Guidelines for performing Systematic Literature Reviews in Software Engineering

Version 2.3

EBSE Technical Report

Software Engineering Group
School of Computer Science and Mathematics
Keele University
Keele, Staffs
ST5 5BG, UK

and

Department of Computer Science
University of Durham
Durham,
UK

9 July, 2007

0. Document Control Section

0.1 Contents

0.2 Document Version Control

Document status	Version Number	Date	Changes from previous version
Draft	0.1	1 April 2004	None
Published	1.0	29 June 2004	Correction of typos Additional discussion of problems of assessing evidence Section 7 “Final Remarks” added.
Revision	1.1	17 August 2005	Corrections of typos.
Major Revision	1.9	25 October 2005	Changed title, added SMC as a reviser, several new sections added, finalisation of major revisions should be version 2.0 Changes summarised below: Added Section 2 – to but EBSE in Context Expanded the reporting of the review processes Added sections on Systematic Mapping and Tertiary reviews in section 4 Updated the Reporting the Review Section Added two final sections, Systematic Mapping Studies and Tertiary Reviews
Further major revisions	2.0	17 March 2007	Revised the section on hierarchy of studies to be consistent with social science viewpoints. Removed some general discussion that was not well-focused on the construction of the guidelines. Revised the section on quality checklists Added lessons learnt from SE articles. Removed final section on Tertiary reviews (seemed unnecessary)
Minor revisions after internal	2.1	27 March 2007	Correction of Typos Inclusion of a Glossary Inclusion of guidelines

review			construction process Minor restructuring – Mapping reviews and tertiary reviews moved into section 3 to avoid interfering with the flow of the guidelines.
Further minor revisions	2.2	4 April 2007	Typos and grammatical corrections. A paragraph on how to read the guidelines included in the Introduction.
Revisions after external review	2.3	20 July	Amendments after external review including the introduction of more examples.

0.3 Document development team

This document was revised by members of the Evidence-Based Software Engineering(EBSE) Project (EP/CS51839/X) which was funded by the UK Economics and Physical Sciences Research Council.

Name	Affiliation	Role
Barbara Kitchenham	Keele University, UK	Lead author
Stuart Charters	Lincoln University, NZ	Second author
David Budgen	University of Durham, UK	EBSE Internal Reviewer
Pearl Brereton	Keele University, UK	EBSE Internal Reviewer
Mark Turner	Keele University, UK	EBSE Internal Reviewer
Steve Linkman	Keele University, UK	EBSE Internal Reviewer
Magne Jørgensen	Simula Research Laboratory, Norway	External reviewer
Emilia Mendes	University of Auckland, New Zealand	External reviewer
Giuseppe Visaggio	University of Bari, Italy	External reviewer

0.4 Executive Summary

The objective of this report is to propose comprehensive guidelines for systematic literature reviews appropriate for software engineering researchers, including PhD students. A systematic literature review is a means of evaluating and interpreting all available research relevant to a particular research question, topic area, or phenomenon of interest. Systematic reviews aim to present a fair evaluation of a research topic by using a trustworthy, rigorous, and auditable methodology.

The guidelines presented in this report were derived from three existing guidelines used by medical researchers, two books produced by researchers with social science backgrounds and discussions with researchers from other disciplines who are involved in evidence-based practice. The guidelines have been adapted to reflect the specific problems of software engineering research.

The guidelines cover three phases of a systematic literature review: planning the review, conducting the review and reporting the review. They provide a relatively high level description. They do not consider the impact of the research questions on the review procedures, nor do they specify in detail the mechanisms needed to perform meta-analysis.

0.5 Glossary

Meta-analysis. A form of secondary study where research synthesis is based on quantitative statistical methods.

Primary study. (In the context of evidence) An empirical study investigating a specific research question.

Secondary study. A study that reviews all the primary studies relating to a specific research question with the aim of integrating/synthesising evidence related to a specific research question.

Sensitivity analysis. An analysis procedure aimed at assessing whether the results of a systematic literature review or a meta-analysis are unduly influenced by a small number of studies. Sensitivity analysis methods involve assessing the impact of high leverage studies (e.g. large studies or studies with atypical results), and ensuring that overall results of a systematic literature remain the same if low quality studies (or high quality) studies are omitted from the analysis, or analysed separately.

Systematic literature review (also referred to as a systematic review). A form of secondary study that uses a well-defined methodology to identify, analyse and interpret all available evidence related to a specific research question in a way that is unbiased and (to a degree) repeatable.

Systematic review protocol. A plan that describes the conduct of a proposed systematic literature review.

Systematic mapping study (also referred to as a scoping study). A broad review of primary studies in a specific topic area that aims to identify what evidence is available on the topic.

Tertiary study (also called a tertiary review). A review of secondary studies related to the same research question.

1. Introduction

This document presents general guidelines for undertaking systematic reviews. The goal of this document is to introduce the methodology for performing rigorous reviews of current empirical evidence to the software engineering community. It is aimed primarily at software engineering researchers including PhD students. It does not cover details of meta-analysis (a statistical procedure for synthesising quantitative results from different studies), nor does it discuss the implications that different types of systematic review questions have on research procedures.

The original impetus for employing systematic literature review practice was to support evidence-based medicine, and many guidelines reflect this viewpoint. This document attempts to construct guidelines for performing systematic literature reviews that are appropriate to the needs of software engineering researchers. It discusses a number of issues where software engineering research differs from medical research. In particular, software engineering research has relatively little empirical research compared with the medical domain; research methods used by software engineers are not as generally rigorous as those used by medical researchers; and much empirical data in software engineering is proprietary.

1.1 Source Material used in the Construction of the Guidelines

The document is based on a review of three existing guidelines for systematic reviews, the experiences of the Keele University and University of Durham Evidence-based Software Engineering project, meetings with domain experts in a variety of disciplines interested in evidence-based practice, and text books describing systematic review principles:

- The Cochrane Reviewer's Handbook [7] and Glossary [8].
- Guidelines prepared by the Australian National Health and Medical Research Council [1] and [2].
- Centre for Reviews and Dissemination (CRD) Guidelines for those carrying out or commissioning reviews [19].
- Systematic reviews in the Social Sciences: A Practical Guide, Mark Petticrew and Helen Roberts [25]
- Conducting Research Literature Reviews. From the Internet to Paper, 2nd Edition, Arlene Fink [11].
- Various articles and texts describing procedures for literature reviews in medicine and social sciences ([20], [13], and [24]).
- Meetings with various domain experts and centres including, the Evidence for Policy and Practice Information and Coordinating Centre (EPPI Centre http://eppi.ioe.ac.uk/cms/) Social Science Research Unit Institute of Education, University of London; CRD York University, Mark Petticrew, Glasgow University; Andrew Booth, Sheffield University
- Experiences from the Evidence Based Software Engineering Project at Keele University and Durham University.

In particular, this document owes much to the CRD Guidelines.

1.2 The Guideline Construction Process

The construction process used for the guidelines was:

- The guidelines were originally produced by a single person (Kitchenham).
- They were then updated by two people (Charters and Kitchenham).
- They were reviewed by members of the Evidence-based Software Engineering project (Brereton, Budgen, Linkman, and Turner).
- After correction, the guidelines were then circulated to external experts for independent review.
- The guidelines were further amended after the review by the external experts.

1.3 The Structure of the Guidelines

The structure of the guidelines is as follows:

- Section 2 provides an introduction to systematic reviews.
- Section 3 explains why social science SLR methodology is appropriate in the context of software engineering research.
- Section 4 specifies the stages in a systematic review.
- Section 5 discusses the planning stages of a systematic review.
- Section 6 discusses the stages involved in conducting a systematic review.
- Section 7 discusses reporting a systematic review.
- Section 8 discusses systematic mapping studies.

Throughout the guidelines we have incorporated examples taken from two recently published systematic literature reviews [21] and [17]. Kichenham et al. [21] addressed the issue of whether it was possible to use cross-company benchmarking datasets to produce estimation models suitable for use in a commercial company. Jørgensen [17] investigated the use of expert judgement, formal models and combinations of the two approaches when estimating software development effort. In addition, Appendix 2 provides a list published systematic literature reviews assessed as high quality by the authors of this report. These SLRs were identified and assessed as part of a systematic literature review of recent software engineering SLRs. The protocol for the review is documented in Appendix 3.

1.4 How to Use the Guidelines

These guidelines are aimed at software engineering researchers, PhD students, and practitioners who are new to the concept of performing systematic literature reviews. Readers who are unsure about what a systematic literature review is should start by reading Section 2.

Readers who understand the principles of a systematic literature review can skip to Section 4 to get an overview of the systematic literature review process. They should then concentrate on Sections 5, 6 and 7, which describe in detail how to perform each review phase. Sections 3 and 8 provide ancillary information that can be omitted on first reading.

Readers who have more experience in performing systematic reviews may find the list of tasks in Section 4, the quality checklists in Tables 5 and 6 and the reporting structure presented in Table 7 sufficient for their needs.

Readers with detailed methodological queries are unlikely to find answers in this document. They may find some of the references useful.

2. Systematic Literature Reviews

A systematic literature review (often referred to as a systematic review) is a means of identifying, evaluating and interpreting all available research relevant to a particular research question, or topic area, or phenomenon of interest. Individual studies contributing to a systematic review are called *primary* studies; a systematic review is a form of *secondary* study.

2.1 Reasons for Performing Systematic Literature Reviews

There are many reasons for undertaking a systematic literature review. The most common reasons are:

- To summarise the existing evidence concerning a treatment or technology e.g. to summarise the empirical evidence of the benefits and limitations of a specific agile method.
- To identify any gaps in current research in order to suggest areas for further investigation.
- To provide a framework/background in order to appropriately position new research activities.

However, systematic literature reviews can also be undertaken to examine the extent to which empirical evidence supports/contradicts theoretical hypotheses, or even to assist the generation of new hypotheses (see for example [14]).

2.2 The Importance of Systematic Literature Reviews

Most research starts with a literature review of some sort. However, unless a literature review is thorough and fair, it is of little scientific value. This is the main rationale for undertaking systematic reviews. A systematic review synthesises existing work in a manner that is fair and seen to be fair. For example, systematic reviews must be undertaken in accordance with a predefined search strategy. The search strategy must allow the completeness of the search to be assessed. In particular, researchers performing a systematic review must make every effort to identify and report research that does not support their preferred research hypothesis as well as identifying and reporting research that supports it.

> *"Indeed, one of my major complaints about the computer field is that whereas Newton could say,* ***"If I have seen a little farther than others, it is because I have stood on the shoulders of giants,"*** *I am forced to say, "Today we stand on each other's feet." Perhaps the central problem we face in all of computer science is how we are to get to the situation where we build on top of the work of others rather than redoing so much of it in a trivially different way. Science is supposed to be cumulative, not almost endless duplication of the same kind of things".*
>
> Richard Hamming 1968 Turning Award Lecture

Systematic literature reviews in all disciplines allow us to stand on the shoulders of giants and in computing, allow us to get off each others' feet.

2.3 Advantages and disadvantages

The advantages of systematic literature reviews are that:

- The well-defined methodology makes it less likely that the results of the literature are biased, although it does not protect against publication bias in the primary studies.
- They can provide information about the effects of some phenomenon across a wide range of settings and empirical methods. If studies give consistent results, systematic reviews provide evidence that the phenomenon is robust and transferable. If the studies give inconsistent results, sources of variation can be studied.
- In the case of quantitative studies, it is possible to combine data using meta-analytic techniques. This increases the likelihood of detecting real effects that individual smaller studies are unable to detect.

The major disadvantage of systematic literature reviews is that they require considerably more effort than traditional literature reviews. In addition, increased power for meta-analysis can also be a disadvantage, since it is possible to detect small biases as well as true effects.

2.4 Features of Systematic Literature Reviews

Some of the features that differentiate a systematic review from a conventional expert literature review are:

- Systematic reviews start by defining a review protocol that specifies the research question being addressed and the methods that will be used to perform the review.
- Systematic reviews are based on a defined search strategy that aims to detect as much of the relevant literature as possible.
- Systematic reviews document their search strategy so that readers can assess their rigour and the completeness and repeatability of the process (bearing in mind that searches of digital libraries are almost impossible to replicate).
- Systematic reviews require explicit inclusion and exclusion criteria to assess each potential primary study.
- Systematic reviews specify the information to be obtained from each primary study including quality criteria by which to evaluate each primary study.
- A systematic review is a prerequisite for quantitative meta-analysis.

2.5 Other Types of Review

There are two other types of review that complement systematic literature reviews: systematic mapping studies and tertiary reviews.

2.5.1 Systematic Mapping Studies

If, during the initial examination of a domain prior to commissioning a systematic review, it is discovered that very little evidence is likely to exist or that the topic is

very broad then a systematic mapping study may be a more appropriate exercise than a systematic review.

A systematic mapping study allows the evidence in a domain to be plotted at a high level of granularity. This allows for the identification of evidence clusters and evidence deserts to direct the focus of future systematic reviews and to identify areas for more primary studies to be conducted. An outline of the systematic mapping study process highlighting the main differences from the standard systematic review process can be found in Section 8.

2.5.2 Tertiary Reviews

In a domain where a number of systematic reviews exist already it may be possible to conduct a tertiary review, which is a systematic review of systematic reviews, in order to answer wider research questions. A tertiary review uses exactly the same methodology as a standard systematic literature review. It is potentially less resource intensive than conducting a new systematic review of primary studies but is dependent on sufficient systematic reviews of a high quality being available. The protocol presented in Appendix 3 is a protocol for a tertiary review.

3. Evidence Based Software Engineering in Context

It is important to understand the relationship of Software Engineering to other domains with regard to the applicability of the Evidence Based paradigm. In doing so, we can identify how procedures adopted from other disciplines (particularly medicine) need to be adapted to suit software engineering research and practice.

Budgen et al. [6] interviewed practitioners in a number of domains that use evidence based approaches to research, and compared their research practices with those of software engineering. Table 1 shows the results of their assessment of the similarity between software engineering research practices and those of other domains. It shows that software engineering is much more similar to the Social Sciences than it is to medicine. This similarity is due to experimental practices, subject types and blinding procedures. Within Software Engineering it is difficult to conduct randomised controlled trials or to undertake double blinding. In addition, human expertise and the human subject all affect the outcome of experiments.

Table 1 Comparing Software Engineering experimental methodology with that of other disciplines

Discipline	Comparison with SE (1 is perfect agreement, 0 is complete disagreement)
Nursing & Midwifery	0.83
Primary Care	0.33
Organic Chemistry	0.83
Empirical Psychology	0.66
Clinical Medicine	0.17
Education	0.83

These factors mean that software engineering is significantly different from the traditional medical arena in which systematic reviews were first developed. For this

reason we have revised these guidelines to incorporate recent ideas from the area of social science ([25], [11]). In addition, the choice of references on which to base these guidelines was informed by our discussions with researchers in these disciplines.

4. The Review Process

A systematic literature review involves several discrete activities. Existing guidelines for systematic reviews have slightly different suggestions about the number and order of activities (see Appendix 1). However, the medical guidelines and sociological text books are broadly in agreement about the major stages in the process.

This document summarises the stages in a systematic review into three main phases: Planning the Review, Conducting the Review, Reporting the Review.

The stages associated with *planning the review* are:

- Identification of the need for a review (See Section 5.1).
- Commissioning a review (See Section 5.2).
- Specifying the research question(s) (See Section 5.3).
- Developing a review protocol (See Section 5.4).
- Evaluating the review protocol (See Section 5.5).

The stages associated with *conducting the review* are:

- Identification of research (See Section 6.1).
- Selection of primary studies (See Section 6.2).
- Study quality assessment (See Section 6.3).
- Data extraction and monitoring (See Section 6.4).
- Data synthesis (See Section 6.5).

The stages associated with *reporting the review* are:

- Specifying dissemination mechanisms (See Section 7.1).
- Formatting the main report (See Section 7.2).
- Evaluating the report (See Section 7.3).

We consider all the above stages to be mandatory except:

- Commissioning a review which depends on whether or not the systematic review is being done on a commercial basis.
- Evaluating the review protocol (5.5) and Evaluating the report (7.3) which are optional and depend on the quality assurance procedures decided by the systematic review team (and any other stakeholders).

The stages listed above may appear to be sequential, but it is important to recognise that many of the stages involve iteration. In particular, many activities are initiated during the protocol development stage, and refined when the review proper takes place. For example:

- The selection of primary studies is governed by inclusion and exclusion criteria. These criteria are initially specified when the protocol is drafted but may be refined after quality criteria are defined.
- Data extraction forms initially prepared during construction of the protocol will be amended when quality criteria are agreed.

- Data synthesis methods defined in the protocol may be amended once data has been collected.

The systematic reviews road map prepared by the Systematic Reviews Group at Berkeley demonstrates the iterative nature of the systematic review process very clearly [24].

5. Planning

Prior to undertaking a systematic review it is necessary to confirm the need for such a review. In some circumstances systematic reviews are commissioned and in such cases a commissioning document needs to be written. However, the most important pre-review activities are defining the research questions(s) that the systematic review will address and producing a review protocol (i.e. plan) defining the basic review procedures. The review protocol should also be subject to an independent evaluation process. This is particularly important for a commissioned review.

5.1 The need for a systematic review

The need for a systematic review arises from the requirement of researchers to summarise all existing information about some phenomenon in a thorough and unbiased manner. This may be in order to draw more general conclusions about some phenomenon than is possible from individual studies, or may be undertaken as a prelude to further research activities.

Examples

Kitchenham et al. [21] argued that accurate cost estimation is important for the software industry; that accurate cost estimation models rely on past project data; that many companies cannot collect enough data to construct their own models. Thus, it is important to know whether models developed from data repositories can be used to predict costs in a specific company. They noted that a number of studies have addressed that issue but have come to different conclusions. They concluded that it is necessary to determine whether, or under what conditions, models derived from data repositories can support estimation in a specific company.

Jørgensen [17] pointed out in spite of the fact that most software cost estimation research concentrates on formal cost estimation models and that a large number of IT managers know about tools that implement formal models, most industrial cost estimation is based on expert judgement. He argued that researchers need to know whether software professionals are simply irrational, or whether expert judgement is just as accurate as formal models or has other advantages that make it more acceptable than formal models.

In both cases the authors had undertaken research in the topic area and had first hand knowledge of the research issues.

Prior to undertaking a systematic review, researchers should ensure that a systematic review is necessary. In particular, researchers should identify and review any existing systematic reviews of the phenomenon of interest against appropriate evaluation criteria. The CRD [19] suggests the following checklist:

- What are the review's objectives?
- What sources were searched to identify primary studies? Were there any restrictions?

- What were the inclusion/exclusion criteria and how were they applied?
- What criteria were used to assess the quality of primary studies?
- How were quality criteria applied?
- How were the data extracted from the primary studies?
- How were the data synthesised?
- How were differences between studies investigated?
- How were the data combined?
- Was it reasonable to combine the studies?
- Do the conclusions flow from the evidence?

The CRD Database of Abstracts of Reviews of Effects (DARE) criteria (http://www.york.ac.uk/inst/crd/crddatabases.htm#DARE) are even simpler. They are based on four questions:

1. Are the review's inclusion and exclusion criteria described and appropriate?
2. Is the literature search likely to have covered all relevant studies?
3. Did the reviewers assess the quality/validity of the included studies?
4. Were the basic data/studies adequately described?

Examples

We applied the DARE criteria both to Kitchenham et al.'s study [21] and to Jørgensen's study [17]. We gave Kitchenham et al.'s study a score of 4 and Jørgensen's study a score of 3.5. Other studies scored using the DARE criteria are listed in Appendix 2.

From a more general viewpoint, Greenlaugh [12] suggests the following questions:

- Can you find an important clinical question, which the review addressed? (Clearly, in software engineering, this should be adapted to refer to an important software engineering question.)
- Was a thorough search done of the appropriate databases and were other potentially important sources explored?
- Was methodological quality assessed and the trials weighted accordingly?
- How sensitive are the results to the way that the review has been done?
- Have numerical results been interpreted with common sense and due regard to the broader aspects of the problem?

5.2 Commissioning a Review

Sometimes an organisation requires information about a specific topic but does not have the time or expertise to perform a systematic literature itself. In such cases it will commission researchers to perform a systematic literature review of the topic. When this occurs the organisation must produce a commissioning document specifying the work required.

A commissioning document will contain or consider the following items (adapted from the CRD guidelines [12])

- Project Title
- Background
- Review Questions

- Advisory/Steering Group Membership (Researchers, Practitioners, Lay members, Policy Makers etc)
- Methods of the review
- Project Timetable
- Dissemination Strategy
- Support Infrastructure
- Budget
- References

The commissioning document can be used both to solicit tenders from research groups willing to undertake the review and to act as a steering document for the advisory group to ensure the review remains focused and relevant in the context.

The commissioning phase of a systematic review is not required for a research team undertaking a review for their own needs or for one being undertaken by a PhD student. If the commissioning stage is not undertaken then the dissemination strategy should be incorporated into the review protocol. As yet, there are no examples of commissioned SLRs in the software engineering domain.

5.3 The Research Question(s)

Specifying the research questions is the most important part of any systematic review. The review questions drive the entire systematic review methodology:

- The search process must identify primary studies that address the research questions.
- The data extraction process must extract the data items needed to answer the questions.
- The data analysis process must synthesise the data in such a way that the questions can be answered.

5.3.1 Question Types

The most important activity during planning is to formulate the research question(s). The Australian NHMR Guidelines [1] identify six types of health care questions that can be addressed by systematic reviews:

1. Assessing the effect of intervention.
2. Assessing the frequency or rate of a condition or disease.
3. Determining the performance of a diagnostic test.
4. Identifying aetiology and risk factors.
5. Identifying whether a condition can be predicted.
6. Assessing the economic value of an intervention or procedure.

In software engineering, it is not clear what the equivalent of a diagnostic test would be, but the other questions can be adapted to software engineering issues as follows:

- Assessing the effect of a software engineering technology.
- Assessing the frequency or rate of a project development factor such as the adoption of a technology, or the frequency or rate of project success or failure.
- Identifying cost and risk factors associated with a technology.
- Identifying the impact of technologies on reliability, performance and cost models.

- Cost benefit analysis of employing specific software development technologies or software applications.

Medical guidelines often provide different guidelines and procedures for different types of question. This document does not go to this level of detail.

The critical issue in any systematic review is to ask the right question. In this context, the right question is usually one that:

- Is meaningful and important to practitioners as well as researchers. For example, researchers might be interested in whether a specific analysis technique leads to a significantly more accurate estimate of remaining defects after design inspections. However, a practitioner might want to know whether adopting a specific analysis technique to predict remaining defects is more effective than expert opinion at identifying design documents that require re-inspection.
- Will lead either to changes in current software engineering practice or to increased confidence in the value of current practice. For example, researchers and practitioners would like to know under what conditions a project can safely adopt agile technologies and under what conditions it should not.
- Will identify discrepancies between commonly held beliefs and reality.

Nonetheless, there are systematic reviews that ask questions that are primarily of interest to researchers. Such reviews ask questions that identify and/or scope future research activities. For example, a systematic review in a PhD thesis should identify the existing basis for the research student's work and make it clear where the proposed research fits into the current body of knowledge.

Examples

Kitchenham et al. [21] had three research questions:

Question 1: What evidence is there that cross-company estimation models are not significantly different from within-company estimation models for predicting effort for software/Web projects?

Question 2: What characteristics of the study data sets and the data analysis methods used in the study affect the outcome of within- and cross-company effort estimation accuracy studies?

Question 3: Which experimental procedure is most appropriate for studies comparing within- and cross-company estimation models?

Jørgensen [17] had two research questions:

1. Should we expect more accurate effort estimates when applying expert judgment or models?
2. When should software development effort estimates be based on expert judgment, when on models, and when on a combination of expert judgment and models?

In both cases, the authors were aware from previous research that results were mixed, so in each case they added a question aimed at investigating the conditions under which different results are obtained.

5.3.2 Question Structure

Medical guidelines recommend considering a question about the effectiveness of a treatment from three viewpoints:

- The population, i.e. the people affected by the intervention.

- The interventions, which are usually a comparison between two or more alternative treatments.
- The outcomes, i.e. the clinical and economic factors that will be used to compare the interventions.

More recently Petticrew and Roberts suggest using the PICOC (Population, Intervention, Comparison, Outcome, Context) criteria to frame research questions [25]. These criteria extend the original medical guidelines with:
Comparison: I.e. what is the intervention being compared with
Context: i.e. what is the context in which the intervention is delivered.

In addition, study designs appropriate to answering the review questions may be identified and used to guide the selection of primary studies.

We discuss these criteria from the viewpoint of software engineering below.

Population
In software engineering experiments, the populations might be any of the following:
- A specific software engineering role e.g. testers, managers.
- A category of software engineer, e.g. a novice or experienced engineer.
- An application area e.g. IT systems, command and control systems.
- An industry group such as Telecommunications companies, or Small IT companies.

A question may refer to very specific population groups e.g. novice testers, or experienced software architects working on IT systems. In medicine the populations are defined in order to reduce the number of prospective primary studies. In software engineering far fewer primary studies are undertaken, thus, we may need to avoid any restriction on the population until we come to consider the practical implications of the systematic review.

Intervention
The intervention is the software methodology/tool/technology/procedure that addresses a specific issue, for example, technologies to perform specific tasks such as requirements specification, system testing, or software cost estimation.

Comparison
This is the software engineering methodology/tool/technology/procedure with which the intervention is being compared. When the comparison technology is the conventional or commonly-used technology, it is often referred to as the “control” treatment. The control situation must be adequately described. In particular “not using the intervention” is inadequate as a description of the control treatment. Software engineering techniques usually require training. If you compare people using a technique with people not using a technique, the effect of the technique is confounded with the effect of training. That is, any effect might be due to providing training not the specific technique. This is a particular problem if the participants are students.

Outcomes
Outcomes should relate to factors of importance to practitioners such as improved reliability, reduced production costs, and reduced time to market. All relevant

outcomes should be specified. For example, in some cases we require interventions that improve some aspect of software production without affecting another e.g. improved reliability with no increase in cost.

A particular problem for software engineering experiments is the widespread use of surrogate measures for example, defects found during system testing as a surrogate for quality, or coupling measures for design quality. Studies that use surrogate measures may be misleading and conclusions based on such studies may be less robust.

Context
For Software Engineering, this is the context in which the comparison takes place (e.g. academia or industry), the participants taking part in the study (e.g. practitioners, academics, consultants, students), and the tasks being performed (e.g. small scale, large scale). Many software experiments take place in academia using student participants and small scale tasks. Such experiments are unlikely to be representative of what might occur with practitioners working in industry. Some systematic reviews might choose to exclude such experiments although in software engineering, these may be the only type of studies available.

Experimental designs
In medical studies, researchers may be able to restrict systematic reviews to primary studies of one particular type. For example, Cochrane reviews are usually restricted to randomised controlled trials (RCTs). In other circumstances, the nature of the question and the central issue being addressed may suggest that certain study designs are more appropriate than others. However, this approach can only be taken in a discipline where the large number of research papers is a major problem. In software engineering, the paucity of primary studies is more likely to be the problem for systematic reviews and we are more likely to need protocols for aggregating information from studies of widely different types.

Examples

Kitchenham et al.[21] used the PICO criteria and defined the question elements as
Population: software or Web project.
Intervention: cross-company project effort estimation model.
Comparison: single-company project effort estimation model
Outcomes: prediction or estimate accuracy**.**

Jørgensen [17] did not use a structured version of his research questions.

5.4 Developing a Review Protocol

A review protocol specifies the methods that will be used to undertake a specific systematic review. A pre-defined protocol is necessary to reduce the possibility of researcher bias. For example, without a protocol, it is possible that the selection of individual studies or the analysis may be driven by researcher expectations. In medicine, review protocols are usually submitted to peer review.

The components of a protocol include all the elements of the review plus some additional planning information:

- Background. The rationale for the survey.
- The research questions that the review is intended to answer.
- The strategy that will be used to search for primary studies including search terms and resources to be searched. Resources include digital libraries, specific journals, and conference proceedings. An initial mapping study can help determine an appropriate strategy.
- Study selection criteria. Study selection criteria are used to determine which studies are included in, or excluded from, a systematic review. It is usually helpful to pilot the selection criteria on a subset of primary studies.
- Study selection procedures. The protocol should describe how the selection criteria will be applied e.g. how many assessors will evaluate each prospective primary study, and how disagreements among assessors will be resolved.
- Study quality assessment checklists and procedures. The researchers should develop quality checklists to assess the individual studies. The purpose of the quality assessment will guide the development of checklists.
- Data extraction strategy. This defines how the information required from each primary study will be obtained. If the data require manipulation or assumptions and inferences to be made, the protocol should specify an appropriate validation process.
- Synthesis of the extracted data. This defines the synthesis strategy. This should clarify whether or not a formal meta-analysis is intended and if so what techniques will be used.
- Dissemination strategy (if not already included in a commissioning document).
- Project timetable. This should define the review schedule.

An example of protocol for a tertiary review is given in Appendix 3. This is a simple survey, so the protocol is quite short. In our experience, protocols can be very long documents. In this case, the protocol is short because the search process is relatively limited and the data extraction and data analysis processes are relatively straightforward.

5.5 Evaluating a Review Protocol

The protocol is a critical element of any systematic review. Researchers must agree a procedure for evaluating the protocol. If appropriate funding is available, a group of independent experts should be asked to review the protocol. The same experts can later be asked to review the final report.

PhD students should present their protocol to their supervisors for review and criticism.

The basic SLR review questions discussed in Section 5.1 can be adapted to assist the evaluation of a systematic review protocol. In addition, the internal consistency of the protocol can be checked 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.

5.6 Lessons learned for protocol construction

Brereton et al. [5] identify a number of issues that researchers should anticipate during protocol construction:

- A pre-review mapping study may help in scoping research questions.
- Expect to revise questions during protocol development, as understanding of the problem increases.
- All the systematic review team members need to take an active part in developing the review protocol, so they understand how to perform the data extraction process.
- Piloting the research protocol is essential. It will find mistakes in the data collection and aggregation procedures. It may also indicate the need to change the methodology intended to address the research questions including amending the data extraction forms and synthesis methods.

Staples and Niazi [27] recommend limiting the scope of a systematic literature by choosing clear and narrow research questions.

6. Conducting the review

Once the protocol has been agreed, the review proper can start. However, as noted previously, researchers should expect to try out each of the steps described in this section when they construct their research protocol.

6.1 Identification of Research

The aim of a systematic review is to find as many primary studies relating to the research question as possible using an unbiased search strategy. The rigour of the search process is one factor that distinguishes systematic reviews from traditional reviews.

6.1.1 Generating a search strategy

It is necessary to determine and follow a search strategy. This should be developed in consultation with librarians or others with relevant experience. Search strategies are usually iterative and benefit from:

- Preliminary searches aimed at both identifying existing systematic reviews and assessing the volume of potentially relevant studies.
- Trial searches using various combinations of search terms derived from the research question.
- Checking trial research strings against lists of already known primary studies.
- Consultations with experts in the field.

A general approach is to break down the question into individual facets i.e. population, intervention, comparison, outcomes, context, study designs as discussed in Section 5.3.2. Then draw up a list of synonyms, abbreviations, and alternative spellings. Other terms can be obtained by considering subject headings used in journals and data bases. Sophisticated search strings can then be constructed using Boolean ANDs and ORs.

Initial searches for primary studies can be undertaken using digital libraries but this is not sufficient for a full systematic review. Other sources of evidence must also be searched (sometimes manually) including:

- Reference lists from relevant primary studies and review articles
- Journals (including company journals such as the IBM Journal of Research and Development), grey literature (i.e. technical reports, work in progress) and conference proceedings
- Research registers
- The Internet.

It is also important to identify specific researchers to approach directly for advice on appropriate source material.

Medical researchers have developed pre-packaged search strategies. Software engineering researchers need to develop and publish such strategies including identification of relevant digital libraries.

A problem for software engineering SLRs is that there may be relatively few studies on a particular topic. In such cases it may be a good idea to look for studies in related disciplines for example, sociology for group working practices, and psychology for notation design and/or problem solving approaches.

Example
Jørgensen [16] investigated when we can expect expert estimates to have acceptable accuracy in comparison with formal models by reviewing relevant human judgement studies (e.g. time estimation studies) and comparing their results with the results of software engineering studies.

6.1.2 Publication Bias

Publication bias refers to the problem that *positive* results are more likely to be published than *negative* results. The concept of *positive* or *negative* results sometimes depends on the viewpoint of the researcher. (For example, evidence that full mastectomies were not always required for breast cancer was actually an extremely positive result for breast cancer sufferers.)

However, publication bias remains a problem particularly for formal experiments, where failure to reject the null hypothesis is considered less interesting than an experiment that is able to reject the null hypothesis. Publication bias is even more of a problem when methods/techniques are sponsored by influential groups in the software industry. For example, the US MoD is an extremely important and influential organisation which sponsored the development of the Capability Maturity Model and used its influence to encourage industry to adopt the CMM. In such circumstances few companies would want to publish negative results and there is a strong incentive to publish papers that support the new method/technique.

Publication bias can lead to systematic bias in systematic reviews unless special efforts are made to address this problem. Many of the standard search strategies identified above are used to address this issue including:

- Scanning the grey literature
- Scanning conference proceedings

- Contacting experts and researchers working in the area and asking them if they know of any unpublished results.

In addition, statistical analysis techniques can be used to identify the potential significance of publication bias (see Section 6.5.7).

6.1.3 Bibliography Management and Document Retrieval

Bibliographic packages such as Reference Manager or Endnote may be useful for managing the large number of references that can be obtained from a thorough literature search.

Once reference lists have been finalised the full articles of potentially useful studies will need to be obtained. A logging system is needed to make sure all relevant studies are obtained.

6.1.4 Documenting the Search

The process of performing a systematic literature review must be transparent and replicable (as far as possible):

- The review must be documented in sufficient detail for readers to be able to assess the thoroughness of the search.
- The search should be documented as it occurs and changes noted and justified.
- The unfiltered search results should be saved and retained for possible reanalysis.

Procedures for documenting the search process are given in Table 2.

Table 2 Search process documentation

Data Source	Documentation
Digital Library	Name of database Search strategy for the database Date of search Years covered by search
Journal Hand Searches	Name of journal Years searched Any issues not searched
Conference proceedings	Title of proceedings Name of conference (if different) Title translation (if necessary) Journal name (if published as part of a journal)
Efforts to identify unpublished studies	Research groups and researchers contacted (Names and contact details) Research web sites searched (Date and URL)
Other sources	Date Searched/Contacted URL Any specific conditions pertaining to the search

Researchers should specify their rationale for:

- The digital libraries to be searched.
- The journal and conference proceedings to be searched.
- The use of electronic or manual searches or a combination of both. Although most text books emphasise the use of electronic search procedures, they are not usually sufficient by themselves, and some researchers strongly advocate the use of manual searches (e.g. Jørgensen, [18]).

6.1.5 Lessons learned for Search Procedures

Brereton et al. [5] identify several issues that need to be addressed when specifying electronic search procedures:

- There are alternative search strategies that enable you to achieve different sorts of search completion criteria. You must select and justify a search strategy that is appropriate for your research question. For example, knowing the publication date of the first article on a specific topic restricts the years that need to be searched. Also, if you are going to restrict your search to specific journals and conference proceedings this needs to be justified.
- We need to search many different electronic sources; no single source finds all the primary studies.
- Current software engineering search engines are not designed to support systematic literature reviews. Unlike medical researchers, software engineering researchers need to perform resource-dependent searches.

In an attempt to perform an exhaustive search Brereton et al. [5] identified seven electronic sources of relevance to Software Engineers:

- IEEExplore
- ACM Digital library:
- Google scholar (scholar.google.com)
- Citeseer library (citeseer.ist.psu.edu)
- Inspec (www.iee.org/Publish/INSPEC/)
- ScienceDirect (www.sciencedirect.com)
- EI Compendex (www.engineeringvillage2.org/Controller/Servlet/AthensService).

However, it may also be necessary to consider SpringerLink to access journals such as Empirical Software Engineering and Springer Conference Proceedings, or SCOPUS (which claims to be the largest database of abstracts and citations).

Examples

Kitchenham et al. [21] used their structured questions to construct search strings for use with electronic databases. The identified synonyms and alternative spellings for each of the question elements and linked them using the Boolean OR e.g.:

Population: software OR application OR product OR Web OR WWW OR Internet OR World-Wide Web OR project OR development

Intervention: cross company OR cross organisation OR cross organization OR multiple-organizational OR multiple-organisational model OR modeling OR modelling effort OR cost OR resource estimation OR prediction OR assessment

Contrast: within-organisation OR within-organization OR within-organizational OR within-organisational OR single company OR single organisation

Outcome: Accuracy OR Mean Magnitude Relative Error

The search strings were constructed by linking the four OR lists using the Boolean AND.

The search strings were used on 6 digital libraries:

- INSPEC
- El Compendex
- Science Direct
- Web of Science
- IEEExplore
- ACM Digital library

The search strings needed to be adapted to suit the specific requirements of the difference data bases. In addition, the researchers searched several individual journals (J) and conference proceedings (C) sources:

- Empirical Software Engineering (J)
- Information and Software Technology (J)
- Software Process Improvement and Practice (J)
- Management Science (J)
- International Software Metrics Symposium (C)
- International Conference on Software Engineering (C)
- Evaluation and Assessment in Software Engineering (manual search) (C)

These sources were chosen because they had published papers on the topic.

In addition, Kitchenham et al. checked the references of each relevant article and approached researchers who published on the topic to ask whether they had published (or were in the process of publishing) any other articles on the topic.

Jørgensen [17] used an existing database of journal papers that he had identified for another review (Jørgensen and Shepperd [15]). Jørgensen and Shepperd manually searched all volumes of over 100 journals for papers on software cost estimation. The journals were identified by reading reference lists of cost estimation papers, searching the Internet, and the researchers own experience. Individual papers were categorised and recorded in a publicly available data base (www.simula.no\BESTweb.

For conference papers, Jørgensen searched papers identified by the INSPEC database using the following search string:

‘effort estimation’ OR ‘cost estimation’) AND ‘software development’.

He also contacted authors of the relevant papers and was made aware of another relevant paper.

Kitchenham et al. used the procedure recommended by most guidelines for performing systematic review. However, it resulted in extremely long search strings that needed to be adapted to specific search engines. Jørgensen [17] used a database previously constructed for a wide survey of software cost estimation. This is an example of how valuable a mapping study can be. He also used a fairly simple search string on the INSPEC database. Kitchenham et al attempted to produce a search string that was very specific to their research question but they still found a large number of false positives. In practice, a simpler search string might have been just as effective.

It is important to note that neither study based its search process solely on searching digital libraries. Both studies had very specific research questions and the researchers were aware that the number of papers addressing the topic would be small. Thus, both studies tried hard to undertake a comprehensive search.

6.2 Study Selection

Once the potentially relevant primary studies have been obtained, they need to be assessed for their actual relevance.

6.2.1 Study selection criteria

Study selection criteria are intended to identify those primary studies that provide direct evidence about the research question. In order to reduce the likelihood of bias, selection criteria should be decided during the protocol definition, although they may be refined during the search process.

Inclusion and exclusion criteria should be based on the research question. They should be piloted to ensure that they can be reliably interpreted and that they classify studies correctly.

Examples
Kitchenham et al. used the following inclusion criteria:

- any study that compared predictions of cross-company models with within-company models based on analysis of single company project data.

They used the following exclusion criteria:

- studies where projects were only collected from a small number of different sources (e.g. 2 or 3 companies),
- studies where models derived from a within-company data set were compared with predictions from a general cost estimation model.

Jørgensen [17] included papers that compare judgment-based and model-based software development effort estimation. He also excluded one relevant paper due to "incomplete information about how the estimates were derived".

Issues:

- Medical standards make a point that it is important to avoid, as far as possible, exclusions based on the language of the primary study. This may not be so important for Software Engineering.
- It is possible that inclusion decisions could be affected by knowledge of the authors, institutions, journals or year of publication. Some medical researchers have suggested reviews should be done after such information has been removed. However, it takes time to do this and experimental evidence suggests that masking the origin of primary studies does not improve reviews [4].

6.2.2 Study selection process

Study selection is a multistage process. Initially, selection criteria should be interpreted liberally, so that unless a study identified by the electronic and hand searches can be clearly excluded based on title and abstract, a full copy should be obtained. However, Brereton et al. [5] point out that "The standard of IT and software engineering abstracts is too poor to rely on when selecting primary studies. You should also review the conclusions."

The next step is to apply inclusion/exclusion criteria based on practical issues [11] such as:

- Language
- Journal
- Authors
- Setting
- Participants or subjects
- Research Design
- Sampling method
- Date of publication.

Staples and Niazi point out that it is sometimes necessary to consider the questions that are not being addressed in order to refine your exclusion criteria [27].

Example
Staples and Niazi's research question was

- Why do organizations embark on CMM-based SPI initiatives?

They also defined complementary research questions that were not being investigated:

- What motivates individuals to support the adoption of CMM-based SPI in an organization?
- Why should organizations embark on CMM-based SPI initiatives?
- What reasons for embarking on CMM-based SPI are the most important to organizations?
- What benefits have organizations received from CMM-based SPI initiatives?
- How do organizations decide to embark on CMM-based SPI initiatives?
- What problems do organizations have at the time that they decide to adopt CMM-based SPI?

This clarified the boundaries of their research question of interest for example they were concerned with the motivations of organisations not the motivations of individuals and they were concerned with why organisations rejected CMM not why the adopted it. They found that this process directly improved and clarified their primary study selection and data extraction process.

Sometimes, researchers undertake a third stage in the selection process based on detailed quality criteria.

Most general SLR text books recommend maintaining a list of excluded studies identifying the reason for exclusion. However, in our experience, initial electronic searches results in large numbers of totally irrelevant papers, i.e. papers that not only do not address any aspect of the research questions but do not even have anything do with software engineering. We, therefore, recommend maintaining a list of excluded papers, only after the totally irrelevant papers have been excluded, in particular, maintaining a record of those candidate primary studies that are excluded as a result of the more detailed inclusion/exclusion criteria.

6.2.3 Reliability of inclusion decisions

When two or more researchers assess each paper, agreement between researchers can be measured using the Cohen Kappa statistic [9]. The initial value of the Kappa statistics should be documented in the final report. Each disagreement must be discussed and resolved. This may be a matter of referring back to the protocol or may involve writing to the authors for additional information. Uncertainty about the inclusion/exclusion of some studies should be investigated by sensitivity analysis.

A single researcher (such as a PhD student) should consider discussing included and excluded papers with their advisor, an expert panel or other researchers. Alternatively, individual researchers can apply a test-retest approach, and re-evaluate a random sample of the primary studies found after initial screening to check the consistency of their inclusion/exclusion decisions.

6.3 Study Quality Assessment

In addition to general inclusion/exclusion criteria, it is considered critical to assess the "quality" of primary studies:

- To provide still more detailed inclusion/exclusion criteria.
- To investigate whether quality differences provide an explanation for differences in study results.

- As a means of weighting the importance of individual studies when results are being synthesised.
- To guide the interpretation of findings and determine the strength of inferences.
- To guide recommendations for further research.

An initial difficulty is that there is no agreed definition of study "quality". However, the CRD Guidelines [19] and the Cochrane Reviewers' Handbook [7] both suggest that quality relates to the extent to which the study minimises bias and maximises internal and external validity (see Table 3).

Table 3 Quality concept definitions

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

Most quality checklists (see Section 6.3.2) include questions aimed at assessing the extent to which articles have addressed bias and validity.

6.3.1 The Hierarchy of Evidence

Medical guidelines suggest that an initial quality evaluation can be based on the type of experiment design being used. Thus, we might rate a randomised controlled trial as more trustworthy than an observational study. This has led to the concept of a hierarchy of evidence with evidence from systematic reviews and randomised controlled experiments at the top of the hierarchy and evidence from quasi-experiments and expert opinion at the bottom of the hierarchy (see [19] and [2]). Researchers can then use these hierarchies to restrict the type of studies they include in their systematic literature review.

Recently, Petticrew and Roberts [25] have suggested that this idea is too simplistic. They point out that some types of design are better than others at addressing different types of question. For example, qualitative studies are more appropriate than randomised experiments for assessing whether practitioners find a new technology appropriate for the type of applications they have to build. Thus, if we want to restrict ourselves to studies of a specific type we should restrict ourselves to studies that are best suited to addressing our specific research questions.

However, there is evidence that observational (e.g. correlation) studies can be unreliable. Medical researchers have often discovered that the results of extremely large scale observational studies have been overturned by the results of randomised controlled trials. A recent example is that of the supposed benefits of vitamin C [22]. Two large scale observational studies had previously suggested that taking vitamin C protected against heart disease. Lawlor et al. [22] suggest that the reason observational studies found a result that could not be observed in randomised trials was that use of vitamin C was a surrogate for other life-style characteristics that protect against heart disease such as exercising and keeping to a healthy diet. This is

an issue that needs to be taken seriously in software engineering where much of our research on topics such as software cost estimation and project success factors are correlation studies. Good observational studies need to consider possible confounding effects, put in place methods to measure them and adjust any analyses to allow for their effect. In particular, they need to include sensitivity analysis to investigate the impact of measured and unmeasured confounders.

6.3.2 Development of Quality Instruments

Detailed quality assessments are usually based on "quality instruments" which are checklists of factors that need to be evaluated for each study. If quality items within a checklist are assigned numerical scales, numerical assessments of quality can be obtained.

Checklists are usually derived from a consideration of factors that could bias study results. The CRD Guidelines [19], the Australian National Health and Medical Research Council Guidelines [1], and the Cochrane Reviewers' Handbook [7] all refer to four types of bias shown in Table 4. (We have amended the definitions (slightly) and protection mechanisms (considerably) to address software engineering rather than medicine.) In particular, medical researchers rely on "blinding" subjects and experimenters (i.e. making sure that neither the subject nor the researcher knows which treatment a subject is assigned to) to address performance and measurement bias. However, that protocol is usually impossible for software engineering experiments.

Table 4 Types of Bias

Type	Synonyms	Definition	Protection mechanism
Selection bias	Allocation bias	Systematic difference between comparison groups with respect to treatment	Randomisation of a large number of subjects with concealment of the allocation method (e.g. allocation by computer program not experimenter choice).
Performance bias		Systematic difference is the conduct of comparison groups apart from the treatment being evaluated.	Replication of the studies using different experimenters. Use of experimenters with no personal interest in either treatment.
Measurement bias	Detection Bias	Systematic difference between the groups in how outcomes are ascertained.	Blinding outcome assessors to the treatments is sometimes possible.
Attrition bias	Exclusion bias	Systematic differences between comparison groups in terms of withdrawals or exclusions of participants from the study sample.	Reporting of the reasons for all withdrawals. Sensitivity analysis including all excluded participants.

The factors identified in Table 4 can be refined into a quality instrument by considering:

- Generic items that relate to features of particular study designs, such as survey designs, experimental designs, and qualitative study designs.
- Specific items that relate to the review's subject area such as the particular method of cross-validation used in a study of cost estimation prediction accuracy.

Checklists are also developed by considering bias and validity problems that can occur at the different stages in an empirical study:

- Design
- Conduct
- Analysis
- Conclusions.

There are many published quality checklists for different types of empirical study. The medical guidelines all provide checklists aimed at assisting the quality assessment undertaken during a systematic literature review as do Fink [11] and Petticrew and Roberts [25]. In addition, Crombie [10] and Greenhalgh [12] also provide checklists aimed at assisting a reader to evaluate a specific article. Shaddish et al. [25] discuss quasi-experimental designs and provide an extensive summary of validity issues affecting them. However, each source identifies a slightly different set of questions and there is no standard agreed set of questions.

For quantitative studies we have accumulated a list of questions from [10], [11], [12], [19] and [25] and organised them with respect to study stage and study type (see Table 5). We do not suggest that anyone uses all the questions. Researchers should adopt Fink's suggestion [11] which is to review the list of questions in the context of their own study and select those quality evaluation questions that are most appropriate for their specific research questions. They may need to construct a measurement scale for each item since sometimes a simple Yes/No answer may be misleading. Whatever form the quality instrument takes, it should be assessed for reliability and usability during the trials of the study protocol before being applied to all the selected studies.

Examples

Kitchenham et al. [21] constructed a quality questionnaire based on 5 issues affecting the quality of the study which were scored to provide an overall measure of *study* quality:

1. Is the data analysis process appropriate?
   1.1 Was the data investigated to identify outliers and to assess distributional properties before analysis?
   1.2 Was the result of the investigation used appropriately to transform the data and select appropriate data points?
2. Did studies carry out a sensitivity or residual analysis?
   2.1 Were the resulting estimation models subject to sensitivity or residual analysis?
   2.2 Was the result of the sensitivity or residual analysis used to remove abnormal data points if necessary?
3. Were accuracy statistics based on the raw data scale?
4. How good was the study comparison method?
   4.1 Was the single company selected at random (not selected for convenience) from several different companies?
   4.2 Was the comparison based on an independent hold out sample (0.5) or random subsets (0.33), leave-one-out (0.17), no hold out (0)? The scores used for this item reflect the researchers opinion regarding the stringency of each criterion.
5. The size of the within-company data set, measured according to the criteria presented below. Whenever a study used more than one within-company data set, the average score was used:

- Less than 10 projects: Poor quality (score = 0)
- Between 10 and 20 projects: Fair quality (score = 0.33)
- Between 21 and 40 projects: Good quality (score = 0.67)
- More than 40 projects: Excellent quality (score = 1)

They also considered the *reporting* quality based on 4 questions:

1. Is it clear what projects were used to construct each model?
2. Is it clear how accuracy was measured?
3. Is it clear what cross-validation method was used?
4. Were all model construction methods fully defined (tools and methods used)?

It is good practice not to include quality of study and quality of reporting scores in a single metric but Kitchenham et al. proposed using a weighted measure giving less weight to the reporting quality score.

Kitchenham et al.'s quality questionnaire was based on the specific nature of the primary studies (such as the method of cross-validation used) as well as more general quality issues (such as sample size, and sensitivity analysis).

Jørgensen [17] did not undertake a specific quality assessment of the primary studies.

Table 5 Summary Quality Checklist for Quantitative Studies

Question	Quantitative Empirical Studies (no specific type)	Correlation (observational studies)	Surveys	Experiments	Source
Design					
Are the aims clearly stated?	X	X	X	X	[11], [10]
Was the study designed with these questions in mind?			X		[25]
Do the study measures allow the questions to be answered?			X	X	[10], [25]
What population was being studied?			X		[25]
Who was included?			X		[12]
Who was excluded?			X		[12]
How was the sample obtained (e.g. postal, interview, web-based)?			X		[10], [12], [25]
Is the survey method likely to have introduced significant bias?			X		[25]
Is the sample representative of the population to which the results will generalise?			X	X	[10], [25]
Were treatments randomly allocated?				X	[10]
Is there a comparison or control group?	X		X	X	[12]
If there is a control group, are participants similar to the treatment group participants in terms of variables that may affect study outcomes?	X		X	X	[10], [12]
Was the sample size justified	X		X	X	[10], [12]
If the study involves assessment of a technology, is the technology clearly defined?	X	X	X	X	[11]
Could the choice of subjects influence the size of the treatment effect?				X	[10], [11], [19],[25]
Could lack of blinding introduce bias?				X	[10]
Are the variables used in the study adequately measured (i.e. are the variables likely to be valid and reliable)?	X	X	X	X	[10], [11], [19],[25]
Are the measures used in the study fully defined?	X	X	X	X	[11]

Are the measures used in the study the most relevant ones for answering the research questions?	X	X	X	X	[11], [19],[25]
Is the scope (size and length) of the study sufficient to allow for changes in the outcomes of interest to be identified?	X		X	X	[19], [12], [25]
Conduct					
Did untoward events occur during the study?	X	X	X	X	[10]
Was outcome assessment blind to treatment group?	X			X	[19], [12], [25]
Are the data collection methods adequately described?	X	X	X	X	[11]
If two groups are being compared, were they treated similarly within the study?				X	[12], [25]
If the study involves participants over time, what proportion of people who enrolled at the beginning dropped out?	X		X	X	[10], [11]
How was the randomisation carried out?				X	[10]
Analysis					
What was the response rate?			X		[10], [25]
Was the denominator (i.e. the population size) reported?			X		[25]
Do the researchers explain the data types (continuous, ordinal, categorical)?	X	X	X	X	[11]
Are the study participants or observational units adequately described? For example, SE experience, type (student, practitioner, consultant), nationality, task experience and other relevant variables.	X	X	X	X	[12], [25]
Were the basic data adequately described?	X	X	X	X	[10]
Have “drop outs” introduced bias?	X		X	X	[11], [12], [25]
Are reasons given for refusal to participate?	X		X	X	[11]
Are the statistical methods described?	X	X	X	X	[10], [11], [19]
Is the statistical program used to analyse the data referenced?	X	X	X	X	[11]
Are the statistical methods justified?	X	X	X	X	[11]
Is the purpose of the analysis clear?	X	X	X	X	[11]
Are scoring systems described?	X			X	[11]
Are potential confounders adequately controlled for in the analysis?	X	X	X	X	[11]
Do the numbers add up across different tables and	X	X	X	X	[10], [11]

subgroups?					
If different groups were different at the start of the study or treated differently during the study, was any attempt made to control for these differences, either statistically or by matching?	X		X	X	[12], [25]
If yes, was it successful?	X		X	X	[25]
Was statistical significance assessed?	X	X	X	X	[10]
If statistical tests are used to determine differences, is the actual p value given?	X	X	X	X	[11]
If the study is concerned with differences among groups, are confidence limits given describing the magnitude of any observed differences?	X		X	X	[11]
Is there evidence of multiple statistical testing or large numbers of post hoc analysis?	X	X	X	X	[10], [25]
How could selection bias arise?	X		X	X	[10], [25]
Were side-effects reported?					[10]
Conclusions					
Are all study questions answered?	X	X	X	X	[11]
What do the main findings mean?	X	X	X	X	[10]
Are negative findings presented?	X	X	X	X	[11]
If statistical tests are used to determine differences, is practical significance discussed?	X	X	X	X	[11]
If drop outs differ from participants, are limitations to the results discussed?	X		X	X	[11]
How are null findings interpreted? (I.e. has the possibility that the sample size is too small been considered?)	X	X	X	X	[10], [12]
Are important effects overlooked?	X	X	X	X	[10]
How do results compare with previous reports?	X	X	X	X	[10]
How do the results add to the literature?	X	X	X	X	[12]
What implications does the report have for practice?	X	X	X	X	[10]
Do the researchers explain the consequences of any problems with the validity/reliability of their measures?	X	X	X	X	[11]

If a review includes qualitative studies, it will be necessary to assess their quality. Table 6 provides a checklist of assessing the quality of qualitative studies.

Table 6 Checklist for qualitative studies

Number	Question	Source
1	How credible are the findings?	[12], [25]
1.1	If credible, are they important?	[12]
2	How has knowledge or understanding been extended by the research?	[12], [25]
3	How well does the evaluation address its original aims and purpose?	[25]
4	How well is the scope for drawing wider inference explained?	[25]
5	How clear is the basis of evaluative appraisal?	[25]
6	How defensible is the research design?	[12], [25], [11]
7	How well defined are the sample design/target selection of cases/documents?	[12], [25], [11]
8	How well is the eventual sample composition and coverage described?	[25]
9	How well was data collection carried out?	[12], [25], [11]
10	How well has the approach to, and formulation of, analysis been conveyed?	[12], [25], [11]
11	How well are the contexts and data sources retained and portrayed?	[25]
12	How well has diversity of perspective and context been explored?	[25]
13	How well have detail, depth, and complexity (i.e. richness) of the data been conveyed?	[25]
14	How clear are the links between data, interpretation and conclusions – i.e. how well can the route to any conclusions be seen?	[25]
15	How clear and coherent is the reporting?	[25]
16	How clear are the assumptions/theoretical perspectives/values that have shaped the form and output of the evaluation?	[12], [25], [11]
17	What evidence is there of attention to ethical issues?	[25]
18	How adequately has the research process been documented?	[25]

6.3.3 Using the Quality Instrument

It is important that researchers not only define the quality instrument in the study protocol but also specify how the quality data are to be used. Quality data can be used in two rather different ways:

1. To assist primary study selection. In this case, the quality data are used to construct detailed inclusion/exclusion criteria. The quality data must be collected prior to the main data collection activity using separate data collection forms.
2. To assist data analysis and synthesis. In this case the quality data are used to identify subsets of the primary study to investigate whether quality differences are associated with different primary study outcomes. The quality data can be collected at the same time as the main data extraction activity using a joint form.

It is of course possible to have both types of quality data in the same systematic review.

Example

Kitchenham et al. [21] used the quality score to investigate whether the results of the primary study were associated with study quality. They also investigated whether some of the individual quality factors (i.e. sample size, validation method) were associated with primary study outcome.

Some researchers have suggested weighting meta-analysis results using quality scores. This idea is **not** recommended by any of the medical guidelines.

If a systematic review includes studies of different types, it is necessary to use an appropriate quality instrument for each study type. In some cases a common set of quality evaluation questions may be suitable for all the quantitative studies included in a systematic review, but if a review includes qualitative and quantitative studies different checklists will be essential.

6.3.4 Limitations of Quality Assessment

Primary studies are often poorly reported, so it may not be possible to determine how to assess a quality criterion. It is tempting to assume that because something wasn't reported, it wasn't done. This assumption may be incorrect. Researchers should attempt to obtain more information from the authors of the study. Petticrew and Roberts [25] explicitly point out that quality checklists need to address methodological quality *not* reporting quality.

There is limited evidence of relationships between factors that are thought to affect validity and actual study outcomes. Evidence suggests that inadequate concealment of allocation and lack of double-blinding result in over-estimates of treatment effects, but the impact of other quality factors is not supported by empirical evidence.

It is possible to identify inadequate or inappropriate statistical analysis, but without access to the original data it is not possible to correct the analysis. Very often software data is confidential and cannot therefore be made generally available to researchers. In some cases, software engineers may refuse to make their data available to other researchers because they want to continue publishing analyses of the data.

6.4 Data Extraction

The objective of this stage is to design data extraction forms to accurately record the information researchers obtain from the primary studies. To reduce the opportunity for bias, data extraction forms should be defined and piloted when the study protocol is defined.

6.4.1 Design of Data Extraction Forms

The data extraction forms must be designed to collect all the information needed to address the review questions and the study quality criteria. If the quality criteria are to be used to identify inclusion/exclusion criteria, they require separate forms (since the information must be collected prior to the main data extraction exercise). If the quality criteria are to be used as part of the data analysis, the quality criteria and the review data can be included in the same form.

In most cases, data extraction will define a set of numerical values that should be extracted for each study (e.g. number of subjects, treatment effect, confidence intervals, etc.). Numerical data are important for any attempt to summarise the results

of a set of primary studies and are a prerequisite for meta-analysis (i.e. statistical techniques aimed at integrating the results of the primary studies).

Data extraction forms need to be piloted on a sample of primary studies. If several researchers will use the forms, they should all take part in the pilot. The pilot studies are intended to assess both technical issues such as the completeness of the forms and usability issues such as the clarity of user instructions and the ordering of questions.

Electronic forms are useful and can facilitate subsequent analysis.

6.4.2 Contents of Data Collection Forms

In addition to including all the questions needed to answer the review question and quality evaluation criteria, data collection forms should provide standard information including:

- Name of Reviewer
- Date of Data extraction
- Title, authors, journal, publication details
- Space for additional notes

Examples

Kitchenham et al. [21] used the extraction form shown in Table 7 (note the actual form also included the quality questions).

Table 7 Data Collection form completed for Maxwell et al., 1998

Data item	Value	Additional notes
Data Extractor		
Data Checker		
Study Identifier	S1	
Application domain	Space, military and industrial	
Name of database	European Space Agency (ESA)	
Number of projects in database (including within-company projects)	108	
Number of cross-company projects	60	
Number of projects in within-company data set	29	
Size metric(s): FP (Yes/No) Version used: LOC (Yes/No) Version used: Others (Yes/No) Number:	FP: No LOC: Yes (KLOC) Others: No	
Number of companies	37	
Number of countries represented	8	European only
Were quality controls applied to data collection?	No	
If quality control, please describe		
How was accuracy measured?	Measures: R^2 (for model construction only) MMRE Pred(25) r (Correlation between estimate and actual)	

Cross-company model		
What technique(s) was used to construct the cross-company model?	A preliminary productivity analysis was used to identify factors for inclusion in the effort estimation model. Generalised linear models (using SAS). Multiplicative and Additive models were investigated. The multiplicative model is a logarithmic model.	
If several techniques were used which was most accurate?	In all cases, accuracy assessment was based on the logarithmic models not the additive models.	It can be assumed that linear models did not work well.
What transformations if any were used?	Not clear whether the variables were transformed or the GLM was used to construct a log-linear model	Not important: the log models were used and they were presented in the raw data form – thus any accuracy metrics were based on raw data predictions.
What variables were included in the cross-company model?	KLOC, Language subset, Category subset, RELY	Category is the type of application. RELY is reliability as defined by Boehm (1981)
What cross-validation method was used?	A hold-out sample of 9 projects from the single company was used to assess estimate accuracy	
Was the cross-company model compared to a baseline to check if it was better than chance?	Yes	The baseline was the correlation between the estimates and the actuals for the hold-out.
What was/were the measure(s) used as benchmark?	The correlation between the prediction and the actual for the single company was tested for statistical significance. (Note it was significantly different from zero for the 20 project data set, but not the 9 project hold-out data set.)	
Within-company model		
What technique(s) was used to construct the within-company model?	A preliminary productivity analysis was used to identify factors for inclusion in the effort estimation model. Generalised linear models (using SAS). Multiplicative and Additive models were investigated. The multiplicative model is a logarithmic model.	
If several techniques were used which was most accurate?	In all cases, accuracy assessment was based on the logarithmic models not the additive models.	It can be assumed that linear models did not work well.
What transformations if any were used?	Not clear whether the variables were transformed or the GLM was used to construct a log-linear model	Not important: the log models were used and they were presented in the raw data form – thus any accuracy metrics were based on raw data predictions.
What variables were included in the within-company model?	KLOC, Language subset, Year	
What cross-validation	A hold-out sample of 9 projects	

method was used	from the single company was used to assess estimate accuracy	
Comparison		
What was the accuracy obtained using the cross-company model?	Accuracy on main single company data set (log model): n=11 (9 projects omitted) MMRE=50% Pred(25)=27% r=0.83 Accuracy on single company hold out data set n=4 (5 projects omitted) MMRE=36% Pred(25)=25% R=0.16 (n.s)	Using the 79 cross-company projects, Maxwell et al. identified the best model for that dataset and the best model for the single company data. The two models were identical. This data indicates that for all the single company projects: n=15 Pred(25)=26.7% (4 of 15) MMRE=46.3%
What was the accuracy obtained using the within-company model?	Accuracy on main single company data set (log model): n=14 (6 projects omitted) R^2=0.92 MMRE=41% Pred(25)=36% r=0.99 Accuracy on single company hold out data set n=6 (3 projects omitted) MMRE=65% Pred(25)=50% (3 of 6) r=0.96	
What measure was used to check the statistical significance of prediction accuracy (e.g. absolute residuals, MREs)?	Estimated and actual effort	
What statistical tests were used to compare the results?	r, correlation between the prediction and the actual	
What were the results of the tests?		
Data Summary		
Data base summary (all projects) for size and effort metrics.	Effort min: 7.8 MM Effort max: 4361 MM Effort mean: 284 MM Effort median: 93 MM Size min: 2000 KLOC Size max: 413000 KLOC Size mean: 51010 KLOC Size median: 22300 KLOC	KLOC: non-blank, non-comment delivered 1000 lines. For reused code Boehm's adjustment were made (Boehm, 1981). Effort was measured in man months, with 144 man hours per man month
With-company data summary for size and effort metrics.	Effort min: Effort max: Effort mean: Effort median: Size min: Size max: Size mean: Size median:	Not specified

Jørgensen [17] extracted design factors and primary study results. Design factors included:

- Study design
- Estimation method selection process
- Estimation models

- Calibration level
- Model use expertise and degree of mechanical use of model
- Expert judgment process
- Expert judgement estimation expertise
- Possible motivational biases in estimation situation
- Estimation input
- Contextual information
- Estimation complexity
- Fairness limitations
- Other design issues

Study results included:

- Accuracy
- Variance
- Other results

Jørgensen's article includes the completed extraction form for each primary study.

6.4.3 Data extraction procedures

Whenever feasible, data extraction should be performed independently by two or more researchers. Data from the researchers must be compared and disagreements resolved either by consensus among researchers or arbitration by an additional independent researcher. Uncertainties about any primary sources for which agreement cannot be reached should be investigated as part of any sensitivity analyses. A separate form must be used to mark and correct errors or disagreements.

If several researchers each review different primary studies because time or resource constraints prevent all primary papers being assessed by at least two researchers, it is important to employ some method of checking that researchers extract data in a consistent manner. For example, some papers should be reviewed by all researchers (e.g. a random sample of primary studies), so that inter-researcher consistency can be assessed.

For single researchers such as PhD students, other checking techniques must be used. For example supervisors could perform data extraction on a random sample of the primary studies and their results cross-checked with those of the student. Alternatively, a test-retest process can be used where the researcher performs a second extraction from a random selection of primary studies to check data extraction consistency.

Examples

Kitchenham et al. [21] assigned one person to be the data extractor who completed the data extraction form and another person to be the data checker who confirmed that the data on extraction form were correct. Because Kitchenham and Mendes co-authored some of the primary studies, they also ensured that the data extractor was never a co-author of the primary study. Any disagreements were examined and an agreed final data value recorded.

As a single researcher, Jørgensen [17] extracted all the data himself. However, he sent the data from each primary study to an author of the study and requested that they inform him if any of the extracted data was incorrect.

6.4.4 Multiple publications of the same data

It is important not to include multiple publications of the same data in a systematic review synthesis because duplicate reports would seriously bias any results. It may be necessary to contact the authors to confirm whether or not reports refer to the same

study. When there are duplicate publications, the most complete should be used. It may even be necessary to consult all versions of the report to obtain all the necessary data.

6.4.5 Unpublished data, missing data and data requiring manipulation

If information is available from studies in progress, it should be included providing appropriate quality information about the study can be obtained and written permission is available from the researchers.

Reports do not always include all relevant data. They may also be poorly written and ambiguous. Again the authors should be contacted to obtain the required information.

Sometimes primary studies do not provide all the data but it is possible to recreate the required data by manipulating the published data. If any such manipulations are required, data should first be reported in the way they were published. Data obtained by manipulation should be subject to sensitivity analysis.

6.4.6 Lessons learned about Data Extraction

Brereton et al. [5] identified two issues of importance during data extraction:

- Having one reader act as data extractor and one act as data checker may be helpful when there are a large number of papers to review.
- Review team members must make sure they understand the protocol and the data extraction process.

6.5 Data Synthesis

Data synthesis involves collating and summarising the results of the included primary studies. Synthesis can be descriptive (non-quantitative). However, it is sometimes possible to complement a descriptive synthesis with a quantitative summary. Using statistical techniques to obtain a quantitative synthesis is referred to as *meta-analysis*. Description of meta-analysis methods is beyond the scope of this document, although techniques for displaying quantitative results will be described. (To learn more about meta-analysis see [7].)

The data synthesis activities should be specified in the review protocol. However, some issues cannot be resolved until the data is actually analysed, for example, subset analysis to investigate heterogeneity is not required if the results show no evidence of heterogeneity.

6.5.1 Descriptive (Narrative) synthesis

Extracted information about the studies (i.e. intervention, population, context, sample sizes, outcomes, study quality) should be tabulated in a manner consistent with the review question. Tables should be structured to highlight similarities and differences between study outcomes.

It is important to identify whether results from studies are consistent with one another (i.e. homogeneous) or inconsistent (e.g. heterogeneous). Results may be tabulated to display the impact of potential sources of heterogeneity, e.g. study type, study quality, and sample size.

Examples
Kitchenham et al. [21] tabulated the data from the primary studies in three separate tables based on the outcome of the primary study: no significant difference between the cross-company model and the within company model, within-company model significantly better than the cross-company model and no statistical tests performed. They also highlighted studies that they believed should be excluded from the synthesis because they were complete replications in terms of the cross-company database and the within company database because they did not offer additional independent evidence.

They concluded that small companies producing specialised (niche) software would not benefit from using a cross-company estimation model. Large companies producing applications of similar size range to the cross-company projects might find cross-company models helpful.

Jørgensen [17] tabulated the studies according to the relative accuracy of the model and the experts. Thus he considered the accuracy of the most accurate expert and least accurate expert compared with the most accurate and least accurate models. He also considered the average accuracy of the models and the experts. He coded the studies chronologically (as did Kitchenham et al.), so it was possible to look for possible associations with study age and outcome.

He concluded that models are not systematically better than experts for software cost estimation, possibly because experts possess more information than models or it may be difficult to build accurate software development estimation models. Expert opinion is likely to be useful if models are not calibrated to the company using them and/or experts have access to important contextual information that they are able to exploit. Models (or a combination of models and experts) may be useful when there are situational biases towards overoptimism, experts do not have access to large amounts of contextual information, and/or models are calibrated to the environment.

6.5.2 Quantitative Synthesis

Quantitative data should also be presented in tabular form including:

- Sample size for each intervention.
- Estimates effect size for each intervention with standard errors for each effect.
- Difference between the mean values for each intervention, and the confidence interval for the difference.
- Units used for measuring the effect.

However, to synthesise quantitative results from different studies, study outcomes must be presented in a comparable way. Medical guidelines suggest different effect measures for different types of outcome.

Binary outcomes (Yes/No, Success/Failure) can be measured in several different ways:

- Odds. The ratio of the number of subjects in a group with an event to the number without an event. Thus if 20 projects in a group of 100 project failed to achieve budgetary targets, the odds would be 20/80 or 0.25.
- Risk (proportion, probability, rate) The proportion of subjects in a group observed to have an event. Thus, if 20 out of 100 projects failed to achieve budgetary targets, the risk would be 20/100 or 0.20.
- Odds ratio (OR). The ratio of the odds of an event in the experimental (or intervention) group to the odds of an event on the control group. An OR equal to one indicates no difference between the control and the intervention group. For undesirable outcomes a value less than one indicates that the intervention was

successful in reducing risk, for a desirable outcome a value greater than one indicates that the intervention was successful in reducing risk.

- Relative risk (RR) (risk ratio, rate ratio). The ratio of risk in the intervention group to the risk in the control group. An RR of one indicates no difference between comparison groups. For undesirable events an RR less than one indicates the intervention was successful, for desirable events an RR greater than one indicates the intervention was successful.
- Absolute risk reduction (ARR) (risk difference, rate difference). The absolute difference in the event rate between the comparison groups. A difference of zero indicates no difference between the groups. For an undesirable outcome an ARR less than zero indicates a successful intervention, for a desirable outcome an ARR greater than zero indicates a successful intervention.

Each of these measures has advantages and disadvantages. For example, odds and odds ratios are criticised for not being well-understood by non-statisticians (other than gamblers), whereas risk measures are generally easier to understand. Alternatively, statisticians prefer odds ratios because they have some mathematically desirable properties. Another issue is that relative measures are generally more consistent than absolute measures for statistical analysis, but decision makers need absolute values in order to assess the real benefit of an intervention.

Effect measures for continuous data include:

- Mean difference. The difference between the means of each group (control and intervention group).
- Weighted mean difference (WMD). When studies have measured the difference on the same scale, the weight given to each study is usually the inverse of the study variance
- Standardised mean difference (SMD). A common problem when summarising outcomes is that outcomes are often measured in different ways, for example, productivity might be measured in function points per hour, or lines of code per day. Quality might be measured as the probability of exhibiting one or more faults or the number of faults observed. When studies use different scales, the mean difference may be divided by an estimate of the within-groups standard deviation to produce a standardised value without any units. However, SMDs are only valid if the difference in the standard deviations reflect differences in the measurement scale, not real differences among trial populations.

6.5.3 Presentation of Quantitative Results

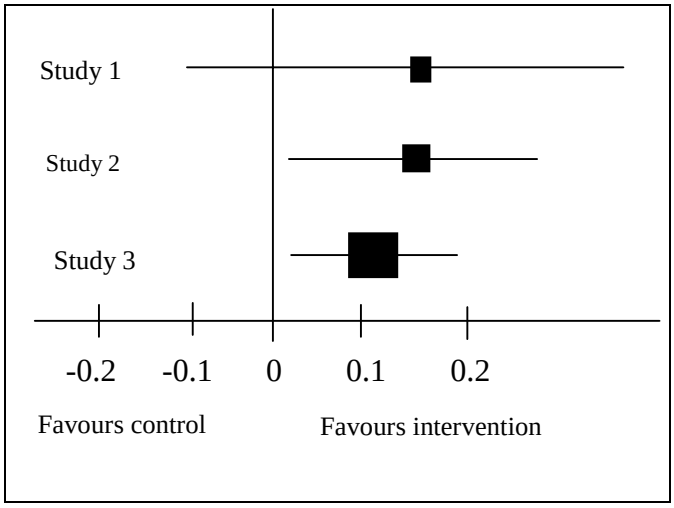
The most common mechanism for presenting quantitative results is a forest plot, as shown in Figure 1. A forest plot presents the means and variance of the difference for each study. The line represents the standard error of the difference, the box represents the mean difference and its size is proportional to the number of subjects in the study. A forest plot may also be annotated with the numerical information indicating the number of subjects in each group, the mean difference and the confidence interval on the mean. If a formal meta-analysis is undertaken, the bottom entry in a forest plot will be the summary estimate of the treatment difference and confidence interval for the summary difference.

Figure 1 represents the ideal result of a quantitative summary, as the results of the studies basically agree. There is clearly a genuine treatment effect and a single overall

summary statistic would be a good estimate of that effect. If effects were very different from study to study, our results would suggest heterogeneity. A single overall summary statistics would probably be of little value. The systematic review should continue with an investigation of the reasons for heterogeneity.

To avoid the problems of post-hoc analysis (i.e. "fishing" for results), researchers should identify possible sources of heterogeneity when they construct the review protocol. For example, studies of different types may have different results, so it is often useful to synthesise the results of different study types separately and assess whether the results are consistent across the different study types.

Figure 1 Example of a forest plot



6.5.4 Qualitative Synthesis

Synthesizing qualitative studies involves trying to integrate studies comprising natural language results and conclusions, where different researchers may have used terms and concepts with subtly (or grossly) different meanings. Noblit and Hare [23] propose three approaches to qualitative synthesis:

- Reciprocal translation. When studies are about similar things and researchers are attempting to provide an additive summary, synthesis can be achieved by "translating" each case into each of the other cases.
- Refutational Synthesis. When studies are implicitly or explicitly refutations of each other, it is necessary to translate both the individual studies and the refutations allowing the refutations to be analysed in detail.
- Line of argument synthesis. This approach is used when researchers are concerned about what they can infer about a topic as a whole from a set of selective studies that look at a part of the issue. This analysis is a two part one. First the individual studies are analysed, then an attempt is made to analyse the set of studies as a whole. This is rather similar to a descriptive synthesis. Issues of importance are identified and the approach to each issue taken by each study is documented and tabulated.

6.5.5 Synthesis of qualitative and quantitative studies

When researchers have a systematic literature review that includes quantitative and qualitative studies, they should:

- Synthesise the quantitative and qualitative studies separately.
- Then attempt to integrate the qualitative and quantitative results by investigating whether the qualitative results can help explain the quantitative results. For example qualitative studies can suggest reasons why a treatment does or does not work in specific circumstances.

As yet we have no published software engineering SLRs that have combined a qualitative survey and a quantitative survey. However, Sutcliffe et al. [28] provide an example of such a study in their survey of children and healthy eating. They performed three syntheses:

1. A statistical meta-analysis of studies which attempted to increase children's consumption of fruit and vegetables.
2. A thematic qualitative synthesis of studies focused on children's views of healthy eating.
3. A "cross-study synthesis" that used the results of the qualitative synthesis to interpret the findings of the meta-analysis.

6.5.6 Sensitivity analysis

Sensitivity analysis is important whether you have undertaken a descriptive or quantitative synthesis. However, it is usually easier to perform as part of a meta-analysis (since quantitative sensitivity analysis techniques are well understood). In such cases, the results of the analysis should be repeated on various subsets of primary studies to determine whether the results are robust. The types of subsets selected would be:

- High quality primary studies only.
- Primary studies of particular types.
- Primary studies for which data extraction presented no difficulties (i.e. excluding any studies where there was some residual disagreement about the data extracted).
- The experimental method used by the primary studies.

When a formal meta-analysis is not undertaken but quantitative results have been tabulated, forest plots can be annotated to identify high quality primary studies, the studies can be presented in decreasing order of quality or in decreasing study type hierarchy order. Primary studies where there are queries about the data extracted can also be explicitly identified on the forest plot, by for example, using grey colouring for less reliable studies and black colouring for reliable studies.

When you have undertaken a descriptive synthesis, sensitivity analysis is more subjective, but you should consider what impact excluding poor quality studies or studies of a particular type would have on your conclusions.

Examples

Jørgensen [17] reported the results of field studies as well as the results of all studies based on the argument that field studies would have more external validity.

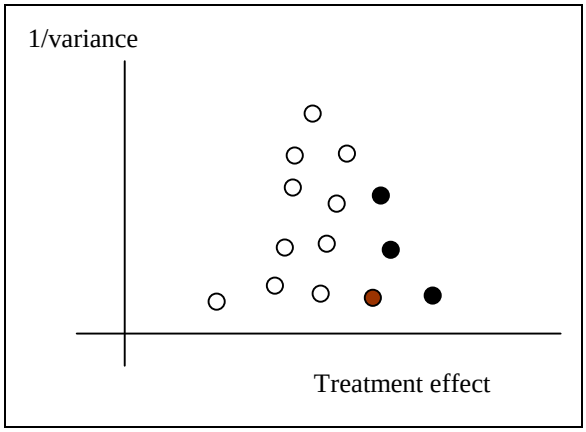
In a study of the Technology Acceptance Model (TAM), Turner et al. [29] investigated the relationship between the TAM variables Perceived Ease of Use (PEU) and PU (Perceived Usefulness) and Actual Use measured subjectively and objectively. As part of their sensitivity

analysis they investigated the impact on their results of removing primary studies authored by the researcher who developed the TAM.

6.5.7 Publication bias

Funnel plots are used to assess whether or not a systematic review is likely to be vulnerable to publication bias. Funnel plots plot the treatment effect (i.e. mean difference between intervention group and control) against the inverse of the variance or the sample size. A systematic review that exhibited the funnel shape shown in Figure 2 would be assumed **not** to be exhibiting evidence of publication bias. It would be consistent with studies based on small samples showing more variability in outcome than studies based on large samples. If, however, the points shown as filled-in black dots were not present, the plot would be asymmetric and it would suggest the presence of publication bias. This would suggest the results of the systematic review must be treated with caution.

Figure 2 An example of a funnel plot



6.5.8 Lessons Learned about Data Synthesis

Brereton et al. [5] identified three issues of importance during data extraction:

- IT and software engineering systematic reviews are likely to be qualitative (i.e. descriptive) in nature.
- Even when collecting quantitative information it may not be possible to perform meta-analysis of IT and software engineering studies because the reporting protocols vary so much between studies.
- Tabulating the data is a useful means of aggregation but it is necessary to explain how the aggregated data actually answer the research questions.

7. Reporting the review (Dissemination)

The final phase of a systematic review involves writing up the results of the review and circulating the results to potentially interested parties.

7.1 Specifying the Dissemination Strategy

It is important to communicate the results of a systematic review effectively. For this reason most guidelines recommend planning the dissemination strategy during the commissioning stage (if any) or when preparing the systematic review protocol.

Academics usually assume that dissemination is about reporting results in academic journals and/or conferences. However, if the results of a systematic review are intended to influence practitioners, other forms of dissemination are necessary. In particular:

1. Practitioner-oriented journals and magazines
2. Press Releases to the popular and specialist press
3. Short summary leaflets
4. Posters
5. Web pages
6. Direct communication to affected bodies.

7.2 Formatting the Main Systematic Review Report

Usually systematic reviews will be reported in at least two formats:

- In a technical report or in a section of a PhD thesis.
- In a journal or conference paper.

A journal or conference paper will normally have a size restriction. In order to ensure that readers are able to properly evaluate the rigour and validity of a systematic review, journal papers should reference a technical report or thesis that contains all the details.

The structure and contents of reports suggested in [19] are presented in Table 8. This structure is appropriate for technical reports and journals. For PhD theses, the entries marked with an asterisk are not likely to be relevant.

7.3 Evaluating Systematic Review Reports

Journal articles will be peer reviewed as a matter of course. Experts review PhD theses as part of the examination process. In contrast, technical reports are not usually subjected to any independent evaluation. However, if systematic reviews are made available on the Web so that results are made available quickly to researchers and practitioners, it is worth organising a peer review. If an expert panel were assembled to review the study protocol, the same panel would be appropriate to undertake peer review of the systematic review report, otherwise several researchers with expertise in the topic area and/or systematic review methodology should be approached to review the report.

The evaluation process can use the quality checklists for systematic literature reviews discussed in Section 5.1.

7.4 Lessons Learned about Reporting Systematic Literature Reviews

Brereton et al. [5] identified two issues of importance during data extraction:

- Review teams need to keep a detailed record of decisions made throughout the review process.
- The software engineering community needs to establish mechanisms for publishing systematic literature reviews which may result in papers that are longer than those traditionally accepted by many software engineering outlets or that have appendices stored in electronic repositories.

Staples and Niazi [27] also emphasize the need to keep a record of what happens during the conduct of the review. They point out that you need to report deviations from the protocol.

With respect to publishing systematic literature reviews, the Journal of Information and Software Technology (http://www.elsevier.com/wps/find/homepage.cws_home) has expressed a willingness to publish systematic literature reviews.

.

Table 8 Structure and Contents of Reports of Systematic Reviews

Section	Subsection	Scope	Comments
Title*			The title should be short but informative. It should be based on the question being asked. In journal papers, it should indicate that the study is a systematic review.
Authorship*			When research is done collaboratively, criteria for determining both who should be credited as an author, and the order of author's names should be defined in advance. The contribution of workers not credited as authors should be noted in the Acknowledgements section.
Executive summary or Structured Abstract*	Context	The importance of the research questions addressed by the review.	A structured summary or abstract allows readers to assess quickly the relevance, quality and generality of a systematic review.
	Objectives	The questions addressed by the systematic review.	
	Methods	Data Sources, Study selection, Quality Assessment and Data extraction.	
	Results	Main finding including any meta-analysis results and sensitivity analyses.	
	Conclusions	Implications for practice and future research.	
Background		Justification of the need for the review. Summary of previous reviews.	Description of the software engineering technique being investigated and its potential importance.
Review questions		Each review question should be specified.	Identify primary and secondary review questions. Note this section may be included in the background section.
Review Methods	Data sources and search strategy		This should be based on the research protocol. Any changes to the original protocol should be reported.
	Study selection		
	Study quality assessment		
	Data extraction		
	Data synthesis		
Included and excluded studies		Inclusion and exclusion criteria. List of excluded studies with rationale for exclusion.	Study inclusion and exclusion criteria can sometimes best be represented as a flow diagram because studies will be excluded at different stages in the review for different reasons.

Results	Findings	Description of primary studies. Results of any quantitative summaries Details of any meta-analysis.	Non-quantitative summaries should be provided to summarise each of the studies and presented in tabular form. Quantitative summary results should be presented in tables and graphs.
	Sensitivity analysis		
Discussion	Principal findings		These must correspond to the findings discussed in the results section.
	Strengths and Weaknesses	Strengths and weaknesses of the evidence included in the review. Relation to other reviews, particularly considering any differences in quality and results.	A discussion of the validity of the evidence considering bias in the systematic review allows a reader to assess the reliance that may be placed on the collected evidence.
	Meaning of findings	Direction and magnitude of effect observed in summarised studies. Applicability (generalisability) of the findings.	Make clear to what extent the results imply causality by discussing the level of evidence. Discuss all benefits, adverse effects and risks. Discuss variations in effects and their reasons (for example are the treatment effects larger on larger projects).
Conclusions	Recommendations	Practical implications for software development.	What are the implications of the results for practitioners?
		Unanswered questions and implications for future research.	
Acknowledgements*		All persons who contributed to the research but did not fulfil authorship criteria.	
Conflict of Interest			Any secondary interest on the part of the researchers (e.g. a financial interest in the technology being evaluated) should be declared.
References and Appendices			Appendices can be used to list studies included and excluded from the study, to document search strategy details, and to list raw data from the included studies.

8 Systematic Mapping Studies

Systematic Mapping Studies (also known as Scoping Studies) are designed to provide a wide overview of a research area, to establish if research evidence exists on a topic and provide an indication of the quantity of the evidence. The results of a mapping study can identify areas suitable for conducting Systematic Literature Reviews and also areas where a primary study is more appropriate. Mapping Studies may be requested by an external body before they commission a systematic review to allow more cost effective targeting of their resources. They are also useful to PhD students who are required to prepare an overview of the topic area in which they will be working. As an example of a mapping study see Bailey et al.'s mapping study which aimed at investigating the extent to which software design methods are supported by empirical evidence [3].

The main differences between a mapping study and systematic review are:

- Mapping studies generally have broader research questions driving them and often ask multiple research questions.
- The search terms for mapping studies will be less highly focussed than for systematic reviews and are likely to return a very large number of studies, for a mapping study however this is less of a problem than with large numbers of results during the search phase of the systematic review as the aim here is for broad coverage rather than narrow focus.
- The data extraction process for mapping studies is also much broader than the data extraction process for systematic reviews and can more accurately be termed a classification or categorisation stage. The purpose of this stage is to classify papers with sufficient detail to answer the broad research questions and identify papers for later reviews without being a time consuming task.
- The analysis stage of a mapping study is about summarising the data to answer the research questions posed. It is unlikely to include in depth analysis techniques such as meta-analysis and narrative synthesis, but totals and summaries. Graphical representations of study distributions by classification type may be an effective reporting mechanism.
- Dissemination of the results of a mapping study may be more limited than for a systematic review; limited to commissioning bodies and academic publications, with the aim of influencing the future direction of primary research.

9 Final remarks

This report has presented a set of guidelines for planning, conducting, and reporting a systematic review. The previous versions of these guidelines were based on guidelines used in medical research. However, it is important to recognise that software engineering research is not the same as medical research. We do not undertake randomised clinical trials, nor can we use blinding as a means to reduce distortions due to experimenter and subject expectations. For this reason, this version of the guidelines has incorporated information from text books authored by researchers from the social sciences.

These guidelines are intended to assist PhD students as well as larger research groups. However, many of the steps in a systematic review assume that it will be undertaken

by a large group of researchers. In the case of a single researcher (such as a PhD student), we suggest the most important steps to undertake are:

- Developing a protocol.
- Defining the research question.
- Specifying what will be done to address the problem of a single researcher applying inclusion/exclusion criteria and undertaking all the data extraction.
- Defining the search strategy.
- Defining the data to be extracted from each primary study including quality data.
- Maintaining lists of included and excluded studies.
- Using the data synthesis guidelines.
- Using the reporting guidelines

In our experience this "light" version of a systematic review is manageable for PhD students. Furthermore, research students often find the well-defined nature of a systematic review helpful both for initial scoping exercises and for more detailed studies that are necessary to position their specific research questions.

10 References

[1] Australian National Health and Medical Research Council. How to review the evidence: systematic identification and review of the scientific literature, 2000. IBSN 186-4960329.

[2] Australian National Health and Medical Research Council. How to use the evidence: assessment and application of scientific evidence. February 2000, ISBN 0 642 43295 2.

[3] Bailey, J., Budgen, D., Turner, M., Kitchenham, B., Brereton, P. and Linkman, S. Evidence relating to Object-Oriented software design: A survey. ESEM07.

[4] Berlin, J.A., Miles, C.G., Crigliano, M.D. Does blinding of readers affect the results of meta-analysis? Online J. Curr. Clin. Trials, 1997: Doc No 205.

[5] Brereton, Pearl , Kitchenham, Barbara A., Budgen, David, Turner, Mark and Khalil, Mohamed. Lessons from applying the systematic literature review process within the software engineering domain. JSS 80, 2007, pp 571-583.

[6] Budgen, David, Stuart Charters, Mark Turner, Pearl Brereton, Barbara Kitchenham and Stephen Linkman Investigating the Applicability of the Evidence-Based Paradigm to Software Engineering, Proceedings of WISER Workshop, ICSE 2006, 7-13, May 2006, ACM Press.

[7] Cochrane Collaboration. Cochrane Reviewers' Handbook. Version 4.2.1. December 2003

[8] Cochrane Collaboration. The Cochrane Reviewers' Handbook Glossary, Version 4.1.5, December 2003.

[9] Cohen, J. Weighted Kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Pychol Bull (70) 1968, pp. 213-220.

[10] Crombie, I.K. The Pocket Guide to Appraisal, BMJ Books, 1996.

[11] Fink, A. Conducting Research Literature Reviews. From the Internet to Paper, Sage Publication, Inc., 2005.

[12] Greenhalgh, Trisha. How to read a paper: The Basics of Evidence-Based Medicine. BMJ Books, 2000.

[13] Hart, Chris. Doing a Literature Review. Releasing the Social Science Research Imagination. Sage Publications Ltd., 1998.

[14] Jasperson, Jon (Sean), Butler, Brian S., Carte, Traci, A., Croes, Henry J.P., Saunders, Carol, S., and Zhemg, Weijun. Review: Power and Information Technology Research: A Metatriangulation Review. MIS Quarterly, 26(4): 397-459, December 2002.
[15] Jørgensen, M., and Shepperd, M. A Systematic Review of Software Development Cost Estimation Studies IEEE Transactions on SE, 33(1), 2006, pp33-53.
[16] Jørgensen, M.A review of studies of expert estimation of software development effort, Journal of Systems and Software, 70, 2002, pp 37-60.
[17] Jørgensen, M. Estimation of Software Development Work Effort: Evidence on Expert Judgment and Formal Models, International Journal of Forecasting, 2007.
[18] Jørgensen, M. Evaluation of guidelines for performing systematic literature reviews in software engineering, version 2.2, 2007
[19] Khan, Khalid, S., ter Riet, Gerben., Glanville, Julia., Sowden, Amanda, J. and Kleijnen, Jo. (eds) Undertaking Systematic Review of Research on Effectiveness. CRD's Guidance for those Carrying Out or Commissioning Reviews. CRD Report Number 4 (2nd Edition), NHS Centre for Reviews and Dissemination, University of York, IBSN 1 900640 20 1, March 2001.
[20] Khan, Khalid, S., Kunz, Regina, Kleijnen, Jos and Antes, Gerd. Systematic Reviews to Support Evidence-based Medicine, The Royal Society of Medicine Press Ltd., 2003.
[21] Kitchenham, B., Mendes, E., Travassos, G.H. (2007) A Systematic Review of Cross- vs. Within-Company Cost Estimation Studies, IEEE Trans on SE, 33 (5), pp 316-329.
[22] Lawlor, Debbie A., George Davey Smith, K Richard Bruckdorfer, Devi Kundu, Shah. Ebrahim Those confounded vitamins: what can we learn from the differences between observational versus randomised trial evidence? The Lancet, vol363, Issue 9422, 22 May, 2004.
[23] Noblit, G.W. and Hare, R.D. Meta-Ethnography: Synthesizing Qualitative Studies. Sage Publications, 1988.
[24] Pai, Madhukar., McCulloch, Michael., Gorman, Jennifer D., Pai, Nitika, Enanoria, Wayne, Kennedy, Gail, Tharyan, Prathap, and Colford, John, M. Jr. Systematic reviews and meta-analyses: An illustrated, step-by-step guide, The National Medical Journal of India, 17(2), 2004, pp 84-95.
[25] Petticrew, Mark and Helen Roberts. Systematic Reviews in the Social Sciences: A Practical Guide, Blackwell Publishing, 2005, ISBN 1405121106
[26] Shadish, W.R., Cook, Thomas, D. and Campbell, Donald, T. Experimental and Quasi-experimental Designs for Generalized Causal Inference. Houghton Mifflin Company, 2002.
[27] Staples, M. and Niazi, M. Experiences using systematic review guidelines. Article available online, JSS.
[28] Sutcliffe, T.J., Harden, K., Oakley, A., Oliver, A., Rees,S., Brunton, R. and Kavanagh, G. Children and Healthy Eating: A systematic review of barriers and facilitators, London, EPPI-Centre, Social Science Research Unit, Institute of Education, University og London, October 2003.
[29] Turner, M., Kitchenham, B., Budgen, D., Charters, S. and Brereton, P. A Systematic Literature Review of the technology Acceptance Model and its Predictive Capabilities, Keele University and University of Durham Joint Technical Report, 2007.

Appendix 1 Steps in a systematic review

Guidelines for systematic review in the medical domain have different view of the process steps needed in a systematic review. The Systematic Reviews Group (UC Berkeley) present a very detailed process model [24], other sources present a coarser process. These process steps are summarised in Table 9, which also attempts to collate the different processes.

Table 9 Systematic review process proposed in different sources

Systematic Reviews Group [24]	Australian National Health and Medical Research Council [1]	Cochrane Reviewers Handbook [7]	CRD Guidance [19]	Petticrew and Roberts [25]	Fink [11]
			Identification of the need for a review. Preparation of a proposal for a systematic review		
Define the question & develop draft protocol Identify a few relevant studies and do a pilot study; specify inclusion/exclusion criteria, test forms and refine protocol.		Developing a protocol	Development of a review protocol		
	Question Formulation	Formulating the problem		Refine questions and define boundaries	Select Research Questions
Identify appropriate databases/sources.	Finding Studies	Locating and selecting studies for reviews	Identification of research	Define Inclusion/Exclusion criteria	Select Bibliographic Databases and Web Sites. Choose Search Terms
Run searches on all relevant data bases and sources. Save all citations (titles/abstracts) in a reference manager. Document search strategy			Selection of studies	Find the primary studies	Find the studies

Researchers (at least 2) screen titles & abstracts. Researchers meet & resolve differences. Get full texts of all articles. Researchers do second screen. Articles remaining after second screen is the final set for inclusion					Apply Practical Screening criteria
Researchers extract data including quality data. Researchers meet to resolve disagreements on data Compute inter-rater reliability. Enter data into database management software	Appraisal and selection of studies	Assessment of study quality	Study quality assessment	Assess study quality	Apply methodological Quality Screen
		Collecting data	Data extraction & monitoring progress		Train Reviewers Pilot the Reviewing Process Do the Review
Import data and analyse using meta-analysis software. Pool data if appropriate. Look for heterogeneity.	Summary and synthesis of relevant studies	Analysing & presenting results	Data synthesis	Synthesize the evidence. Explore heterogeneity and publication bias	Synthesize the results Produce a descriptive review or perform meta-analysis
Interpret & present data. Discuss generalizability of conclusions and limitations of the review. Make recommendations for practice or policy, & research.	Determining the applicability of results. Reviewing and appraising the economics literature.	Interpreting the results	The report and recommendations. Getting evidence into practice.	Disseminate the results	

Appendix 2 Software Engineering Systematic Literature Reviews

Software engineering SLRs published between 2004 and June 2007 that scored 2 or more on University of York, CRD DARE scale as assessed by staff working on the Keele University and Durham University EBSE project.

Author	Date	Title	Reference Details	Topic type	Topic area	Quality Score
Barcelos, R.F., and Travassos, G.H.	2006	Evaluation approaches for Software Architectural Documents: A systematic Review	Ibero-American Workshop on Requirements Engineering and Software Environments (IDEAS). La Plata, Argentina.	Technology evaluation	Software Architecture Evaluation Methods	2.5
Dyba, T; Kampenes, V.B. and Sjoberg, D.I.K..	2006	A systematic review of statistical power in software engineering experiments	Information and Software Technology, 48(8), pp 745-755.	Research trends	Power in SE Experiments	2.5
Glass, R.L., Ramesh, V., and Vessey, I	2004	An Analysis of Research in Computing Disciplines	CACM, Vol. 47, No. 6, pp89-94.	Research Trends	Comparative trends in CS, IS and SE	2
Grimstad, S., Jorgensen, M. and Molokken-Ostvold, K	2006	Software effort estimation terminology: The tower of Babel	Information and Software Technology, 48 (4), pp 302-310	Technology	Cost Estimation	3
Hannay, J E., Sjøberg, D.I.K and Dybå. T	2007	A Systematic Review of Theory Use in Software Engineering Experiments	IEEE Trans on SE, 33 (2), pp 87-107.	Research trends	Theory in SE Experiments	2.5
Jørgensen, M	2004	A review of studies on expert estimation of software development effort,	Journal of Systems and Software, 70 (1-2), pp37-60.	Technology	Cost Estimation	3
Jørgensen, M., and Shepperd, M.	2007	A Systematic Review of Software Development Cost Estimation Studies	IEEE Transactions on SE, 33(1), pp33-53.	Research trends	Cost Estimation	3
Kampenes, V.B., Dybå, T., Hannay, J.E. and Sjøberg, D.I.K. (	2007	A systematic review of effect size in software engineering experiments.	Information and Software Technology, In press.	Research trends	Effect size in SE experiments	2.5
Mair, C. and Shepperd, M.	2005	The consistency of empirical comparisons of regression and analogy-based software project cost prediction	International Symposium on Empirical Software Engineering	Technology evaluation	Cost Estimation	2
Mendes, E.	2005	A systematic review of Web engineering research.	International Symposium on Empirical Software Engineering	Research Trends	Web Research	2

Moløkken-Østvold, K.J., Jørgensen, M. Tanilkan, S.S., Gallis,H., Lien, A.C. and Hove, S.E.	2004	Survey on Software Estimation in the Norwegian Industry	Proceedings Software Metrics Symposium.	Technology evaluation	Cost Estimation	2
Petersson, H., Thelin, T, Runeson, P, and Wohlin, C.	2004	Capture-recapture in software inspections after 10 years research – theory, evaluation and application	Journal of Systems and Software, 72, 2004, pp 249-264	Technology evaluation	Capture-recapture in Inspections	2.5
Runeson, P., Andersson, C., Thelin, T., Andrews, A. and Berling, T.	2006	What do we know about Defect Detection Methods?	IEEE Software, 23(3) 2006, pp 82-86.	Technology evaluation	Testing methods	2
Sjoeberg, D.I.K., Hannay, J.E., Hansen, O., Kampenes, V.B., Karahasanovic, A., Liborg, N.K. and Rekdal, A.C.	2005	A survey of controlled experiments in software engineering	IEEE Transactions on SE, 31 (9), 2005, pp733-753.	Research trends	SE experiments	2
Zannier, C, Melnick, G. and Maurer, F.	2006	On the Success of Empirical Studies in the International Conference on Software Engineering	ICSE06, pp 341-350	Research Trends	Empirical studies in ICSE	3.5

Appendix 3 Protocol for a Tertiary study of Systematic Literature Reviews and Evidence-based Guidelines in IT and Software Engineering

Barbara Kitchenham, Pearl Brereton, David Budgen, Mark Turner, John Bailey and Stephen Linkman

Background

At ICSE04, Kitchenham et al. (2004) Suggested software engineering researchers should adopt "Evidence-based Software Engineering" (EBSE). EBSE aims to apply an evidence-based approach to software engineering research and practice. The ICSE paper was followed-up by a paper at Metrics05 (Jørgensen et al., 2005) and an article in IEEE Software (Dybå et al., 2005).

Following these papers, staff at the Keele University School of Computing and Mathematics proposed a research project to investigate the feasibility of EBSE. This proposal was funded by the UK Economics and Physical Science Research Council (EPSRC). The proposal was amended to include the Department of Computer Science, University of Durham when Professor David Budgen moved to Durham. The EPSRC have now funded a joint Keele and Durham follow-on project (EPIC).

The purpose of the study described in this protocol is to review the current status of EBSE since 2004 using a tertiary study to review articles related to EBSE in particular articles describing Systematic Literature reviews (SLRs)

Evidence-based research and practice was developed initially in medicine because research indicated that expert opinion based medical advice was not as reliable as advice based on scientific evidence. It is now being adopted in many domains e.g. Criminology, Social policy, Economics, Nursing etc. Based on Evidence-based medicine, the goal of Evidence-based Software Engineering is:

> "To provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision making process regarding the development and maintenance of software." (Dybå et al., 2005)

In this context evidence is defined as a synthesis of best quality scientific studies on a specific topic or research question. The main method of synthesis is a Systematic Literature Review (SLR). In contrast to an ad hoc literature review, an SLR is a methodologically rigorous review of research results.

Research Questions

The research questions to be addressed by this study are:

- How much EBSE activity has there been since 2004?
- What research topics are being addressed?
- Who is leading EBSE research?

- What are the limitations of current research?

Search Process

The search process is a manual search of specific conference proceedings and journal papers since 2004. The nominated journals and conferences are shown in the following Table.

Sources to be Searched

Source	Responsible
Information and Software Technology (IST)	Kitchenham
Journal of Systems and Software	Kitchenham
IEEE Transactions on Software Engineering	Kitchenham
IEEE Software	Kitchenham
Communications of the ACM (CACM)	Brereton
ACM Surveys	Brereton
Transactions on Software Engineering Methods (TOSEM)	Brereton
Software Practice and Experience	Budgen & Kitchenham
Empirical Software Engineering Journal (ESEM)	Budgen
IEE Proceedings Software (now IET Software)	Kitchenham
Proceedings International Conference on Software Engineering (ICSE 04, 05, 06, 07)	Linkman & Kitchenham & Brereton
Proceedings International Seminar of Software Metrics (Metrics04, Metrics05)	Kitchenham & Brereton
Proceedings International Seminar on Empirical Software Engineering (ISESE 04, 05, 06)	Kitchenham & Brereton

Specific researchers will also be contacted directly:
Dr Magne Jørgensen
Professor Guilherme Travassos.

Inclusion criteria

Articles on the following topics, published between Jan 1st 2004 and June 30th 2007, will be included

- Systematic Literature Reviews (SLRs) i.e. Literature surveys with defined research questions, search process, data extraction and data presentation
- Meta-analyses (MA)

Exclusion Criteria

The following types of papers will be excluded

- Informal literature surveys (no defined research questions, no search process, no defined data extraction or data analysis process).
- Papers discussing process of EBSE.
- Papers not subject to peer-review.

When an SLR has been published in more than one journal/conference, the most complete version of the survey will be used.

Primary study selection process

The results will be tabulated as follows:

- Number of papers per year per source
- Number of candidate papers per year per source
- Number of selected papers per year per source.

The relevant candidate and selected studies will be selected by a single researcher. The rejected studies will be checked by another researcher. We will maintain a list candidate papers that were rejected with reasons for the rejection.

Quality Assessment

Each SLR will be evaluated using the York University, Centre for Reviews and Dissemination (CDR) Database of Abstracts of Reviews of Effects (DARE) criteria (http://www.york.ac.uk/inst/crd/crddatabase.htm#DARE). The criteria are based on four questions:

- Are the review's inclusion and exclusion criteria described and appropriate?
- Is the literature search likely to have covered all relevant studies?
- Did the reviewers assess the quality/validity of the included studies?
- Were the basic data/studies adequately described?

The questions are scored as follows:

- Question 1: Y (yes), the inclusion criteria are explicitly defined in the paper, P (Partly), the inclusion criteria are implicit; N (no), the inclusion criteria are not defined and cannot be readily inferred.
- Question 2: Y, the authors have either searched 4 or more digital libraries and included additional search strategies or identified and referenced all journals addressing the topic of interest; P, the authors have searched 3 or 4 digital libraries with no extra search strategies, or searched a defined but restricted set of journals and conference proceedings; N, the authors have search up to 2 digital libraries or an extremely restricted set of journals.
- Question 3: Y, the authors have explicitly defined quality criteria and extracted them from each primary study; P, the research question involves quality issues that are addressed by the study; N no explicit quality assessment of individual papers has been attempted.

- Question 4: Y Information is presented about each paper; P only summary information is presented about individual papers; N the results of the individual studies are not specified.

The scoring procedure is Y=1, P=0.5 and N or Unknown=0.

The data will be extracted by one researcher and checked by another.

Data Collection

The data extracted from each paper will be:

- The source (i.e. the conference or journal).
- The year when the paper was published. Note if the paper was published in several difference sources both dates will be recorded and the first date will be used in any analysis. This is necessary in order to track the EBSE activity over time.
- Classification of paper
    - Type (Systematic Literature Review SLR, Meta-Analysis MA).
    - Scope (Research trends or specific research question).
- Main software engineering topic area.
- The author(s) and affiliation (organisation and country).
- Research question/issue.
- Whether the study referenced an EBSE paper or the SLR Guidelines (Kitchenham, 2004).
- Whether the study resulted in evidence-based practitioner guidelines.
- The number of primary studies used in the SLR/MA
- Summary of paper.
- Quality score for the study.

The data will be extracted by one researcher and checked by another.

Data Analysis

The data will be tabulated (ordered alphabetically by the first author name) to show the basic information about each study. The number of studies in each major category will be counted.

The tables will be reviewed to answer the research questions and identify any interesting trends or limitations in current EBSE-related research as follows:

- Question 1 How much EBSE activity has there been since 2004? This will be addressed by simple counts of the number of EBSE related papers per year.
- Question 2 What research topics are being addressed? This will be addressed by counting the number of papers in each topic area. We will identify whether any specific topic areas that have a relatively large number of SLRs.
- Question 3 Who is leading EBSE research? We will investigate whether any specific organisation of researches have undertaken a relatively large number of SLRs.
- Question 4 What are the limitations of current research? We will review the range of SE topics, the scope of SLRs and the quality of SLRs to determine

whether there are any observable limitations. We will also investigate whether the quality of studies is increasing over time by plotting the quality score against the first publication date, and whether the quality of studies has been influenced by the SLR guidelines (by comparing the average quality score of SLRs that referenced the guidelines with the average score of SLRs that did not reference the guidelines).

Dissemination

The results of the study should be of interest to the software engineering community as well as researchers interested in EBSE. For that reason we plan to report the results on a Web page. We will also document the full result of the study in a joint Keele University and University of Durham technical report. A short version of the study will be submitted to IEEE Software.

References

1. Barbara Kitchenham, Tore Dybå and Magne Jørgensen. (2004) Evidence-based Software Engineering. Proceedings of the 26th International Conference on Software Engineering, (ICSE ’04), IEEE Computer Society, Washington DC, USA, pp 273 – 281 (ISBN 0-7695-2163-0).
2. Kitchenham, B. Procedures for Performing Systematic Reviews. Joint Technical Report, Keele University TR/SE-0401 and NICTA 0400011T.1, July 2004.
3. Tore Dybå, Barbara Kitchenham, and Magne Jørgensen. Evidence-based Software Engineering for Practitioners, IEEE Software, Volume 22 (1) January, 2005, pp58-65.
4. Magne Jørgensen, Tore Dybå, and Barbara Kitchenham. Teaching Evidence-Based Software Engineering to University Students, 11th IEEE International Software Metrics Symposium (METRICS'05), 2005, p. 24.