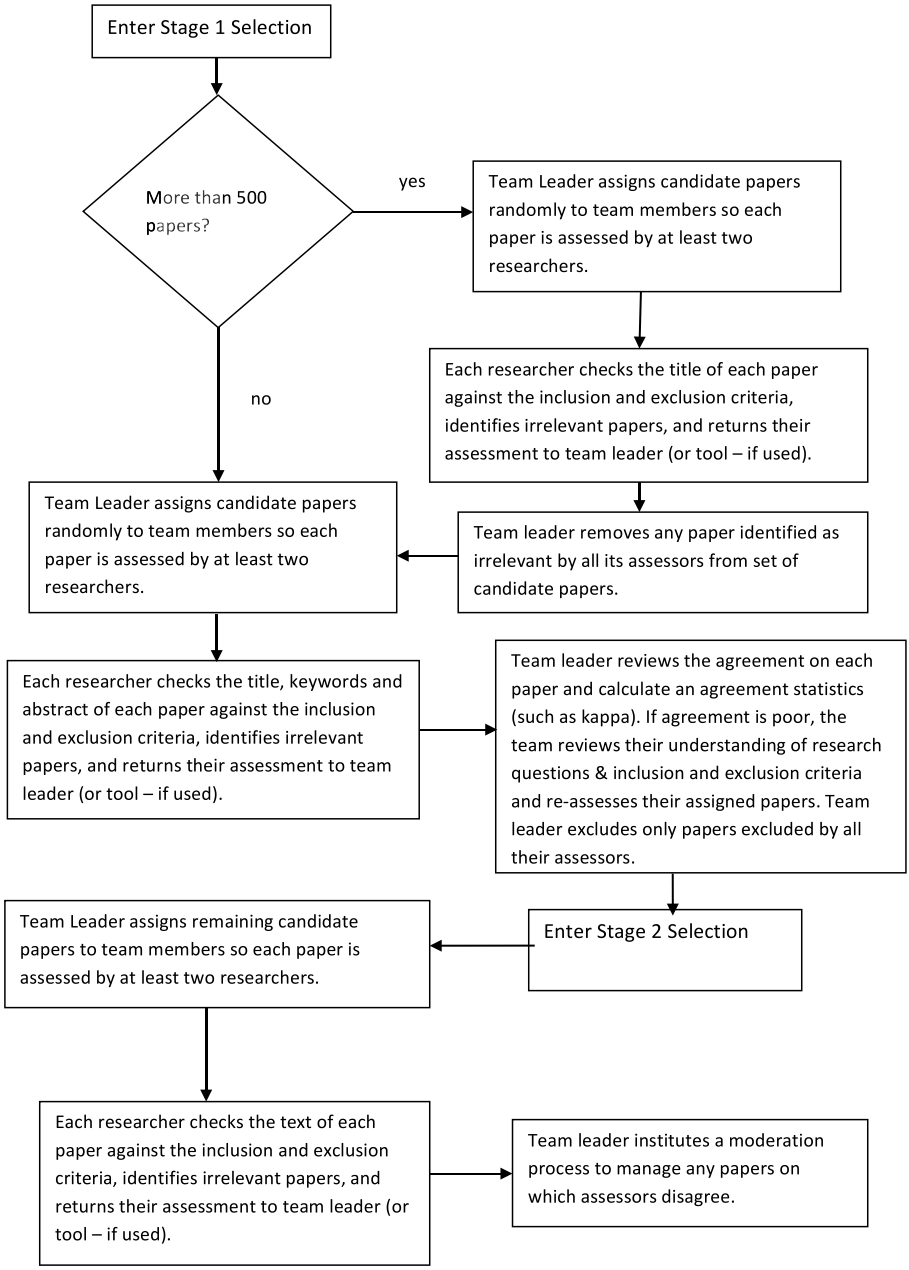

FIGURE 22.7: The team-based primary study selection process.

For a mapping study, it may be appropriate to apply the detailed inclusion/exclusion process only to papers where researchers disagreed about their

relevance during Stage 1. That is, if all researchers agreed that a paper was relevant during Stage 1, it is not necessary to check the paper again against all the inclusion and exclusion criteria. However, for a systematic review it is usually better to apply the inclusion and exclusion criteria explicitly to each paper that passes the initial screening.

After Stage 2 screening, there may be some disagreements among researchers about the inclusion of specific papers. At this point, it is important to record the agreement among researchers in terms of an agreement statistic. If agreement is very poor, the team leader may need to initiate a procedure to investigate whether there is a systematic problem with the selection process. For example:

- If all the researchers show poor agreement, the team leader could call a team meeting to discuss the inclusion and exclusion criteria, in case there are previously overlooked ambiguities or other problems with the criteria.
- If the problem appears to involve a particular researcher, the team leader might initiate some additional training before asking the researcher to reassess their inclusion/exclusion decisions.

After identifying all disagreements, the team leader needs to institute a moderation process to gain an agreement on the relevance of each disputed paper. The moderation process can involve:

- Discussion among the researchers who assessed the paper.
- Assessment of the paper by another researcher.
- A trial data extraction to confirm whether or not the required data can be obtained from the paper.

It is sometimes necessary for the team leader to make a final decision but it is preferable for the researchers to come to a mutually agreed decision.

Points to remember

- The main study selection process usually involves two or three stages.
- If you have a very large number of papers, base initial inclusion/exclusion assessments on title alone, then assess the retained papers based on the keywords and abstract of the remaining papers, and finally assess the retained papers based on their contents.
- With a relatively small number of papers, base the initial inclusion/exclusion assessment on the title, abstract and keywords and then assess the retained papers based on their contents.
- Ancillary searches usually require a separate selection process.

22.6.2 Selection processes for lone researchers

The team-based process cannot be followed by a lone researcher or a PhD student. If you are a lone researcher, you should adopt a test-retest approach whereby you assess the papers once and then, at a later time, assess them again (preferably not in exactly the same order). A substantial disagreement should prompt you to review your research questions and inclusion and exclusion criteria.

If you are a PhD student, you can involve your supervisors by asking them to assess a random selection of papers. This also gives your supervisors an opportunity to provide you with feedback. Again a substantial disagreement would be an indication that you (or your supervisor) misunderstand some aspects of your research question(s) or inclusion and exclusion criteria. If you and your supervisors plan to publish the results of the systematic review or mapping study, it is appropriate for the supervisors to act as members of the research team to ensure the selection process is of an appropriately high quality.

Points to remember

- Lone researchers should use test-retest to validate their inclusion/exclusion decisions.
- PhD students should ask their supervisors to apply inclusion/exclusion criteria to a proportion of the candidate papers and assess the level of agreement.

22.6.3 Selection process problems

If you are doing a broad mapping study rather than a focussed systematic review you may find you have a very large number of candidate primary studies (that is, many thousands of articles rather than a few hundreds). If, in addition, you are a lone researcher or the leader of a small review team, you may find this number of studies impossible to screen in the time available for the review.

Assuming that the search process has been performed correctly, you have two main options for the preliminary stage in the process:

- Recruiting more members to your review team, but bear in mind any new recruits will require time to get up to speed on the planned review procedures.
- Using a text mining tool to identify the set of papers that are most likely to be relevant to your research questions and excluding papers with a low probability of relevance.

If you are at the end of your selection process and still have a very large number of primary studies (for example, many hundreds of papers) that are now *confirmed* as being relevant to your research questions, you may anticipate a potential problem with managing the primary study analysis and synthesis process. In this, case options include:

- Recruiting more review team members, but bear in mind that, the later in the process that you decide to recruit more team members, the more difficult it will be for them to get up to speed on the planned review procedures.
- Revising your research questions, which is possible if your research questions are answered by different subsets of the primary study. You may be able to use a text mining tool to look for primary study clusters.
- Basing selection on a random sample of primary studies, using stratified sampling if the primary studies cluster by research question, domain or topic. Several systematic studies of research methods in software engineering have been based on a sampling strategy (for example, Glass et al. (2004) and Zelkowitz & Wallace (1998))

Points to remember

- Managing an extremely large number of candidate primary studies is difficult and time consuming.
- Consider contingencies for managing large numbers of primary studies during the planning process.

22.6.4 Papers versus studies

An important issue for a systematic review is the relationship between papers and individual studies. Software engineering papers often exhibit overlaps:

- There may be several different papers reporting the same study. This can occur if there is a conference version of the paper followed by an extended journal version of the paper.
- It is possible that the results of a large study may be published in a series of different papers.
- It is also possible that a single paper may report the results of several independent studies.

It is important that a systematic review does not double-count study results, particularly if some form of statistical meta-analysis is to be performed.

Thus, after the completion of the primary study selection, a research team must review papers that have similar titles and authors and assess the relationship between the papers and individual studies. We advise you to keep a record of papers related to a specific study (for completeness and auditability), but ensure your results are reported against the individual study.

Furthermore, all papers should be scanned for the possibility of multiple studies. This is by no means a straightforward procedure since authors may regard some studies as independent that you consider to be a single study. Issues that occur are:

- Researchers may report both a pilot experiment and a main experiment. In some cases, it may be appropriate to ignore the pilot experiment. This is likely to be appropriate if the authors report many changes to the research methods as a result of the pilot experiment, or the pilot experiment was based on a very small sample. In other cases, it may be appropriate to treat the pilot study and main study as one study with two blocks.

- Researchers may report several case studies, and it is unclear whether their design is a multi-case case study or several independent case studies. If the case studies have the same research questions and used the same methodology then our advice is to treat the study as one multi-case case study. If the case study methodologies are very different, for example, ethnography in one case and semi-structured interview in another, we would suggest treating the paper as reporting two independent case studies.

Often there is no obvious "right" answer to the number of independent studies. You need to report your basic approach (for example, two researchers will discuss each case) and the decisions you make in each individual case.

In most cases, this is less of a problem for mapping studies, since they are usually able to work at the paper level irrespective of the relationship between papers and individual studies. However, it is often useful to identify duplicate reports of the same piece of research, particularly if the aim of the mapping study is to identify whether sufficient research is available for a detailed systematic review. Furthermore, if a mapping study aims to assess the quality of research, you may need to consider quality at the study level rather than the paper level.

Points to remember

- The relationship between studies or pieces of research and published paper is many-to-many.

- For systematic reviews, it is important not to over-count (or under-count) the number of independent studies.

22.6.5 The interaction between the search and selection processes

Although the search and selection activities are different, the process of undertaking those activities is often entwined. If you intend to use backwards snowballing, you cannot do such a search until you have selected a set of primary studies from the set of candidate primary studies you found during your main search process. A similar issue arises if you want to write to authors who have published a number of primary studies to seek other as yet unpublished results.

Thus, after performing your main selection process be it automated or manual or a mixture, you need to suspend the search process and enter the selection process to screen the current set of candidate papers. After the screening process is complete, you will need to reactivate the search process in order to check the references of the current set of primary studies or identify the most frequently cited authors.

Clearly if you find more candidate primary studies, the selection process needs to be re-activated.

Point to remember

Ancillary searches often depend on having an existing list of candidate studies, which means they cannot be started until an initial round of the search and selection process has been completed

---

22.7 Validating the search and selection process

The team leader should assess the validity of the search process against the criteria specified in the protocol. This information should be reported in the methods section of the final report.

The team leader should expect to:

- Justify the comprehensiveness of the search process given the type of review, that is, whether it is a quantitative or qualitative systematic review or a mapping study.
- Report the agreement achieved during the Stage 2 selection process, prior to any moderation process (see Figure 22.7).
- Confirm that all papers that were known before the start of the selection process were found by the search process and selected during the selection process.

- *Optional*. Consider validating the search process using textual analysis tools (see Chapter 13). Such tools can be used to check the frequency of the use of main keywords to investigate whether included papers that and excluded papers that differed with respect to usage of keywords and if they did, whether any papers might have been misidentified. Tools can also be used to review the extent to which included and excluded papers cross-reference one another. Again, such an analysis can be used to identify any papers that might have been misclassified. Any papers that may have been misclassified can be reviewed again and their classification revised if necessary. This approach is particularly useful for single researchers.

- If any known papers were kept separate for validation purposes, the team leader must report the coverage of these papers. For a systematic review, coverage should be 100% if the study involves a comparison of two technologies.

- If any previous systematic reviews or mapping studies were kept separate for validation purposes, the team leader must identify the primary studies reported by the previous reviews that should have been found by the current review. The team leader should report the number of such primary studies that were missed by the current review.

Of course, if the final two validation exercises discover missing papers they must be added to the set of primary studies.

Points to remember

- You will need to justify your overall search and selection process.
- You will need to provide evidence that your search process was effective.
- You should report the values of agreement statistics to confirm selection process was effective.

22.8 Quality assessment

The main decisions that need to be made during quality assessment are:

1. Deciding whether or not a quality assessment is necessary.
2. Deciding appropriate quality assessment criteria.

3. Deciding how the quality assessment will be used to support the goals of the review.

4. Deciding how the quality assessment will be managed.

The results of these decisions should be documented in the protocol. Each of these issues is discussed below.

22.8.1 Is quality assessment necessary?

For any systematic review, quality assessment should be considered mandatory. It is important to ensure that the results of any aggregation are based on *best* available evidence. This means either simply excluding poor quality studies from the aggregation, or investigating the impact of excluding such studies.

For mapping studies, quality assessment is not required, unless one of the aims of the mapping study is to assess the quality of existing studies. This can happen in the case of tertiary studies investigating the methodology used in systematic reviews.

22.8.2 Quality assessment criteria

There are two aspects to quality assessment:

- Assessing the quality of individual primary studies.
- Assessing the overall strength of evidence of the review findings.

These are discussed in the following sections.

22.8.2.1 Primary study quality

Quality assessment of primary studies is usually done by means of a quality instrument comprising a number of questions related to the goals, design, conduct and results of each study. The questions are referred to as *quality criteria*. The quality instrument is often referred to as a quality checklist. A checklist for a particular study type is usually made up of questions related to:

- The goals, research questions, hypotheses and outcome measures.
- The study design and the extent to which it is appropriate to the study type.
- Study data collection and analysis and the extent to which they are appropriate given the study design.

- Study findings, the strength of evidence supporting those findings, the extent to which the findings answer the research questions, and their value to researchers and practitioners.

Quality assessment criteria usually depend on the *type* of primary study being evaluated since factors that determine a good example of one type of study may be irrelevant for a different type of study. For example, factors that identify a good quality experiment such as random allocation to treatment, and sufficient experimental units to achieve a reasonably high power are different from the factors that determine a good case study such as an appropriate choice of *case*, and consideration of alternative explanations for the case study findings.

Quality assessment criteria also depend on the *subject type*. For studies that compare human-intensive methods or techniques, the subjects will be human beings and checklists can be adapted from the many recommendations available in the medical and healthcare domain. We advise you to look at some of the published checklists, choose the one(s) most appropriate for your systematic review, and adapt it (if necessary) to your own study. For example, checklists for randomized controlled trials (which are field experiments), qualitative studies, and systematic reviews can be found at the Critical Appraisal Skills Programme (CASP) website[2] or the SURE Critical Appraisal Checklists[3]. A version of the randomised trials checklist suitable for software engineering experiments is shown in Figure 22.9. In addition, Runeson et al. (2012) provide checklists specifically designed to help researchers undertaking and reading software engineering case studies.

In contrast, for studies that compare or evaluate algorithms or tools which are *technology-intensive* studies, specialised checklists will need to be constructed, see for example Figure 22.8 which is adapted from a checklist developed by Kitchenham, Burn & Li (2009), and includes suggestions for scoring each question.

For non-comparative or qualitative systematic reviews, or if a large variety of study types are found, a more general quality assessment instrument may be appropriate. The quality criteria proposed by Dybå & Dingsøyr (2008*a*) based on the CASP checklist for qualitative studies has been used in several software engineering systematic reviews. However, if you have a large number of different study types but only one quality checklist, you need to consider the study type as well as answers to the checklist questions when using the quality assessment (see Section 22.8.2.2 below).

22.8.2.2 Strength of evidence supporting review findings

Dybå & Dingsøyr (2008*b*) recommend using the GRADE approach (Guyatt, Oxman, Vist, Kunz, Falck-Ytter, Alonso-Coello & Schünemann

[2]http://www.casp-uk.net/find-appraise-act/appraising-the-evidence/
[3]http://www.cardiff.ac.uk/insrv/libraries/sure/checklists.html

Question No	Question	Scoring
1	Are the goals of the experiment clear	No=0, Partly=0.5, Yes=1
2	Were the research questions and hypotheses defined?	No=0, Partly=0.5, Yes=1
3	Was there any replication, for example, multiple test objects, multiple test sets?	Yes=1, No=0 – If No this is not an experiment and should be considered a case study, feasibility study or example.
4	Are the study measures valid?	None=0 / Some (0.33) / Most (0.75) / All (1)
5	If test cases were required by the Test Treatment, how were the test cases generated?	Not applicable (reduce number of questions by 1) By the experimenters (Yes=0) By an independent third party (Yes=0.5) Automatically (Yes=0.75) By industry practitioners when the test object was created (Yes=1)
6	How were Test Objects generated?	Small programs (Yes=0) Derived from industrial programs but simplified (Yes=0.5) Real industrial programs but small. (Yes=0.75) Real industry programs of various sizes including large programs (Yes=1)
7	How were the faults/modifications found?	Not applicable (reduce number of questions by 1 and go to question 8) Naturally occurring Yes=1, go to question 8 If No go to questions 7a
7a	For seeded faults/modifications, how were the faults identified?	Faults introduced by the experimenters (Yes=0), Independent third party (Yes=0.25) Generated automatically (Yes=0.5)
7b	For seeded faults/modifications, were the type and number of faults/modifications introduced justified?	Type & Number: Yes (0.5) Type or Number (Yes=0.25) No=0
8	Did the statistical analysis match the study design?	No=(0), somewhat (0.33) Mostly (0.66), Completely (1)
9	Was any sensitivity analysis done to assess whether results were due to a specific test object or a specific type of fault/modification?	Yes=1 / Somewhat=0.5 / No=0
10	Were limitations of the study reported either during the explanation of the study design or during the discussion of the study results?	No=0, Somewhat=0.5, Extensively=1
11	Were the findings clearly reported?	No=0, Partly=0.5, Fully-1
12	Are the findings of value to industry or researchers?	No=0, Somewhat=0.5, Extensively=1

FIGURE 22.8: Quality criteria for studies of automated testing methods.

2008) to assess the strength of the evidence for recommendations. The GRADE approach is mainly used to assess the strength of *recommendations* when a decision has to be made concerning the *adoption* of a recommendation in a particular situation. However, it can also be considered for assessing the strength of evidence associated with individual findings.

GRADE defines strength of evidence in terms of the confidence we have that further research will or will not change the estimate of effect size:

1. High confidence means that further research is unlikely to change the estimate.
2. Moderate confidence means further research may change the estimate.
3. Low confidence means further research is likely to change the estimate.

Broad issue	Detailed Issue	Factors to consider
Are the results of the experiment valid? Initial screening questions	Did the experiment address a clearly focussed issue?	Is the population study defined? Are the methods (including the control) well defined? Are the outcomes appropriate?
	Was the assignment of participants to methods randomised?	How was assignment carried out? Was the allocation concealed from researchers and subjects?
	Were all participants who took part in the study accounted for at its conclusion?	Were participants analysed in the groups to which they were assigned? How were partial results from dropouts handled? Were dropout rates related to the method?
Validity – detailed questions	Were any of the experimenters "blind to treatment"?	Did the researchers running the experiment know who was in which treatment group? Were the outputs independent of the method and if so were markers/evaluators blind to the method used by each participant?
	Were the groups similar at the start of the experiment?	Have other factors been considered such as SE experience, knowledge of the different methods?
	Aside from the method, were groups treated equally?	Were both groups given appropriate training? Were the trainers equivalently skilled in the method(s) they taught? Did the trainers have a vested interest in the success of one of the methods?
What are the results?	How large was the treatment effect?	What outcomes were measured? Is the primary outcome clearly specified? What effect sizes were found for each outcome?
	How precise was the estimate of the treatment effect?	What are the confidence limits?
Will the results help locally?	Can the results be applied in your context?	Are the participants similar to the intended population, for example, being practitioners rather than students?
	Were all industrially important outcomes considered?	Is there other information you would like to have seen? If not does this affect the value of the experiment?
	Are the benefits worth the risks and costs?	Will likely cost savings outweigh adoption costs?

FIGURE 22.9: Quality criteria for randomised experiments.

4. Very Low confidence means the estimate is very uncertain.

GRADE also considers factors that may decrease or increase confidence in the strength of evidence. Factors that decrease the strength of evidence relate to poor methodological quality, inconsistent findings, sparse data, or reporting bias found in individual studies. Factors that increase the strength of evidence relate to very large effect sizes, only having confounders that would decrease

the effect size, or evidence of a "dose response gradient" (which means that more of the treatment results in a better outcome).

Formulating GRADE evidence in terms of effect sizes and study bias indicates that the method is intended to apply primarily to randomised controlled trials (which are controlled field experiments) or systematic reviews of such studies. However, such studies are extremely rare in software engineering. We must often make do with much weaker forms of study and more varied types of empirical study

To apply the GRADE concept to findings from software engineering systematic reviews, specific findings from the review will need to be discussed in the light of the methodological quality of the related primary studies and their study type(s). For example, if a specific finding is supported only by feasibility studies, even if they are high quality feasibility studies, the evidence for that finding must be considered to be very weak.

Study types that provide very weak evidence are:

- Feasibility studies, including small experiments (that is, experiments with very few subjects), and small-scale examples.
- Lessons learned studies.
- Before-after within-subject quasi-experiments which are the weakest form of quasi-experiment.

Types of study that provide slightly stronger evidence (although still relatively weak) include:

- Post-hoc re-working of large-scale examples (often mistakenly called *case studies*).
- Post-hoc analyses of industry datasets (for example, correlation and regression studies).

Types of study that provide moderately strong evidence include:

- Laboratory-based experiments and quasi-experiments. Good quality studies of these types are likely to give a reliable indication of whether or not an effect size is significant and the *direction* of the effect size. However, they are likely to give biased estimates of the *magnitude* of the effect size (probably overestimates) because the studies did not take place in an industrial context where other factors such as time scale pressure, team dynamics, task complexity, task dependencies, and personal motivation influence outcomes. Also, if the studies involve students rather than professionals or the software engineering task is particularly straightforward, confidence in the evidence should be downgraded.
- Industry case studies, preferably using multiple cases, which can provide reasonably reliable evidence.

The most trustworthy form of evidence in software engineering comes from industry-based field studies including:

- Randomised field experiments. Such designs represent the most reliable form of empirical study but can seldom be performed in software engineering contexts.
- Field-based quasi-experiments based on cross-over designs, interrupted time-series, regression discontinuity, or differences-in-differences designs (Shadish et al. 2002) which can provide highly reliable evidence.

22.8.3 Using quality assessment results

There is little point in collecting data about primary study quality if you have no plan as to how such data will be used. There are several possibilities:

- Specific quality criteria may be used as part of the inclusion criteria to screen out low quality studies.
- The quality score (that is, the sum of numerical values assigned to the answer of each quality question) for each primary study may be used to identify poor quality studies. Then, the impact of the poor quality studies on the results of the review findings can be assessed to investigate whether poor quality studies are causing the results to be biased.
- The quality data may be assessed to see if there are systematic problems with primary study quality, for example it may be problematic if most or all of the studies use student participants.
- Specific quality criteria may be used as moderating factors (that is, factors that might explain the differences among study results) in meta-analysis, for example whether or not the empirical study was an experiment (formal or quasi) or a less rigorous form of study type.

Finally, as mentioned above, the quality score of studies can be used as part of an assessment of the strength of evidence supporting individual findings.

22.8.4 Managing the quality assessment process

Many quality criteria require subjective assessment, for example, any criterion that asks whether something was *appropriate*. To reduce the problem of bias associated with subjective assessments, it is customary for at least two researchers to assess the quality criteria of each paper and for disagreements to be moderated.

The general process used for managing quality assessment in a team-based systematic review is shown in Figure 22.10. The specific process you decide to adopt should be documented in the systematic review protocol.

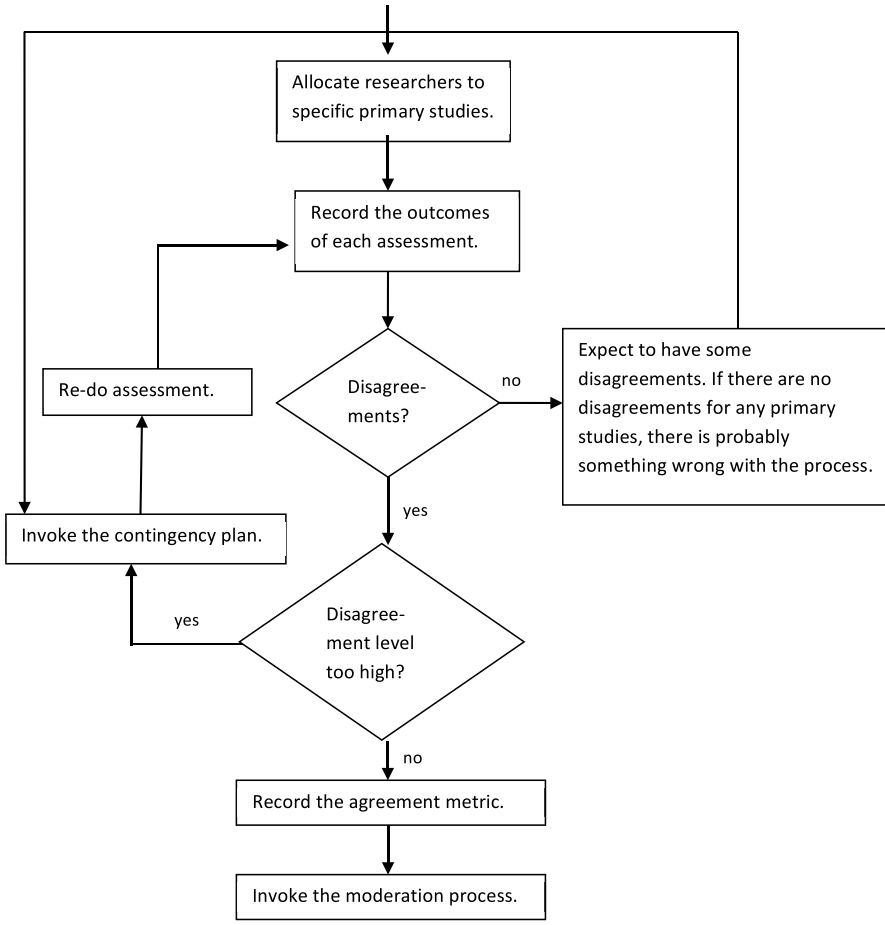

FIGURE 22.10: Process for managing team-based quality assessment.

22.8.4.1 A team-based quality assessment process

For team-based systematic reviews, the team leader should assign at least two researchers to assess the quality (and study type, if necessary) of each primary study. Each researcher should complete the quality evaluation form independently, then the results of the evaluations for a specific primary study should be compared and disagreements recorded (this may be done by the team leader, a systematic review management tool, or the two assigned researchers working together).

If there are disagreements, some form of moderation must take place as defined in the systematic review protocol. Options include:

- Assigning a third person to assess the quality of the primary study and discuss the assessments with the two original researchers.

- Asking the researchers assigned to the primary study to work together to arrive at an agreed assessment.
- If more than two assessments are available, aggregating the assessment into a combined score (for example, by taking the mean).

The team leader should expect to report initial agreement rates using an appropriate agreement measure. The agreement between independent assessors is used to assess the validity the evaluation process. If the agreement for individual primary studies is very low, the quality criteria may not be well understood, so there should be a contingency plan ready. Options for the contingency plan should include:

- Calling a team meeting to discuss the quality criteria and the reasons for disagreements.
- Additional training for specific members of the team.

The team leader is responsible for deciding which option should be chosen given the specific circumstances. It should also be noted that unusually high agreement can also be a sign of misunderstandings among team members. The team leader should be prepared to invoke the contingency plan if this condition arises.

The actual quality evaluation process may take place as part of the general data collection process or may precede the data collection phase. If some of the quality criteria are being used as inclusion/exclusion criteria, it is better to complete quality evaluation before beginning data extraction.

22.8.4.2 Quality assessment for lone researchers

If you are a lone researcher or a PhD student, the main problem you will have is validating your quality assessment.

If you are a PhD student, options include:

- Requesting your supervisors to assess a random selection of primary studies and comparing the results.
- Re-assessing of a random selection (or all) of the primary studies after a suitable elapsed time and calculating the test-retest agreement.

In both cases, you should specify in the protocol what constitutes a dangerous level of disagreement, and have ready a contingency plan to deal with this possibility. Any disagreements identified during this validation exercise need to be resolved. This should usually be done by discussing each case with a supervisor.

If you are a lone researcher you must also specify how to validate your quality assessment process. This will usually require a test-retest assessment based on all the primary studies and a justification of the process for reaching an agreed evaluation for any disagreements (which might be taking the mean score) or recording a justification for each revised score.

Points to remember

- The subjective nature of many quality criteria makes quality assessment far from simple.
- You should be clear about how the quality assessment will be used.
- If your primary studies include many different study types and you use a general-purpose quality checklist, keep a record of the study type as well. In this case you should use the quality assessment information and study type to assess the reliability of individual findings.
- You should report agreement statistics to indicate the reliability of the quality assessment process.

22.9 Data extraction

Data extraction and data synthesis are phases where the differences between quantitative systematic reviews, qualitative systematic reviews and mapping studies are most significant. You need to consider the type of review you are undertaking both when planning your data extraction process and when conducting data extraction.

22.9.1 Data extraction for quantitative systematic reviews

The data extraction process is most well-defined for quantitative systematic reviews. You should be in a position to define in advance the data you intend to extract from each paper in order to answer your research question(s). However, this presupposes that you know enough about the topic area and the available literature to determine whether a meta-analysis is feasible and if so what effect sizes are most appropriate. If this is not the case, you will need to defer formalising the data extraction and analysis processes until you have selected the primary studies and identified the statistical designs used in the studies and the nature of the outcome metrics they report.

22.9.1.1 Data extraction planning for quantitative systematic reviews

Once you have adequate knowledge of the primary studies, the decisions you need to make to identify the data you need to collect are shown in Figure 22.11.

You will need to decide whether you intend to undertake a formal meta-analysis or a more qualitative-style of analysis. Even with a quantitative systematic review, you will not be able to do a formal meta-analysis:

- If the primary studies use different treatment combinations, and there are insufficient studies that compare the same pair of treatments.
- If your outcome measures include many different incompatible metrics. For example, the quality of a program (or a maintenance change made to a program) might be evaluated using static complexity measures, number of residual errors, subjective quality assessments, or test coverage statistics. You may be able to identify whether differences are statistically significant which is all that is needed for vote-counting, but more detailed meta-analysis methods may not be possible.

Whether you are aiming for vote-counting or a full meta-analysis, data extraction will be based on:

- Basic information about the study, including the treatments being compared and the outcome metrics reported.
- The quantitative outcomes of the experiments being included in the review, as required for the type of meta-analysis being planned (see Chapter 11 and Table 22.1). This would include all the metrics needed to construct the specified effect size such as values of any test statistics, the probability level achieved by the test(s), sample sizes, mean values, standard deviations.
- Contextual information that can be used in any meta-analysis to investigate any heterogeneity among primary studies or support a detailed qualitative analysis (Chapter 11 and Table 22.2). Note, however, that some relevant contextual information may already have been specified in the quality criteria.

After defining the data to extract, you will need to construct a data collection form. This can be a paper form, a spreadsheet, or database form. The data collection form and any necessary associated documentation should define the data being extracted and provide clear guidelines for data extractors. To complete the planning process, the form should be trialled using some known primary studies.

All members of the review team who are expected to extract data should take part in the trial and report any problems with the data collection form to the team leader. Any problems with the form should be resolved prior to finalizing the protocol. A procedure both for checking individual extraction forms and for resolving any disagreements should be defined in the review protocol.

TABLE 22.1: Common Effect Sizes Used in Meta-Analysis

Type	Formula	Definition	Variants
Significance level	p-value	Probability obtained from a statistical test	
Point serial correlation for between groups design	$r = \frac{\sum(x_{ij}-\overline{x})(y_{ij}-\overline{y})}{\sqrt{\sum(x_{ij}-\overline{x})^2 \sum(y_{ij}-\overline{y})^2}}$	Pearson correlation where $x_{ij} = 0$ for group 1 and $x_{ij} = 1$ for group 2 and y_{ij} is the outcome value for observation j in group i.	R-squared for ANOVA designs. Standard Pearson correlation for regression or correlation studies.
Standardized mean difference for numerical outcome metrics	$d = \frac{(m_1 - m_2)}{stdev}$	Difference between the means of observations in each group divided by an appropriate standard deviation.	Cohen's g, Hedge's d, and Glass's Δ.
Odds ratio for counts and probability outcome metrics	$O = \frac{p_1(1-p_2)}{p_2(1-p_1)}$	Ratio of odds related to one group and odds related to a second group. Odds are the probability of an event divided by 1 minus the probability.	Log odds which is the logarithm of the Odds ratio.

TABLE 22.2: Contextual Information Appropriate for Meta-Analysis

Type	Options	Value
Study type	Experiment, Case study, Quasi-experiment, Survey, Benchmarking, Data mining, Lessons learnt	Provides information about constraints on study rigour.
Participants	Students, Practitioners, Consultants, Academics	Indicates the population to which results apply.
Materials	Programs, Software specifications, Test cases	For benchmarking studies or testing studies, define type of systems to which results apply.
Settings	University course, Training course, Industry	Indicates realism of setting.
Task	Task time, Task complexity	Indicates realism of task.

22.9.1.2 Data extraction team process for quantitative systematic reviews

Once the data collection form has been designed and tested, the actual data extraction process should be fairly straightforward. For a team-based review, the team leader must assign two team members to each primary study and monitor the data extraction process (see Figure 22.9.1.1). Note that data extraction may be done at the same time as extracting quality data.

However, it is always possible that a primary study could be found that performed a novel analysis and presented its results in a manner that was not anticipated in the protocol. This should be reported to the team leader, who needs to halt further data extraction in case there are other examples of such analyses among the primary studies. Data extraction should only be restarted when it is clear how to deal with papers using the new type of analysis. This might involve amending the data collection form and/or providing additional training/guidelines for data extractors.

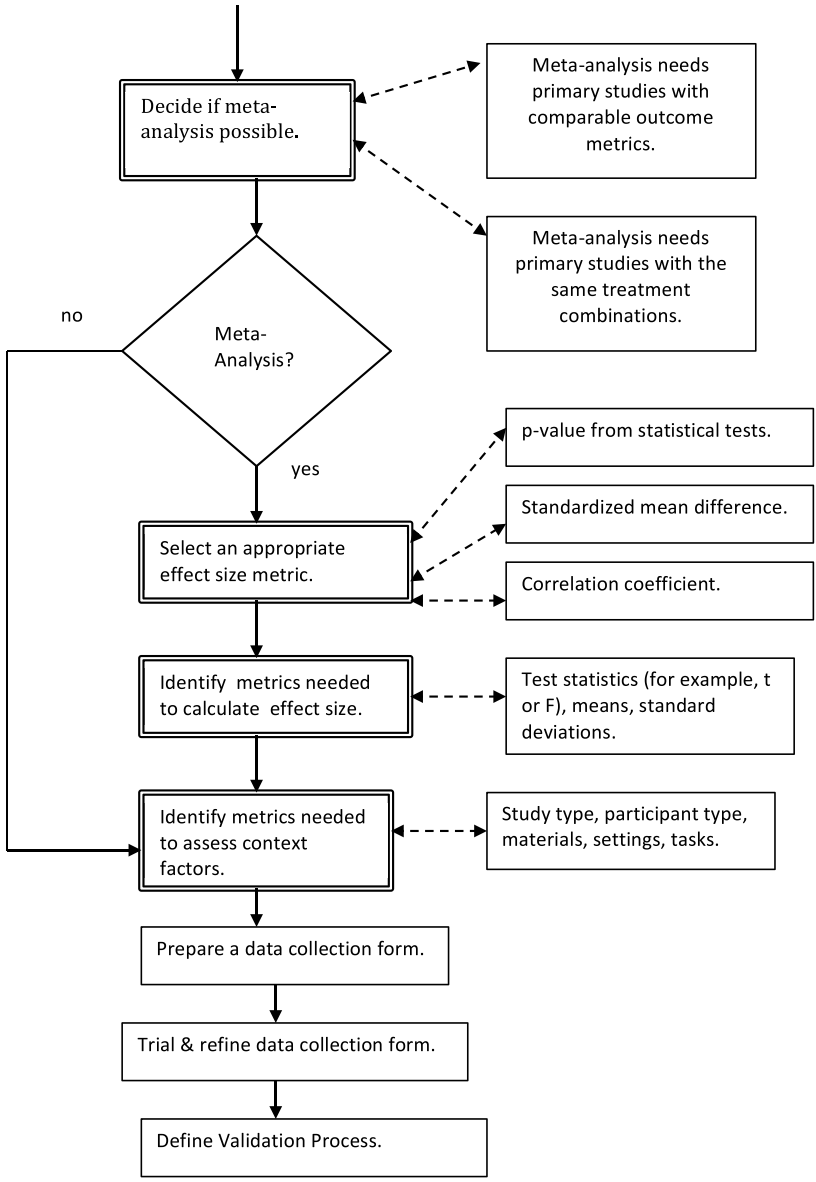

FIGURE 22.11: Initial planning decisions for quantitative systematic reviews.

22.9.1.3 Quantitative systematic reviews data extraction process for lone researchers

If you are a lone researcher, you should use a test-retest approach to validate that all the data was correctly extracted.

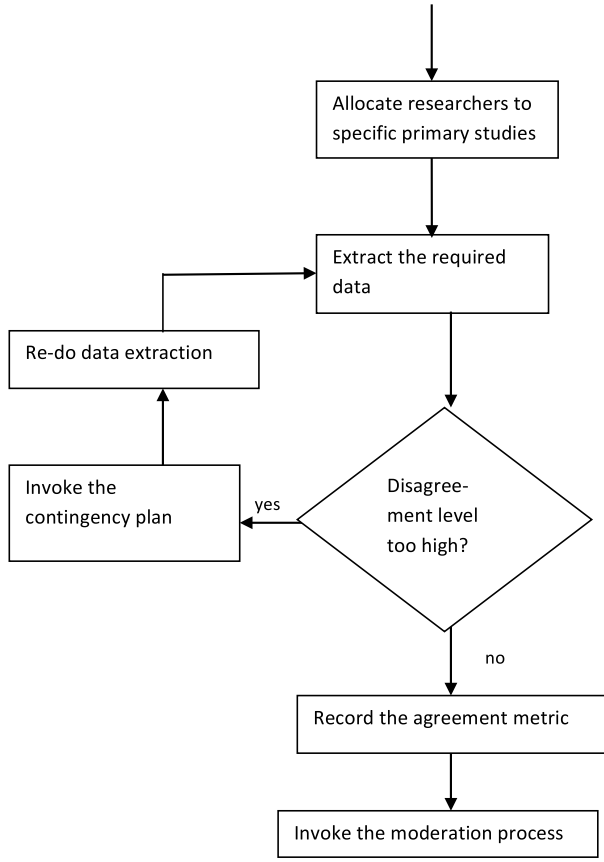

FIGURE 22.12: Quantitative systematic reviews data extraction process.

PhD students can either use test-retest, or ask their supervisors to act as independent extractors. Note, however, performing a full systematic review would be a major task for a PhD student and should be suitable for a journal publication. In such circumstances supervisors should be prepared to act as members of the review team and ensure all the extracted data is properly validated.

22.9.2 Data extraction for qualitative systematic reviews

Qualitative systematic reviews are usually based on data extracted from qualitative primary studies. Qualitative studies are primary studies that:

- used semi-structured or unstructured interviews, or
- were based on observations researchers made about software developers, software teams or managers and their work processes, or

- were based on subjective opinion surveys.

They are the most difficult type of review from the viewpoint of data extraction and data synthesis. This is because such reviews are looking for *textual* information provided by the study authors of issues such as risks, cost benefits, motivators, barriers to adoption, definitions of terminology and other themes or concepts related to the research topic. Problems arise because authors of primary studies may use different terms for the same concept or the same terms for different concepts. This means the textual information required from each primary study cannot usually be defined in advance. Furthermore, data extraction and data synthesis become inextricably linked as you attempt to identify and define core terms including all homonyms and synonyms.

22.9.2.1 Planning data extraction for qualitative systematic reviews

In most cases you should expect the data extraction process to centre on creating a database of evidence, similar in concept to a case study database (Yin 2014). You should specify, in the protocol, the tool that will be used to hold the data. The data itself would usually include textual information extracted from each primary study, a reference to the place in the primary study the text was found, and a comment indicating the relevance of the text (if you have several different research questions, you should identify which research question(s) it addresses.)

Some general points can be made:

- The more specific your research questions are, the easier data extraction and synthesis will be, since you will be able to specify relevant themes prior to starting data extraction. In contrast, if you have some general high level topic such as "Global Software Development" or "Cloud-Based software engineering" and intend looking for unspecified "themes" from the information reported by the primary study authors, your task will be much more difficult.

- The more familiar you are with the topic area, the more likely you are to be able to identify appropriate research questions and interesting themes in advance.

22.9.2.2 Data extraction process for qualitative systematic reviews

It is difficult to organise a team-based data extraction/synthesis process. Since any primary study could introduce a new homonym or synonym, none of the primary studies can be considered as independent for the purposes of data extraction and synthesis. Currently our personal experience and experiences reported by other researchers have consisted only of two-person teams where either one person does all the extraction and defines a set of terms which the other member of the team then checks, or both team members jointly read

and extract data from each paper agreeing terminology together. In both cases, the extraction is likely to involve considerable iteration as new terms and concepts are identified and need to be reconciled with the data extracted from previously reviewed primary studies. There are several other options:

- You could use a text analysis tool such as *NVivo* to identify relevant areas of text across all the studies, but as yet there have not been any large-scale software engineering systematic reviews that have reported using such tools. If you are a lone researcher or post-graduate student, the use of a textual analysis tool is a particularly attractive option.
- All team members could read all the primary studies to get a sound overview of the topic area, and then work together to define appropriate themes and agreed terminology before undertaking a systematic data extraction and synthesis process.

However, as yet we have no definitive evidence as to which process is the most effective.

Whether you are a lone researcher or a member of a team, we advise you to read some of the papers and trial various data extraction and synthesis processes on some known studies before making any firm decisions about how to organise data extraction and synthesis.

22.9.3 Data extraction for mapping studies

Mapping studies are generally about finding and classifying the literature related to a specific topic area. Reviewers need to specify a set of characteristics that define the nature of the topic area. In the sense that characteristics might seem similar to themes, there may appear to be some overlap with qualitative systematic reviews, the differences are:

- In a mapping study, the type of study (for example, theoretical or empirical) is a means of classifying the primary study. In contrast, a qualitative systematic review will usually use the type of study as an inclusion/exclusion criterion (for example, including only empirically-based qualitative studies) or as part of a quality assessment of the primary studies.
- A mapping study is not usually concerned with the outcomes of empirical studies whereas a qualitative systematic review aims to aggregate information from the outcomes of qualitative primary studies.

22.9.3.1 Planning data extraction for mapping studies

Mapping studies aim to organise and classify the literature on a specific topic area. This process is done by identifying a set of *features* (sometimes

referred to as *attributes*, or *characteristics*) that describe the research goals and methods employed in the topic area. A feature is often specified as a set of mutually exclusive categories to which a primary study can belong, for example, the *research type* might be one feature with categories *case study*, *experiment*, *quasi-experiment*, *opinion survey*, *lessons learnt*, *personal opinion*, etc. In this case, the feature is represented as a nominal scale metric. Features may also be ordinal scale, for example, the feature *concept definition* might have one of the values *fully defined*, *partially defined*, *undefined*. Other features may be or integer-valued or real-values. Features may relate to one another in hierarchies, for example the category *experiment* belonging to the feature *research type*, might be a sub-feature with categories *fully randomised*, *randomised block*, *latin square*, *n by m factorial*, etc. The features required for a mapping study are related to the specific topic area and the research questions.

The planning activities for a mapping study are shown in Figure 22.13. You need to specify in the protocol the features you will use to classify each primary study, as discussed in Section 22.4.2.2. The major problem with mapping studies is that it may be difficult to identify in advance all the features of interest. We suggest a multiple-phase data extraction process.

Firstly some information needed to answer some of your research questions will already be available from the primary study citation information, for example:

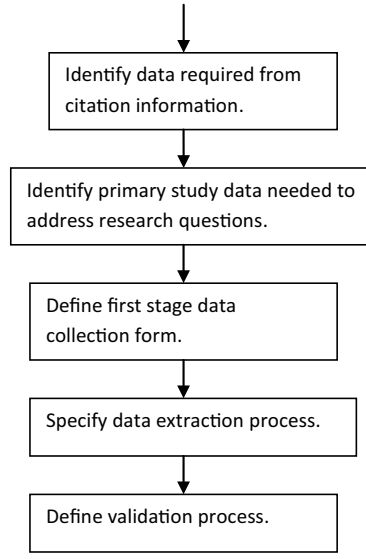

FIGURE 22.13: Planning mapping studies.

- Date of Publication
- Publication type (journal, conference, workshop, technical paper)
- Publication source (journal, conference, workshop name)
- Authors' names, affiliation and country

This information can be specified in the protocol and should be in a suitable format as the outcome of the primary study selection process.

Next, some information you need will be derived from your research questions and can be specified in the protocol. This information will be the basis of the first stage of data extraction. This will include some features and some free-format textual information such as:

- The type of study (using, for example, the categories proposed by Wieringa et al. (2006)).
- The goal(s) of the paper.
- The specific topic(s) or subtopic(s) being addressed in the paper.
- Any issues of interest raised in the paper.

Other information of interest may be identified during the data extraction process.

22.9.3.2 Data extraction process for mapping studies

The overall process for mapping study data extraction is shown in Figure 22.14. The first task is to ensure that all the citation information is held in a format suitable for analysis.

The first stage of data extraction uses the standard data extraction process with at least two members of the team manually extracting data from each primary study.

The team leader should then convene a team meeting to discuss whether there are any more trends or general topics of interest against which to classify your primary studies. If more features are identified, the team leader will need to amend the protocol to include some additional research questions and organise a second round of data extraction to classify the primary studies against the newly defined features. A further round of data extraction would follow the usual data extraction process (that is, two extractors and a process for moderating disagreement). This iterative process continues until no more interesting topics are identified.

Note in the case of mapping studies, the quality of individual primary studies is rarely evaluated and data extraction can usually begin as soon as study selection is completed.

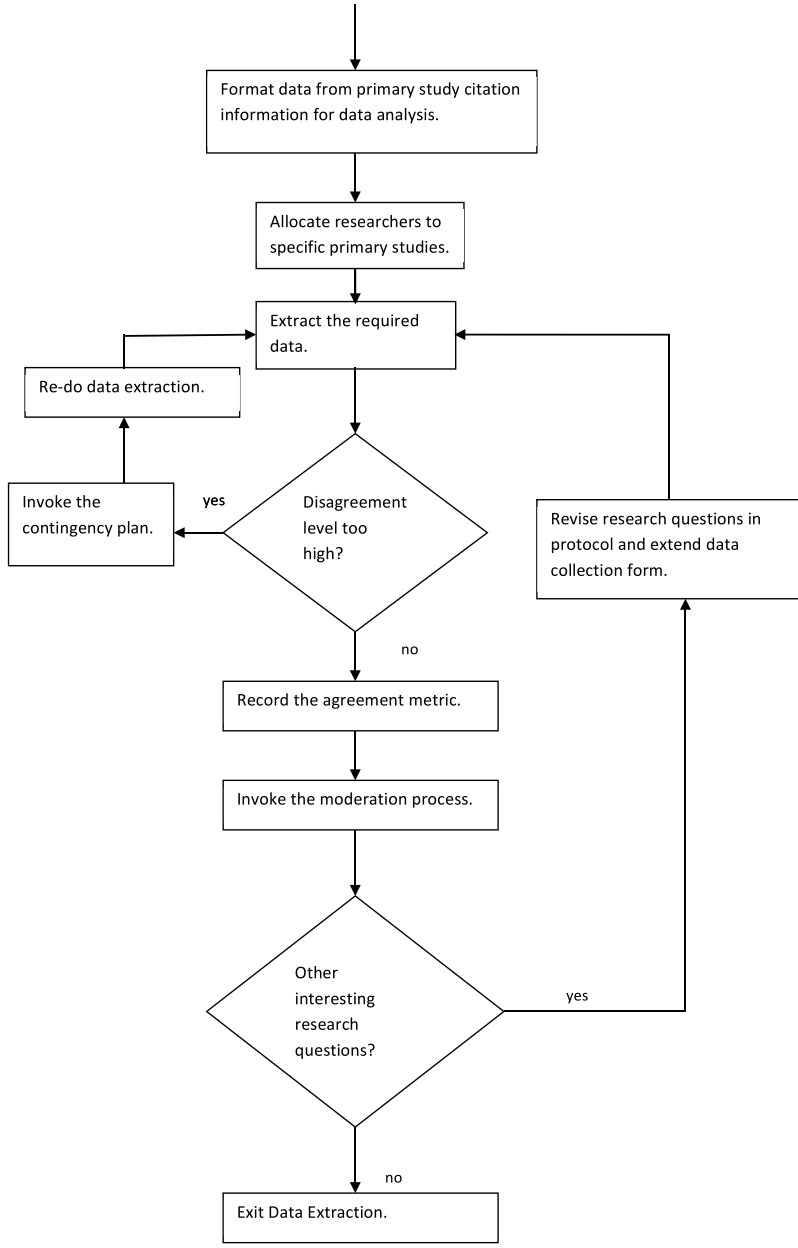

FIGURE 22.14: Mapping study data extraction process.

22.9.4 Validating the data extraction process

For quantitative systematic reviews, the team leader should expect to report initial agreement rates using an appropriate agreement measure. Similar to the quality evaluation process, the agreement between independent assessors is used to assess the validity of the data extraction process. If the agreement for individual primary studies is very low, the data form may not be well understood, so there should be a contingency plan ready. Options for the contingency plan should include:

- Calling a team meeting to discuss the data form and the reasons for disagreements.
- Additional training for specific members of the team.

The team leader is responsible for deciding which option should be chosen given the specific circumstances.

For mapping studies, the process is similar to that for quantitative systematic reviews but if the data extraction process requires two (or more) separate data extraction steps, the agreement measures should be evaluated for each step separately. Felizardo et al. (2010) have suggested using visual text mining tools to support the classification of primary studies in mapping studies. This may be a useful validation approach, particularly for lone researchers and postgraduate students.

It is not clear how data extraction for qualitative primary studies should be validated. It might be possible to use a tool such as *NVivo* to check whether it finds the same textual elements as the review team, but we are not aware of any systematic reviews that reported using this approach.

22.9.5 General data extraction issues

Whatever type of review you are doing, you are supposed to employ critical reasoning when reading primary studies. If you identify some interesting trend or characteristic common to many studies, but it is not mentioned in the protocol, do not ignore it. You should notify your team leader. The team leader needs to decide whether the data extraction process (and the protocol) need to be enhanced to collect information about the newly identified characteristic.

If you are a postgraduate student doing a mapping study as a starting point for your PhD, you need to recognise that the main aim is not to classify a set of primary studies. A mapping study should help you to find the most relevant studies to read, and classifying them may allow you to present a well-organised literature review in your thesis. However, the main aim is for you to read and understand the topic area you intend to study. The secondary aim is for you to understand how to conduct a systematic literature search.

Points to remember

- For quantitative systematic reviews the data extraction process should, in principle, be fully defined in the protocol. However, you need to be alert to any circumstances that indicate an omission in the protocol and be prepared to amend the protocol if necessary.
- For mapping studies, it is not always possible to define all the trends and topics of interest in the protocol. You should expect to iterate the data extraction process if new trends or topics of interest are identified during the data extraction process.
- For qualitative systematic reviews, it is difficult to define the data extraction process or the data synthesis process in advance. You are usually only able to decide whether or not to use a textual analysis tool, and specify the basic strategy that will be used.
- For qualitative systematic reviews, data extraction and data synthesis cannot be regarded as independent processes. In general, you should expect data extraction and synthesis to be iterative including re-reading papers and re-evaluating definitions of terms and themes.

22.10 Data aggregation and synthesis

Like data collection, data synthesis depends on the type of review you are undertaking. There is some disagreement among quantitative and qualitative researchers about the use of the terms "synthesis" and "aggregation". Qualitative researchers use the term "aggregation" to describe results obtained either from statistical analysis or simple counts, whereas "synthesis" is used to refer to analyses that *interpret* findings from qualitative studies. However, quantitative meta-analysts also refer to their statistical analyses as "synthesis". For the purposes of these guidelines we will refer to meta-analysis and qualitative meta-synthesis using the term "synthesis" and we use the term "aggregation" to refer to analysing results from a mapping study.

22.10.1 Data synthesis for quantitative systematic reviews

Your data synthesis process will depend on your data collection plan. This will either be vote counting with a qualitative analysis or a full meta-analysis. For more details on how to perform a meta-analysis consult Chapter 11; vote counting is discussed in Chapter 10.

22.10.1.1 Data synthesis using meta-analysis

Meta-analysis is a statistical method to synthesise the results from primary studies that have reported a statistical analysis of the same (or very similar) hypotheses. For a full meta-analysis, we recommend the process shown in Figure 22.15. It comprises the following steps:

1. Calculate the specified effect size for a specific outcome metric and its variance from the extracted data. If several different outcome metrics are reported in the primary studies, you will need to analyse each outcome metric separately.

2. Tabulate the effect size, its variance and the number of observations per study.

3. Check whether any of the primary studies should be considered dependent replications (that is, primary studies performed by the same researchers, using the same subjects types, and materials). If any of the studies are dependent, you will need to decide how to incorporate these studies. This will involve either using a multi-level analysis model or integrating the dependent studies into a single large-scale experiment.

4. Import the data to an appropriate statistical tool, for example, the R *metafor* package, see Viechtbauer (2010).

5. Perform a random effects analysis to identify whether there is significant heterogeneity among the primary studies, see Viechtbauer (2007). Note, a fixed effects analysis is only appropriate when you have a very small number of effect sizes to aggregate or you are sure that the effect sizes all come from very close replications.

6. If heterogeneity is not significant, data synthesis is completed (subject to appropriate sensitivity analysis, as discussed below) and the overall mean and variance derived from aggregating the primary study results provide the best estimate of the difference between the treatment outcomes.

7. If there is significant heterogeneity, you will need to investigate whether any of the context factors and/or specific quality criteria might have influenced the outcomes and could be used as explanatory factors in a *moderator analysis*. This will require a mixed-effects analysis. Synthesis will depend on whether statistically significant explanatory factors (often referred to as *moderators*) can be found.

8. You will need to consider sensitivity analysis for example, assessing the impact of removing each primary study turn, the impact of high influence studies, and the impact of removing low quality primary studies.

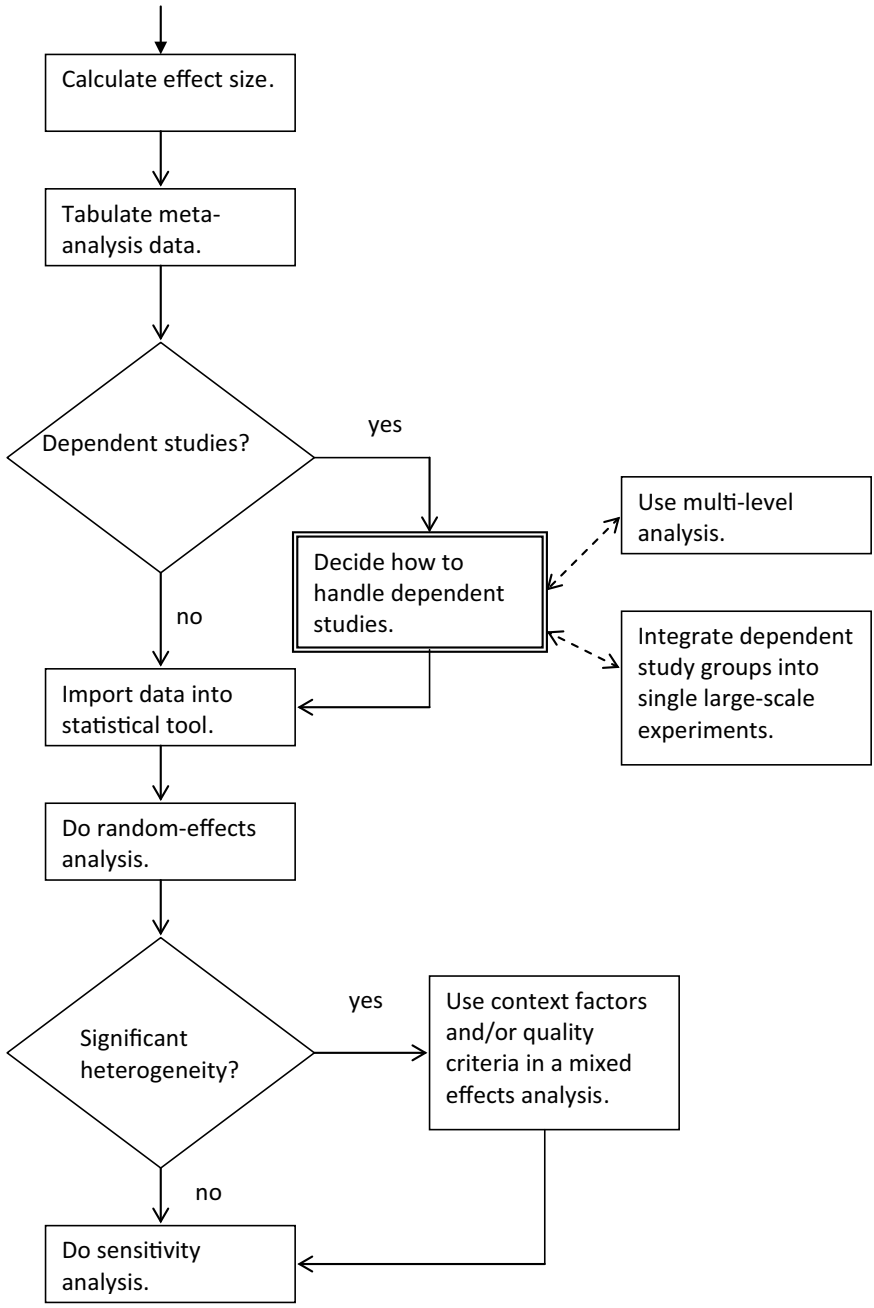

FIGURE 22.15: Meta-analysis process.

22.10.1.2 Reporting meta-analysis results

The R *metafor* package provides the standard meta-analysis graphics to report results (see Chapter 11). The most important graphics are:

- Forest plots to show the mean and variance of the effect size metric for each primary study as well as the overall mean and variance. An example of a forest plot is shown in Figure 22.16. Such plots can also be used to show the impact of significant moderating factors.
- Funnel plots to investigate publication bias (which is the tendency for only papers with significant results to be published). Funnel plots plot the effect size metric for each study against its standard error. The effect size of studies with large standard errors should be more varied than the effect size of studies with relatively small standard errors. The white funnel on the funnel plot shows the acceptable variation for the data points. The funnel plot application can also estimate the number of *missing* studies and display the funnel plot with estimated missing data points filled-in. An example of a funnel plot is shown Figure 22.17. It is based on the same data as the forest plot and includes three extra data points (shown by 3 white filled dots in the left-side of the plot) which correspond to estimated missing data points. The funnel plot also shows

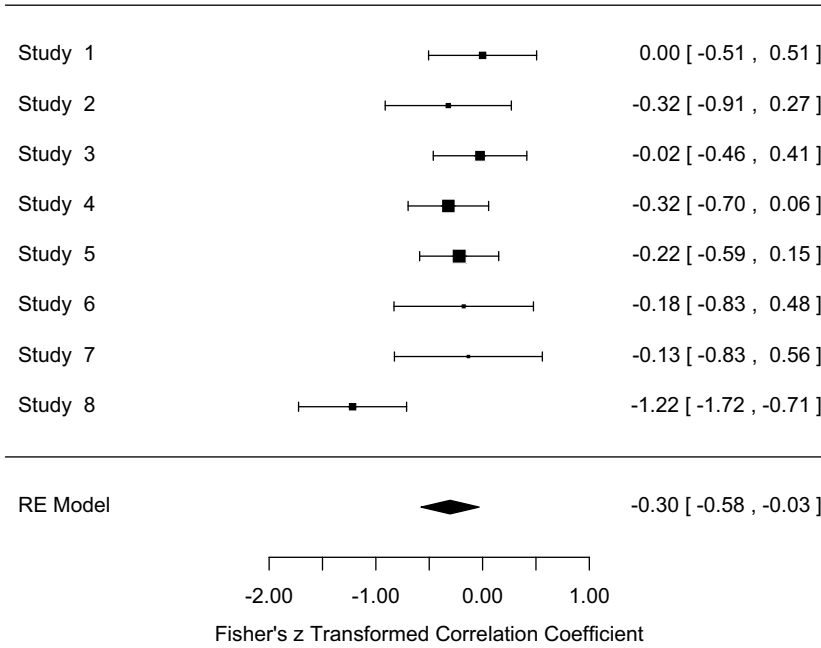

FIGURE 22.16: Forest plot example.

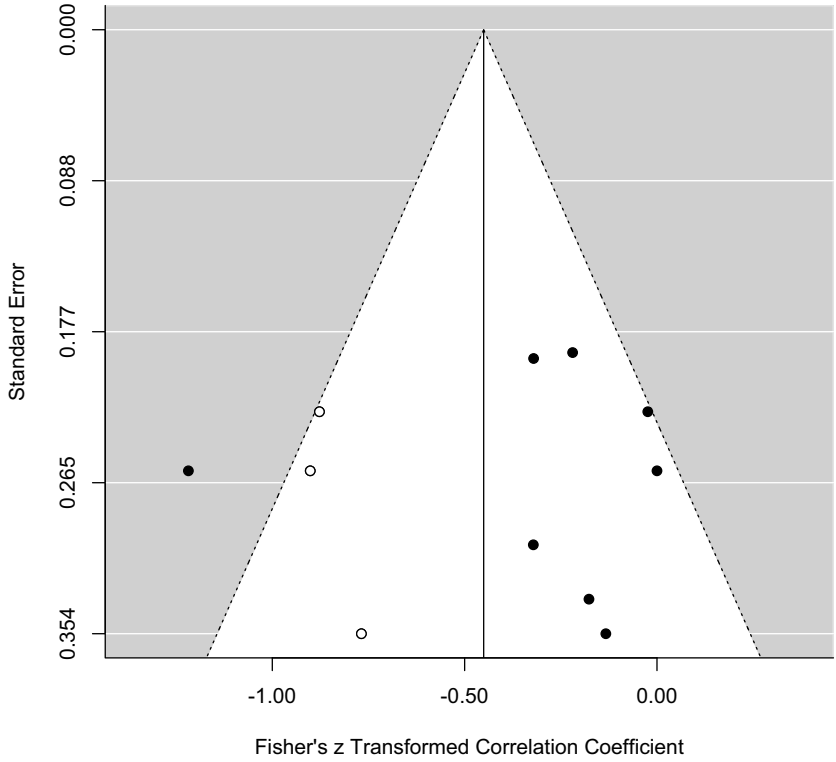

FIGURE 22.17: Funnel plot example.

that one data point is outside the acceptable range (that is, the black dot on the left-hand side of the plot).

Sensitivity analysis can be performed using the influential case diagnostic procedures. In addition Q-Q (Quantile-Quantile) plots can be used to investigate whether the distribution of the primary studies is normal.

22.10.1.3 Vote counting for quantitative systematic reviews

If you cannot do a full meta-analysis because there is too much diversity among the outcome measures or experimental methods, we recommend using a vote counting approach. Vote counting is based on counting the number of primary studies that found a significant affect, and if they constitute the majority, assuming that there is a true effect.

Most meta-analysts object to the use of vote counting for two reasons:

1. Assuming that a study is valid, a significant effect indicates a true effect. However, a non-significant effect does not indicate that there is no effect, because it can be due to low power.

2. Vote counting does not consider the size of effect, so it is not clear whether an effect is of practical importance as well as being significant.

However, in practice, the technique is useful for synthesizing software engineering studies, particularly when it is integrated with some form of moderator analysis. That is, you look for additional factors that might explain differences in primary study outcomes. This is similar to moderator analysis in a meta-analysis, but is qualitative rather than quantitative.

Before considering moderator factors, you need to define the outcome of each primary study. Popay et al. (2006) suggest a five-point scale to describe the outcome:

1. Significantly favours intervention
2. Trends towards intervention
3. No difference
4. Trends towards control
5. Significantly favours control.

They also recommend reporting any effect sizes that can be calculated, not the significance of the outcome.

Some of the synthesis methods used for qualitative primary studies support vote counting (see Chapter 10 and Table 22.3). In particular, the display methods used for Qualitative Cross-Case Analysis (Miles et al. 2014) can also be used to display the results of vote counting and qualitative moderator analysis. The outcomes of each primary study can be tabulated, together with the identified moderator factor values for the study. The display can be organised so that primary studies with the same outcome are kept together. Alternatively, if there is a good reason for displaying the results in a different order (for example, based on the date of the study), it is useful to colour-code the entries according to the outcome. You should then look for any trends among the moderator factors that seem consistent with favourable outcomes. In some cases, it may be possible to use more sophisticated methods such as Comparative Analysis (Ragin 1989) or Case Survey Analysis (Yin & Heald 1975) to analyse vote counting and moderator factor data.

22.10.2 Data synthesis for qualitative systematic reviews

As far as planning is concerned, you should specify in the protocol, the type of synthesis method you intend to use (see Chapter 10 and Table 22.3, for an overview of qualitative methods that have been used in software engineering studies).

In most cases, data synthesis will be integrated with data extraction. You should expect an iterative process whereby the results of reading, extracting and synthesizing data from some primary studies influence the data extraction

TABLE 22.3: Synthesis Methods for Qualitative Analysis

Type	Description
Narrative Synthesis (Popay et al. 2006)	The results and any trends are reported as a textual narrative. Narrative synthesis must be supported by tabulating results or it is very difficult to demonstrate traceability from research questions to the data, to aggregated results that answer the research questions.
Thematic Analysis (Thomas & Harden 2008)	Cruzes & Dybå (2011*a*) define a 5 stage process starting with reading the text and identifying specific segments of text. The segments of text are labelled and coded, then the codes are analysed to reduce overlaps and define themes. Themes are analysed to create higher-order themes and/or models of the phenomenon being studied. Note some themes are likely to be defined in advance as a result of the research questions, while others may arise as a result of reading the primary studies.
Comparative Analysis (Ragin 1989)	List and categorises cases and attempts to assess what inferences the data supports using boolean algebra. For example, looks for factors that are consistently associated with favourable outcomes AND not associated with unfavourable outcomes.
Meta-Ethnography (Noblit & Hare 1988)	7 stage process in which interpretations and explanations reported in the primary studies are translated into one another. Translations may result in agreement among studies, contradictions among studies, or may form parts of a coherent argument.
Case Survey (Yin & Heald 1975)	This is similar to Comparative Analysis but is appropriate when there are a large number primary studies. Individual primary study results and context information are classified and tabulated looking for commonalities and differences.
Qualitative Cross-case Analysis (Miles et al. 2014)	This uses matrices to report textual and quantitative information from each primary study. The matrices allow similarities and differences among primary studies to be identified.
Metasummary (Sandelowski et al. 2007)	This is a quantitatively oriented aggregation method for analysing thematic analysis papers and opinion surveys. Counts of themes such as risks, motivators, barriers to adoption are made on a primary study basis irrespective of the number of participants in each study.

and synthesis of subsequent primary studies and may initiate a re-assessment of some of the primary studies which have already been synthesised. If you are using a textual analysis tool you need to decide whether it will be used during the initial data extraction and synthesis process or as part of the validation process. The basic data extraction and qualitative synthesis process involves:

1. Identifying textual elements (which can be phrases, sentences, paragraphs, items in tables) in each primary study. The textual element should be stored in your data collection form (which can be a database, document or spreadsheet) with associated information identifying where in the document the element was found, and the research question(s) that it addresses.

2. Each textual element is coded, that is, allocated a single word or phrase that defines its content.

3. Codes are cross-checked for consistency across different primary studies and the data extracted by different team members.

4. Codes may be used for context analysis, that is, the frequency of occurrence of the individual codes are counted for research questions.

5. Codes may be used to create a model of the topic of interest by grouping codes together into related higher-level characteristics and themes. Then the relationships among those higher level characteristics and themes are investigated.

In these guidelines it is not possible to describe every approach to qualitative synthesis. You will find more detailed information in Chapter 10. We recommend reviewing existing software engineering systematic reviews that have used qualitative approaches. A good starting point is a paper written by Cruzes & Dybå (2011*b*) which cross-references software engineering systematic reviews to the qualitative synthesis methods they used. For specific qualitative methods, Cruzes & Dybå (2011*a*) provide a detailed explanation of thematic analysis, while Da Silva et al. (2013) present a worked example of using meta-ethnography to synthesise four primary studies. In addition, Cruzes et al. (2014) present an example of synthesising two case studies using three different methods: thematic analysis, qualitative cross case analysis and narrative analysis.

22.10.3 Data aggregation for mapping studies

Mapping study data collection involves identifying important features that describe the characteristics of the primary studies and identifying the appropriate metrics to measure those features. Mapping study aggregation involves tabulating the primary study features. In the case of nominal and ordinal scale features, you should count the number of primary studies in the different categories. For numerical features, you should use standard statistical measures of

location and scale (for example, mean and standard deviations, and graphical representations such as box plots or histograms). It is often useful to represent the data in two-way tables that show the relationship between two different categorical features.

For features represented as nominal and ordinal scale metrics, two graphical representation are particularly useful:

1. Trend plots that have counts of primary studies in a specific category as the y-variable and a year as the x-variable. It is sometimes useful to have more than one y-variable. For example, trend plots might be used to show the number of primary studies of different types per year.

2. Bubble plots that are graphical representations which allow you to view information from two two-way tables on the same diagram when the tables share a nominal scale measure. An example is shown in Figure 22.18. In this diagram, the y-variable is called "Variability Context Factor" and has six categories. There are two x-variables, one called "Contribution Facet" which has five categories, and the other called "Research Facet" which has six categories. The bubble containing the value 21 identifies that 21 studies were categorised as having a y-category of "Requirement Variability" and a "Research Focus" category of "Solution". The diagram shows the total number of primary studies classified by "Research Facet" was 128 (which is the number next to the name of the x-variable) and that 21 primary studies corresponds to 16.4% of those studies. The fact that there are different numbers of primary studies classified against each variable indicates that either that some primary studies were not classified against one of the x-valuables or some studies were classified in more than one category for the same x-variable. Note this is not a three-dimensional table because it does not show the distribution of the third nominal variable conditional on the values of the two other variables.

More information about synthesis for mapping studies can be found in Chapter 9.

22.10.3.1 Tables versus graphics

Although graphical representations of the data are important for showing the distribution of primary studies, it is also important to tabulate the results. Without a table showing the values of the categories for each primary study (or a publicly available on-line supporting database), other researchers cannot make constructive use of the results of a mapping study.

22.10.4 Data synthesis validation

There are no standard validation procedures for data synthesis. You must aim to ensure that there is a clear link from the research questions to the data and then to the syntheses that answers those research questions.

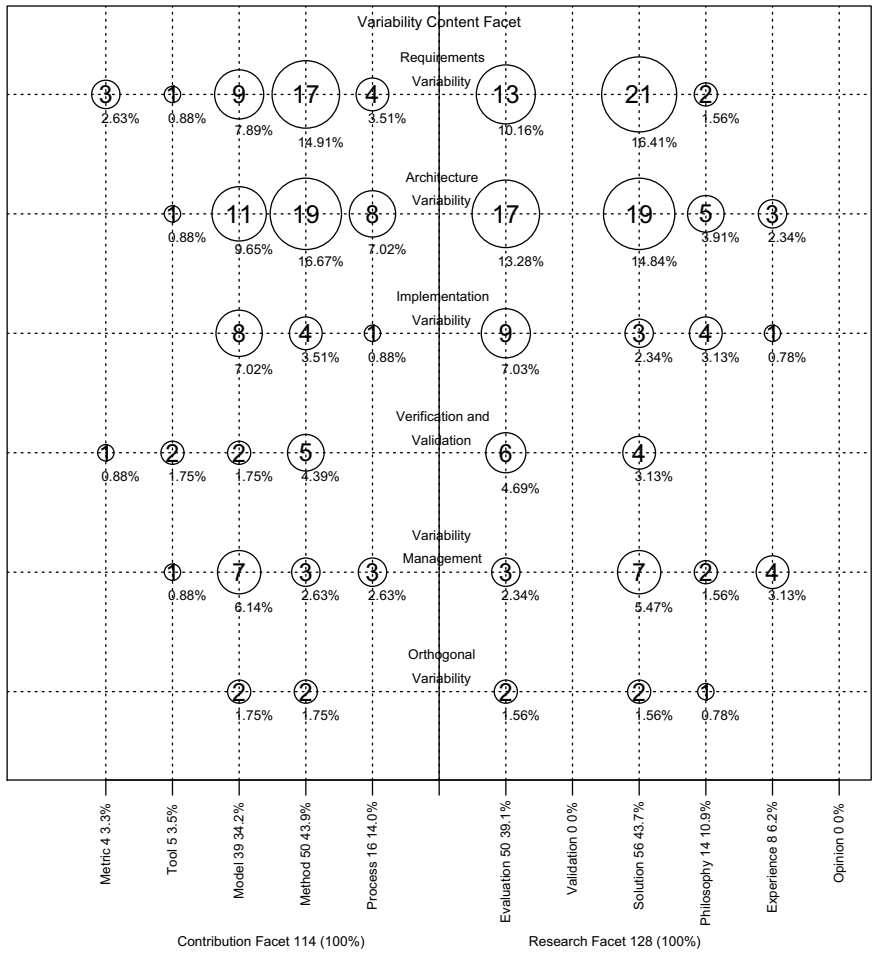

FIGURE 22.18: Bubbleplot example.

General points to remember

- Quantitative systematic reviews can sometimes be analysed using meta-analysis but it may still be necessary to use qualitative synthesis if individual primary studies use non-comparable measurements or treatments.
- Qualitative synthesis is difficult. You should aim to maintain traceability from the data to the synthesis.
- Mapping study aggregation is relatively straightforward.
- If a mapping study is intended for publication in a conference or journal papers, you must make sure that all the primary study data including

the citations and classification information is available to the reader. For very large mapping studies, this may mean publishing an on-line database holding the citations and feature data for each primary study.

22.11 Reporting the systematic review

There are three things you need consider when reporting your results:

1. Who do you expect to be interested in the results of the systematic review and what format of the report(s) do they need?
2. The format of each type of report you need to write.
3. How you plan to validate the report.

22.11.1 Systematic review readership

Systematic reviews (in contrast to mapping studies) should consider two main types of reader: researchers and practitioners. A reader of an academic journal or conference paper will expect to see a full description of the methodology, as well as a report of the results of the study with traceability between the data and the analysis. Practitioners, however, will be more concerned about the implication of the result for software engineering practice. It is also the case that practitioners are more likely to read short magazine articles than ponderous academic papers.

Thus, there is an argument for writing both an academic paper and a separate article for practitioners. Remember, however, to reference any related journal paper in the practitioner article and vice versa. You also need to ensure that you do not violate any originality or copyright requirements of the publication outlets.

For mapping studies, the main readership will be researchers. It is important to include information about each primary study including the full citations, as well as how it was classified against each feature. For conference papers, it may be difficult to fit in all the data for each primary study. In such cases you need to consider providing ancillary data on-line (either an extended technical report or a database).

22.11.2 Report structure

With respect to the structure of a systematic review report, PRISMA, the current guideline for reporting systematic reviews and meta-analysis studies in health care, has been published in several open-access papers (one example being Liberati et al. (2009)).

The high-level structure for a PRISMA report is very similar to the standard scientific format which comprises the Title and Abstract followed by the IMRAD sections (Introduction, Methods, Results, and Discussion). In software engineering, we usually have a separate *Conclusions* section but PRISMA makes *Conclusions* a subset of the *Discussion* section and adds a final section called *Funding*.

We recommend using the basic PRISMA structure with a few minor changes:

1. Title: Identify the topic of the study and that it is a systematic review, meta-analysis or mapping study.

2. Abstract: Structured abstract, including headings for background, objectives, methods, results, conclusions.

3. Introduction: Justification for the review and the research question(s).

4. Background: Any information needed to understand the topic of the systematic review.

5. Methods: Indicate where the protocol can be accessed. Report search and selection process including databases searched, search terms and inclusion and exclusion criteria - consider what needs to be in the body of the paper and what can be put into appendices. Report data collection process and data items collected including quality data and the data items needed to answer the research questions. For meta-analysis, identify the principal summary measures and methods of aggregation. For other forms of review, describe the data analysis and synthesis process. Report how quality data will be used and any other means of identifying possible biases. Describe any additional analysis such as sensitivity analysis or subgroup analysis.

6. Results: Report the study selection process including number of studies excluded at each major stage, preferably with a flow diagram. Report the identified primary study characteristics. Discuss the quality of individual studies and any systematic biases. Present data syntheses using the graphical methods discussed in Section 22.10. Report the results of any additional analyses.

7. Discussion: provide a summary the evidence for each research question and any additional analyses. Discuss the limitations of the review at the primary study level and at the review level.

8. Conclusions: Provide a general interpretation of the result in the context of any other evidence. Provide recommendations for researchers and practitioners.

9. Acknowledgements: Identify the funding agency (if any). Thank any researchers who made useful contributions to the study but were not part of the research team, for example external reviewers of the protocol or final report.

10. Appendices: Report the search strings used for individual digital sources. Specify the data collection form. Provide a list of "near miss" papers, for example, candidate primary studies excluded from the review after the second screening process. Report any quality checklists used.

If you are intending to report your results in a practitioner magazine, you should expect to use a much less formal presentation. You should aim to explain the topic covered by the review and why it is important. It is unnecessary to provide any detailed description of the methodology. When presenting the findings you should explain what the implications are for practice and the confidence that can be placed in the findings.

22.11.3 Validating the report

All report authors have a responsibility to read and review the report, with the aim of ensuring:

- The research questions are clearly specified and fully answered.
- The research methodology is fully and correctly reported.
- There is traceability from the research questions to data collection, data synthesis and conclusions.
- All the tables and figures used to present the results are correct and internally consistent.
- In the case of systematic reviews, the conclusions are written clearly and are targeted both at researchers and practitioners.

If possible, for systematic reviews, you should find someone to act as an independent reviewer of the report. For example, within a research group or a university department, you can try to encourage colleagues to undertake independent reviews on a quid-pro-quo basis. An independent reviewer would probably find useful the following questions, which are based on review assessment questions suggested by Greenhalgh (2010):

- Q1: Does the review address an important software engineering problem?
- Q2: Was a thorough search done of the appropriate databases(s) and were other sources considered?
- Q3: Was methodological quality assessed and primary studies weighted accordingly?

- Q4: How sensitive are the results to the way the review was done?
- Q5: Have numerical results been interpreted sensibly in the context of the problem?

Points to remember

- Identify your target audience and write appropriately.
- Make sure to report your methodology fully.
- For reviews including a large number primary studies, consider providing an online database to hold the information collected about each primary study.
- Particularly for systematic reviews, consider the implications for practitioners.
- All authors need to review the final report carefully.
- For systematic reviews, an independent reviewer can be very helpful.
- For journal or conference papers reporting mapping studies, you may need to provide citation and classification information in a separate technical report or online database.