Empir Software Eng (2010) 15:618–653
DOI 10.1007/s10664-010-9134-8

Refining the systematic literature review process—two participant-observer case studies

Barbara A. Kitchenham · Pearl Brereton · Mark Turner · Mahmood K. Niazi · Stephen Linkman · Rialette Pretorius · David Budgen

Published online: 25 June 2010

Editor: James Miller

Abstract Systematic literature reviews (SLRs) are a major tool for supporting evidence-based software engineering. Adapting the procedures involved in such a review to meet the needs of software engineering and its literature remains an ongoing process. As part of this process of refinement, we undertook two case studies which aimed 1) to compare the use of targeted manual searches with broad automated searches and 2) to compare different methods of reaching a consensus on quality. For Case 1, we compared a tertiary study of systematic literature reviews published between January 1, 2004 and June 30, 2007 which used a manual search of selected journals and conferences and a replication of that study based on a broad automated search. We found that broad automated searches find more studies than manual restricted searches, but they may be of poor quality. Researchers undertaking SLRs may be justified in using targeted manual searches if they intend to omit low quality papers, or they are assessing research trends in research methodologies. For Case 2, we analyzed the process used to evaluate the quality of SLRs. We conclude that if quality evaluation of primary studies is a critical component of a specific SLR, assessments should be based on three independent evaluators incorporating at least two rounds of discussion.

Keywords Systematic literature review · Mapping studies · Broad search · Targeted search · Manual search · Automated search · Case study · Quality evaluation process

1 Introduction

In 2004–5, Kitchenham, Dybå and Jørgensen wrote three papers suggesting that the concept of evidence-based practice, (as initially developed in medicine, and subsequently adopted

B. A. Kitchenham (✉) · P. Brereton · M. Turner · M. K. Niazi · S. Linkman
School of Computing and Mathematics, Keele University, Stoke-on-Trent ST5 5BG, UK
e-mail: b.a.kitchenham@cs.keele.ac.uk

R. Pretorius · D. Budgen
School of Engineering and Computing Sciences, Durham University, Durham DH1 3LE, UK

Table 1 Stages in evidence-based software engineering

Stages	Activities
1	Converting the need for information (about development and maintenance methods, management procedures etc.) into an answerable question.
2	Tracking down the best evidence with which to answer that question.
3	Critically appraising that evidence for its validity (closeness to the truth), impact (size of the effect), and applicability (usefulness in software development practice).
4	Integrating the critical appraisal with our software engineering expertise and with our stakeholders' values and circumstances.
5	Evaluating our effectiveness and efficiency in executing Steps 1–4 and seeking ways to improve them both for next time.

by many different disciplines including economics, psychology, social science and most health care disciplines) should be adopted in software engineering (Kitchenham et al. 2004; Dybå et al. 2005; Jørgensen et al. 2005). By analogy with medicine, they suggested that evidence-based software engineering (EBSE) should be concerned with the aggregation of empirical evidence and should use systematic literature reviews (SLRs) as a methodology for performing unbiased aggregation of empirical results. Based on the stages in evidence-based medicine, Kitchenham et al. (2004) suggested equivalent stages in EBSE which are shown in Table 1. Stage 5 is about seeking ways to improve the way in which we undertake evidence-based software engineering and is the rationale for this paper. In particular, we believe it is important to assess the impact of different SLR procedures used by software engineering researchers, in order to improve the advice given to researchers in our own SLR guidelines (Kitchenham and Charters 2007).[1]

We are currently undertaking a program of case study-based research that is aimed at better understanding the role of systematic literature reviews (SLRs) in software engineering (Brereton et al. 2007). This is part of the Evidence-based Practices Informing Computing (EPIC) project which is funded by the UK Engineering and Physical Sciences Research Council.

This study uses numerous terms adopted from evidence-based medicine that are not widely used in empirical software engineering. We, therefore, provide a glossary of terms in Table 2, but will also provide a definition when we first use them.

SLRs are secondary studies (i.e. studies that are based on analyzing previous research) used to find, critically evaluate and aggregate all relevant research papers (referred to as primary studies) on a specific research question or research topic. The methodology is intended to ensure that the review is unbiased, rigorous and auditable. The basic SLR methodology is similar, irrespective of discipline using it; although medical standards emphasize *meta-analysis* (a means of statistically aggregating the results from different studies of the same phenomena) more than other disciplines (see for example Fink 2005; Petticrew and Roberts 2005; Khan et al. 2003; Kitchenham and Charters 2007).

We are using the *participant-observer* case study approach as our main research methodology for investigating software engineering SLRs. A participant-observer case study is a case study where some (or all) of the individuals conducting the case being observed are also part of the case study research team. The cases in each case study are

[1] This and other related technical reports are available from our website: www.ebse.org.uk

Table 2 Glossary of terms

Term	Meaning
Automated search	A search of digital libraries and electronic indexing systems using search strings aimed at finding candidate primary studies.
Broad search	A search intended to identify as many candidate primary studies as possible either by searching a large number of journals and conference proceedings manually and/or by automated searches of a several different digital sources including digital libraries and electronic indexing systems. Such a search may include searching references in primary studies and direct appeal to experts in the field (Fink, 2002).
Case study	An empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident (Yin, 2003).
Grey literature	Grey literature refers to papers that have not been published in a source where there is a full peer review process, including papers such as technical reports, PhD and masters theses. In the context of this paper, we consider workshop papers and book chapters to be grey literature since such papers may not have been subject to a full editorial review process. Such papers may (or may not) have been reviewed but it is likely that there was no requirement that the authors respond to the reviewers' comments.
Mapping study or scoping study	A SLR aimed at identifying and classifying the research available in a specific topic area.
Manual search	A search carried out on specific journals and conference proceedings by a researcher who identifies whether the paper is a candidate primary study based on title, keywords and abstract. (Note the journal or conference proceedings may be online or hard copies.)
Meta analysis	A SLR where the outcomes of primary studies are aggregated quantitatively.
Participant-observer case study	A case study in which some or all of the case study researchers are involved in the phenomenon being studied.
Primary study	One of the set of studies used in an SLR to answer the SLR's research question.
Publication bias	The problem that journals and conferences are more likely to accept for publications studies that show a positive effect of some method/procedure than papers that show no effect.
Quality criteria	A set of concepts (usually in the form of questions) that are used to assess the quality of a primary study.
Quality data	The answers to the quality criteria for each primary study.
Quality score	After assigning numerical values to individual quality questions, the total score for a primary study is the sum of the individual numerical values.
Restricted / targeted search	A search that targets a specific set of journals and/or conference proceedings on the assumption that these are the most important and or best quality sources of candidate primary studies.
Secondary study	A study that is based on an analysis of other research papers. SLRs, Meta-analyses and Mapping studies are forms of secondary study.
Systematic Literature Review (SLR)	A methodology for finding and aggregating empirical studies in order to answer a defined research question, which aims to be unbiased, auditable and repeatable.
Tertiary study	A study that is based on an analysis of secondary study results. In other words a form of secondary study where the primary studies are in fact secondary studies.

SLRs performed by the EPIC research group who act as SLR participants as well as case study researchers (see Fig. 1). We report two case studies in this paper.

Case Study 1 (CS1) reports the progress of an SLR aimed at extending an existing *tertiary* study (i.e. a systematic review of secondary studies) that surveyed SLRs in the time period 1st Jan 2004 to 30th June 2007. The original tertiary study (referred to as TS1) restricted its search process to a set of 13 journals and conferences (Kitchenham et al. 2009a). The case directly observed in CS1 extends the search to a large number of digital libraries (this SLR will be referred to as TS2).

Medical guidelines for performing SLRs recommend broad search procedures including *automated searches* (i.e. searches that apply search strings electronically to digital libraries and electronic research indexing systems) and efforts to identify any relevant *grey literature* (Khan et al. 2003), where grey literature refers to papers that have not been published in a source where there is a full peer review process. However, SE researchers have taken somewhat different approaches to the SLR search process in different published SLRs. For example, some researchers have restricted their searches to specific digital libraries (Juristo et al. 2006) or a specific set of journals and conference proceedings (Sjøberg et al. 2005). In addition, Jørgensen and Shepperd (2007) have strongly advocated the use of *manual searches* (i.e. searchers where the individual researchers scan the contents of selected journals and conference proceedings) as opposed to automated searches.

CS1, which is based on extending an existing tertiary study using a broader search process, gives us the opportunity to investigate the following research questions:

- RQ1 (Breadth of Literature Search): To what extent is the adoption of an extended search space vital for answering detailed research questions?
- RQ2 (Importance of Grey Literature): To what extent is the grey literature necessary for SE SLRs?
- RQ3 (Manual versus Automated Search Strategies): Are automated search strategies preferable to manual search strategies in the SE domain?

Case Study 2 (CS2) investigates the process used to evaluate the quality of primary studies. The medical guidelines make it clear that it is important to identify which are the most trustworthy primary studies in terms of the rigor of the study methodology. However, although the medical standards recommend using two researchers per primary study and addressing disagreements, they do not reference any evidence-based guidelines for the quality evaluation process.

There is a considerable literature on the reviewing processes used to assess journals and conference papers which has been summarized by Weller (2001) and updated by Bornmann

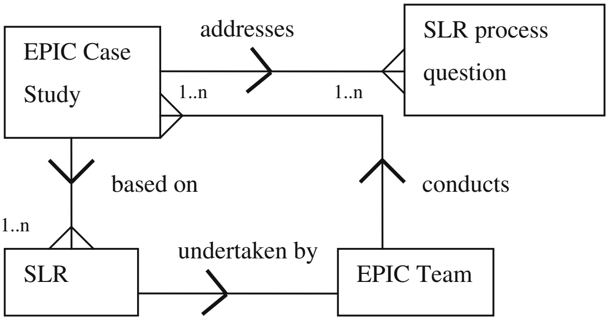

Fig. 1 EPIC case study methodology

et al. (2010). Overall, it seems that there is a need for multiple reviewers (usually more than 2) supported by well-defined evaluation criteria, where more objective criteria lead to better inter-rater reliability than more subjective criteria. Other researchers have reviewed the literature on assessing research and investigated the use of *bibliometric* indicators, such as number of publications and number of citations (Martin and Irvine 1983; Martin 1996). The problem with such indicators is that they relate to *impact* not *quality*. Martin suggests bibliometrics need to be used in conjunction with peer review. However, we found no studies that assessed the procedures for evaluating the quality of individual papers for the purposes of systematic reviews. Since we believe quality assessment of individual primary studies is important, we undertook Case Study 2 to investigate several different quality evaluation procedures.

CS2 was based on a tertiary study cataloguing and classifying systematic literature reviews published in the time period 1st July 2007–30th June 2008 (Kitchenham et al. 2009a). This SLR will be referred to as TS3. In TS1 we adopted an extractor and checker process where a single researcher answered the *quality criteria* (i.e. questions about the study related to aspects of study quality) for each SLR and the answers were checked by another researcher. However, the results of another SLR suggested that the extractor/checker process was not always effective (Turner et al. 2008). In TS2, we used the median of three independent extractions (see Section 2). However, we were unsure how accurate a subjective assessment based on the median would be. In TS3 we undertook a multi-stage quality extraction process that allowed us to compare a variety of different quality extraction methods to investigate the following research question:

> RQ4 (Quality process): Among the processes investigated, what is the best process for assessing the quality of primary studies?

We describe our methodology in Section 2, and present the results of CS1 in Section 3 and the results of CS2 in Section 4. We discuss our results and present our conclusions in Section 5. In using the participant-observer methodology, we are also aiming to assess its appropriateness for our SLR evaluation studies, and for software engineering evaluation studies where a formal experiment is not practical.

CS1 was initially discussed in a paper presented at the Empirical Software Engineering and Measurement conference (Kitchenham et al. 2009b). The additional contribution of this paper is that we discuss quality evaluation options in CS2 and, in addition to recommendations concerning quality evaluation procedures, this case study also provides a second example of our observer-participant case study methodology. Other examples of our case study approach can be found in Kitchenham et al. (2009c) and Kitchenham et al. (2010a).

2 Methodology

We investigated our research questions using participant-observer case studies based primarily on the methodology proposed by Yin (2003). A participant-observer case study has several advantages:

1. There is no problem about access to information about the case.
2. There is no requirement to liaise with external organisations or individuals.

However, there is the possibility of bias when investigating an issue in which the researchers have a vested interest. With respect to vested interests, we are very much in favour of the use of systematic literature reviews and our research questions are based on the assumption that SLRs are useful. However, we merely seek to determine the most appropriate procedures for performing SLRs not to investigate their value as a scientific methodology, so our personal bias will have limited impact in these case studies. In addition, for Case Study 1, we have specified our methodology clearly in a protocol prior to undertaking the case study (Kitchenham et al. 2008a). In the next sections, we describe the methodology used for both case studies.

2.1 Case Study 1 (CS1)

As recommended by Yin (2003), the case study process was planned in advance. The initial research questions of interest were defined in advance for our entire research program, but were refined when we found a specific "case" that was able to address some of those research questions. The format of our case study protocol was based on a case study protocol template (Brereton et al. 2008).

2.1.1 The Case and Basic Design

The "case" in this study is a *tertiary* SLR i.e. it is a study that uses secondary studies (SLRs) as the object of study. It is also a *mapping study.* A mapping study is a form of systematic review that asks general questions about research in a topic area (e.g. what do we know about topic *x*) rather than specific questions about research outcomes (e.g. is method *a* better than method *b*). The original tertiary mapping study, TS1 (Kitchenham et al. 2009a), restricted its search process to a manual search of a set of 13 journals and conferences during the time period 1st Jan 2004 to 30th June 2007. The selected sources were based on those used by Sjøberg et al. (2005). In the subsequent tertiary study, TS2, we replicated TS1 using a broad automated search process which searched both digital libraries and on-line indexing systems. Since we have a baseline "case", with which to compare the results of the broad automated search, our case study design can be categorised as a *multi-case* case study. Furthermore since we investigated specific SLR tasks (i.e. searching, selection and analysis) and analysed individual conclusions, we also regard this as an *embedded* case study (i.e. a case study that investigates various sub-elements of the case).

2.1.2 CS1 Propositions

The Research Questions and their related *propositions* (using the terminology adopted by Yin, 2003) are shown in Table 3. Propositions in case studies play a similar role to hypotheses in formal experiments. They are developed from our understanding of the subject under investigation. These propositions were developed when we started scoping our overall research program (Brereton et al. 2007), but were revised when the specific case was identified.

2.1.3 CS1 Roles

As shown in Fig. 1, the EPIC team was involved both in the case study and in the SLR. We assigned roles as follows:

- SLR Supervisor: David Budgen
- SLR Research team including a Research Assistant (Rialette Pretorius) responsible for most of the SLR activities with support from Pearl Brereton, Barbara Kitchenham, Stephen Linkman, Mark Turner, Mahmood Niazi

Table 3 Case study propositions and their rationale

Case Study 1 Research Questions	ID	Propositions	ID	Rationale and Limitations
(Breadth of Literature Search): To what extent is the adoption of an extended search space vital for answering detailed research questions?	RQ1	A broad automated search will identify more relevant primary studies than a restricted manual search.	P1.1	Finding more primary studies is the basic argument for broad searches.Limitation: We have investigated only broad automated searches compared with restricted but targeted manual searches. Our study therefore does not generalize to any type of broad search strategy.
		Additional primary studies found by a broad automated search will change the conclusions of the study, even if low quality studies are removed.	P1.2	An argument in favour of targeted searches is that they will find high quality studies. The implication of this is that additional studies found by broad searches (outside the scope of the targeted searches) may be dominated by poor quality studies that should be rejected anyway. The argument in favour of broad searches is the risk of missing high quality studies that would change study results found by targeted searches. This proposition addresses these conflicting viewpoints.
(Importance of Grey Literature): To what extent is the grey literature necessary for SE SLRs?	RQ2	Primary studies not published in journals or conference proceedings are of equivalent quality to other primary studies.	P2.1	This proposition addresses directly the quality of studies found outside a targeted search. Limitation: We have not systematically searched the grey literature but have simply considered the type of paper (journal, conference, workshop, book chapter) as a surrogate for whether the paper can be considered grey literature or not.
(Manual versus Automated Search Strategies): Are automated search strategies preferable to manual search strategies in the SE domain?	RQ3	Broad automated searches require less effort than restricted manual searches.	P3.1	An argument in favour of automated searches is that they are more efficient than manual searches. This proposition addresses this issue.Limitation: The comparison of manual and automated searches is confounded with the scope of the search. The manual search targeted a specific set of conference proceedings and journals.

- Case Study leader (Kitchenham)
- Case Study Team (Budgen, Brereton, Linkman, Niazi, Turner).

The SLR supervisor was responsible for organizing the SLR and supervising the Research Assistant (RA) responsible for performing the SLR. He was also responsible for ensuring that the RA collected information about the SLR process as required by the case study. (We name the roles of each researcher because we believe that attempts at anonymity are inappropriate in participant-observer case studies.)

The case study team leader was responsible for constructing the case study protocol (Kitchenham et al. 2008a) and the SLR protocol (Kitchenham et al. 2008b). Members of the case study team including the case study leader provided research support for the SLR process (i.e. assisting as required with primary study identification, quality data extraction and SLR data extraction). The case study team was responsible for documenting the results of the case study.

The data extraction and preliminary selection process for the SLR took much longer than expected, so Kitchenham took over the organization of the SLR after the RA's internship finished. The RA completed the initial search process, and initial screening of papers to remove obviously inappropriate studies. Kitchenham organized the two further screening activities and the data extraction and aggregation processes.

2.1.4 Case Limitations

We note that this "case" is not a typical software engineering SLR:

- It is a tertiary study, not a conventional secondary study.
- The subject of the study is the SLR methodology not a software technology.
- It is a mapping study, not a conventional SLR looking at detailed a research questions(s).
- It is a study where, due to the topic, relatively few additional primary studies were expected.

In addition, the case only compares broadening the search using an automated search of six digital sources, with manual search of journal and conference proceedings that are a subset of articles referenced by three electronic sources (ACM, IEEE and SCOPUS). This differs from the broad manual search proposed by Jørgensen and Shepperd (2007). The results of this case study must, therefore, be interpreted carefully in the light of the specific case.

2.1.5 SLR Methodology Used for the Two Tertiary Studies

The SLR used in this case study replicated the original SLR, in terms of research questions, but extended the search space by undertaking an automated search of four digital libraries and two general indexing services. A comparison of the procedures used in the two tertiary studies is shown in Table 4. In this section, we specify the individual search strings and the quality criteria.

The search strings used for all sources except SCOPUS were a set of 15 simple strings:

1. "Software engineering" AND "review of studies"
2. "Software engineering" AND "structured review"
3. "Software engineering" AND "systematic review"
4. "Software engineering" AND "literature review"

Table 4 Comparison of the procedures used in the two tertiary studies

SLR Process	TS1 Procedures	Comments	TS2 Procedures	Comments
Research Questions	RQ1: How much EBSE activity has there been since 2004?		RQ1: How many SLRs were published between 1st January 2004 and 30th June 2008?	The first three research questions addressed exactly the same issues (although the time span differed for RQ1).
	RQ2: What research topics are being addressed?		RQ2: What research topics are being addressed?	For RQ4, TS1 identified certain limitations that were reviewed in TS2.
	RQ3: Who is leading EBSE research?		RQ3: Which individuals and organizations are most active in SLR-based research?	For RQ5, TS1 analyzed the quality of the SLRs even though it was not a formal research question. The quality analysis was upgraded to a formal research question in TS2.
	RQ4: What are the limitations of current research?		RQ4: Are the limitations of SLRs, as observed in the original study, still an issue?	
			RQ5: Is the quality of SLRs improving?	
Primary Search Process	Manual search of 13 conferences and journals selected because they were used by other researchers assessing research methods (Sjøberg et al., Glass et al.) or we knew included published SLRs.	Four researchers performed manual searches based on availability and access to source.	Automated search of 4 digital libraries and two indexing services IEEE Computer Society Digital Library; ACM; Citeseer; SpringerLink; SCOPUS; Web of Science. Pretorius performed the majority of the searches. Kitchenham undertook the SCOPUS search.	This set of sources is wider than that suggested by Hannay et al. who suggested using ACM, IEEE, ISI Web of Science and EI Compendex with the aim of minimising duplicate references. We used Citeseer instead of Google Scholar because we found the interface better. Like Google Scholar Citeseer uses a web crawler to find documents and a filter to reject papers that do not meet certain academic criteria. It also shares some problems with Google Scolar: 1. It is difficult to know what journals and conferences it searches.2. Searches may not be repeatable.

			Pretorius tested a variety of different search strings (on all sources except SCOPUS). The final selection of search strings was the selection that identified the maximum number of known studies (i.e. those found in the first tertiary study). The search strings are reported in Section 2.1.5	
Secondary Search Process	Searched website of predominant EBSE researcher & e-mailed another researcher known to be undertaking SLRs.		None	We did not search primary studies citations because in contrast to primary studies in conventional SLRS which are likely to reference other papers on the same topic, we did not expect systematic literature reviews to reference other SLRs. That is, we assumed different SLRs would not address exactly the same topic, so would have no need to cite one another.
Number of papers searched	2506	This included all papers excluding editorials and keynotes in the journals and conferences included in the search.	1757	The number of papers returned by the final set of search strings.
Selection Process	Selection was performed by researchers while doing the manual search.	Literature surveys rejected by original researcher were noted by the first reviewer and subsequently checked by another member of the research team.	The selection process involved the following stages: 1. Pretorius screened the outcomes of the searches to remove obviously irrelevant papers (based on title, keywords and abstract) or duplicate references. This left 161 candidate papers. 2. Kitchenham organised two further rounds of screening.	The process is fully described in Kitchenham et al. (2010b). The selection process was generally much more formal than TS1 although the initial screening was based on the opinion of one person.

Table 4 (continued)

SLR Process	TS1 Procedures	Comments	TS2 Procedures	Comments
			a. The first round was based on title, abstract & keywords. Kitchenham reviewed every paper, the other papers were reviewed by two other researchers assigned randomly from the pool of 5 (excluding Pretorius). The emphasis was on removing obviously irrelevant papers. b. The second round used a similar process but the full paper was read and a final consensus agreement was reached for each paper.	
Inclusion Criteria	The paper should: 1. Be a literature survey or meta-analysis or include a literature survey. 2. Have a research question, a defined search process and a defined extraction process.		The paper should: 1. Be a literature survey or meta-analysis or include a literature survey. 2. Have a defined search process. 3. Be related to software engineering rather than computer science or information systems.	Less stringent inclusion criteria than the initial tertiary study with respect to literature survey rigour.
Exclusion Criteria	The paper should not: Be an abstract or PowerPoint presentation. Be an informal literature review with no research question, no defined search process or no defined data extraction process. Be published prior to 2004 or after June 30th 2008.		The paper should not: Be an abstract or PowerPoint presentation. Be an opinion survey. Be an informal literature review with no defined search process. Be published prior to 2004 or after June 30th 2007.	The search process covered an extra year compared with the first case.

	Use a language other than English.		Use a language other than English.	
Number of Candidate Primary Studies	32	All candidate studies found by the manual search. Duplicate reports were referenced in the same study.	161 (after preliminary screening) 119 papers after 2nd round of screening 34 papers after the third round of screening.	If duplicate reports were published in different sources only the most recent/complete journal publication was included.
Agreed Primary Studies	20	Including two studies identified by the secondary search process.	14 in time period 1st January 2004 to June 30th 2007. 20 in time period 1st July 2007 to 30th June 2008.	Only the 14 studies in the same time period as the TS1 are considered in the case study.
Data Extraction Process (including quality extraction)	Kitchenham extracted the information from each paper. The extractions were checked by one of the other researchers (assigned at random).		Three researchers extracted information from each paper. Two researchers were assigned at random from the pool of four researchers while Kitchenham reviewed all the papers. The median value was taken to represent the consensus view.	The data and quality extraction was more rigorous in TS2.
Analysis	The data were tabulated and aggregated to answer the research questions and analyse the quality results (Kitchenham et al., 2009).	The results were reviewed and each individual result statement identified for the purposes of the case study.	The data were tabulated and aggregated to answer the research questions (Kitchenham et al. 2010a).	The data were analysed to address the case study research questions.

5. “Software engineering” AND “literature analysis”
6. “Software engineering” AND “in-depth survey”
7. “Software engineering” AND “literature survey”
8. “Software engineering” AND “meta analysis”
9. “Software engineering” AND “past studies”
10. “Software engineering” AND “subject matter expert”
11. “software engineering” AND “analysis of research”
12. “Software engineering” AND “empirical body of knowledge”
13. “Evidence-based software engineering” OR “evidence based software engineering”
14. “Software engineering” AND “overview of existing research”
15. “Software engineering” AND “body of published research”

The SCOPUS search compressed the set of strings into two more complex queries. A variety of search strings based on terminology used in the SLRs found in TS1 were tested. The SLRs found in TS1 were used to validate the search strings with the final selection being based on the one that found the largest number of known SLRs.

Quality evaluation for both TS1 and TS2 used the Centre for Reviews and Dissemination DARE criteria (2007), which are based on four quality criteria questions:

- Q1: Are the review’s inclusion and exclusion criteria described and appropriate?
- Q2: Is the literature search likely to have covered all relevant studies?
- Q3: Did the reviewers assess the quality/validity of the included studies?
- Q4: Were the basic data/studies adequately described?

We scored each question on a scale of Yes (1), No (0), Partly (0.5) and summed the scores. The DARE criteria concern the rigour of a systematic review in terms of the extent to which it is repeatable (Q1), complete (Q2), able to deliver trustworthy conclusions (Q2 and Q3) and auditable (Q4). These criteria can only be assessed subjectively although we used some guidelines to improve the reliability of the subject evaluation, 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 searched up to 2 digital libraries or an extremely restricted set of journals. Note that scoring question 2 also requires the quality evaluator to consider whether the digital libraries were appropriate for the specific SLR.
- 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.

2.1.6 CS1 Data Collection

We saw no reason why the RA should be blinded to the goals of this case study, so she was given a copy of the case study protocol to ensure that she was aware of data collection requirements placed on her by the demands of the case study.

To address the case study research questions the following data were collected:

- The number of new primary studies identified by the automated search. This was collected by the RA as part of the search process.
- The type of new primary studies (i.e. journal papers, conference papers, book chapters, workshop papers, technical reports). This information was provided by the digital libraries when the citations were found. It was extracted by Kitchenham as part of the SLR data extraction process.
- Information about each change to the results and conclusions of the study due to the additional literature. This information was collected by Kitchenham as part of the study aggregation and reporting process.
- Time taken to complete SLR tasks. This was collected by the RA and SLR Team as part of the agreed SLR process.

The data analysis and interpretation procedures for the case study were specified in advance in our protocol, but managed both to be too simplistic (at the detailed propositions level) and too complex (at the interpretation level) to cope with the data we collected. In particular, counting the number of primary studies found by each search strategy was more complicated than we had expected, so simple interpretations based on the percentage of studies missed was inappropriate.

2.2 Case Study 2 (CS2)

The “case” in CS2 is the quality evaluation stage of a tertiary mapping SLR extending TS2 to cover the time period 1st July 2007 to 30th June 2008. The SLR will be referred to as tertiary study 3 (TS3).

Unlike CS1 which was planned in advance, CS2 was an opportunistic case study. When we considered the process we would use to extract the quality data for TS3, we decided to use three researchers per paper and this gave us the opportunity to evaluate different methods of combining quality data.

2.2.1 The Case and Basic Design

TS3 used the outcome of the search process of TS2 i.e. the 26 papers found by the broad automated search but not analysed as part of TS2. CS2 concentrates on the application of the data extraction process to the quality data. Thus, the case study can be described as a single case embedded case study.

2.2.2 CS2 Propositions and Roles

The basic propositions for the research question RQ4 (Among the processes investigated, what is the best process for assessing the quality of primary studies?) were:

1. P4.1: Aggregating evaluations from researchers improves accuracy. This is the basic justification for using two or more individuals to extract data (particularly subjective data).

2. P4.2: Numerical aggregation is worse than aggregation based on discussion. Medical guidelines advocate using discussion to arrive at consensus, so a standard assumption must be that discussion is important for achieving consensus.
3. P4.3: The more rigorous the evaluation method the more time consuming it will be. We assume that aggregations based on a number of extractions and discussions will be more effort intensive than simple numerical aggregation.

The roles in CS2 were the same as those in CS1.

2.2.3 Limitations of the Case

There are two main differences between quality criteria evaluation in TS3 and quality evaluation in standard software engineering SLRs:

1. The quality criteria used in CS2 are questions for SLRs not quality criteria for conventional primary studies.
2. There were only 4 quality questions whereas examples of quality criteria used in conventional SE SLRs included 11 or 12 questions (e.g. Kitchenham et al. 2007; Dybå, and Dingsøyr 2008).

2.2.4 Methodology Used for TS3

The basic search process, study selection process, and quality criteria were the same as those used in TS2 (in fact the search process and initial selection process for TS3 was integrated with the TS2 search process). After analysing the 14 SLRs for CS2, we undertook data extraction for the 26 TS3 primary studies. However, after further scrutiny of the SLRs, we included 21 papers in the final data extraction activity (see Kitchenham et al. 2010b for full details). Two of the papers were published after 30th June 2008 (Beecham et al. 2008; Gómez et al. 2008) and were excluded from the SLR report but are included in this analysis. There were two major differences between the data extraction process used in TS2 and TS3:

1. In TS3 subjective data was extracted using a multi-step process (see below)
2. Non-subjective data (i.e. author names, affiliations, publication information etc.) were collected by one person (Kitchenham).

The multi-step process (which we refer to as a "consensus and minority report" process) was applied to the four quality questions and four other data items (type of SLR, the focus of the SLR, the number of primary studies and the topic of the SLR). In terms of quality assessment, the process involved:

- Three researchers each providing an assessment of each quality question. The first two researchers were assigned at random to each paper from the TS2 researchers (excluding Kitchenham and Pretorius). Kitchenham acted as the third researcher for all papers.
- Two rounds of agreement. Firstly, the two researchers agreed a joint assessment (called the *initial consensus*); then they reviewed the third researcher's extraction and revised their joint assessment, if it was required to create the *final consensus*.

An additional difference between the quality assessments in this study compared with the TS2 is that each researcher was instructed to provide a justification for his/her score for each quality question when they first extracted data from the SLR. We also refined the guidelines for assessing question 3. If the SLR reported a quality evaluation but did not incorporate the quality evaluation into the aggregation process, we scored question 3 as No (0).[2] This is somewhat harsh but it is pointless to perform a quality evaluation of primary studies and then ignore the results when attempting to aggregate the results.

2.2.5 CS2 Data Collection and Analysis

The multi-step evaluation process led to six quality scores for each SLR, numbered as follows:

1. Quality score produced by researcher R1.
2. Quality score by researcher R2.
3. Joint quality score produced by R1 & R2 after consultation (Initial consensus).
4. Quality score produced by researcher R3 (Minority report)
5. Overall quality score made by researchers R1 & R2 after reviewing the quality score produced by R3 (Final consensus).
6. The median of the three initial quality scores.

Kitchenham extracted the six quality scores for each primary study as part of the data extraction process. Data analysis was planned on the assumption that the final consensus would represent the best possible evaluation of the quality of each primary study and the results obtained from other steps in the process would be compared to the final consensus.

3 Results of Case Study 1 (CS1)

We present CS1 results and analysis in this section. The data extracted from each additional paper found during TS2 are summarized in Table 5. The original papers are reported in Kitchenham et al. (2009a). In Table 5 the “Article” column indicates whether the study was an SLR or a mapping study (MS). The “EBSE” column identifies whether the paper referenced either of the Evidence-Based Software Engineering papers (Kitchenham et al. 2004; Dybå et al. 2005) or the SLR Guidelines technical reports (Kitchenham and Charters 2007; Kitchenham 2004). We refer to such papers as “EBSE-related” papers. The “PG” column indicates whether the study recommended practitioner guidelines. The Total Quality score is the value obtained using the DARE criteria (Centre for Reviews and Dissemination2007). “Survey type” indicates whether the study addressed a specific research question (RQ) or was concerned with general research trends (RT). For SLRs the survey type is usually RQ and for mapping studies it is usually RT, but there can be exceptions.

3.1 RQ1—Breadth of Search

3.1.1 P1.1: A Broad Automated Search Will Identify More Relevant Primary Studies than a Restricted Manual Search

The comparison of the original restricted search and the extended broad automated search are shown in Table 6 and are shown diagrammatically in Fig. 2. Although overall the

[2] This happened only once.

Table 5 Data extraction for papers found by TS2

Authors	Year	Article	Total Quality Score	EBSE	Type	Number of Primary Studies	PG	Survey type	Topic
Grimstad et al.	2005	SLR	2.5	Y	Conference	8	N	RQ	Cost estimation
Jørgensen	2005	SLR	2	N	Journal	70	Y	RQ	Cost estimation
Davis et al.	2006	SLR	3.5	Y	Conference	26	Y	RQ	Requirements Engineering
Shepperd	2007	MS	1.5	Y	Workshop	653	N	RT	Cost estimation
Mair et al.	2005	SLR	2.5	Y	Workshop	50	N	RT	Cost estimation
Yalaho	2006	MS	1.5	N	Workshop	57	N	RT	Outsourcing
Kagdi et al.	2006	MS	1.5	N	Journal	80	N	RT	Mining Software Repositories
Segal et al.	2005	MS	1.5	N	Workshop	119	N	RT	Empirical Software Engineering
Höst et al.	2005	SLR	2	Y	Conference	13	N	RT	Empirical SE methods & Inspections experiment
Hosbond and Nielsen	2005	MS	2.5	N	Conference	105	N	RT	Mobile Systems Development
Shaw and Clements	2006	MS	1	N	Tchnical Report	750	N	RT	Software Architecture
Davis et al.	2007	MS	1.5	N	Book Chapter	4089	N	RT	Requirements Engineering
Höfer and Tichy	2007	MS	1.5	N	Book Chapter	133	N	RT	Empirical Software Engineering
Feller et al.	2006	MS	2	N	Book Chapter	155	N	RT	Open Source Software Development

proposition that broad automated searches find more papers than restricted searches is supported, the results are more complicated than the simple proposition suggests.

The broad automated search missed 5 studies included in TS1. Two of those studies were not discovered by the manual search process (Barcelos and Travassos 2006; Jørgensen

Table 6 How the SLRs used in the tertiary studies were identified

Paper counts	Number
Studies used in TS1 found by manual search	18
Studies used in TS1 not found by manual search	2
Total Studies found by TS2	29
Studies found by searches both in TS1 and in TS2	15
Extra studies found by TS2	14
Studies found by TS1 but not TS2	5
Studies found by TS2 that should have been found by TS1	3
Extra studies found by TS2 but not directly by the broad automated search	1

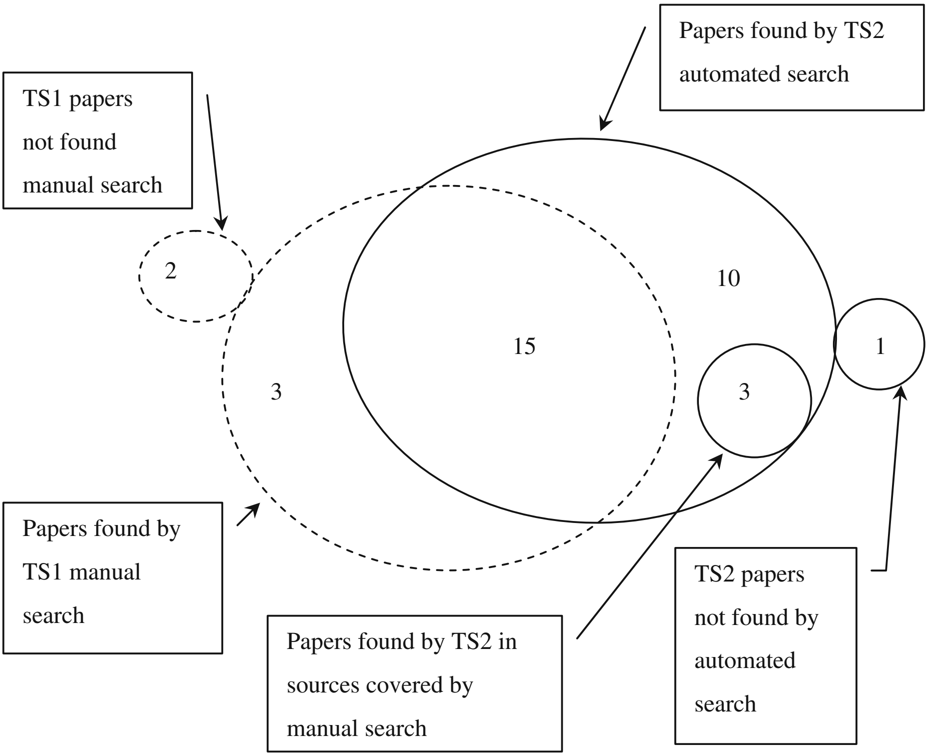
Papers found by TS2 automated search
TS1 papers not found manual search
2
10
15
1
3
3
Papers found by TS1 manual search
Papers found by TS2 in sources covered by manual search
TS2 papers not found by automated search

Fig. 2 Venn diagram comparing TS1 and TS2 search results

2007). They were found by contacting one researcher (Travassos) and searching the web site of another (Jørgensen), both of whom we knew were involved in undertaking SLRs. For completeness, most SLR standards advise contacting researchers known to be active in the field to identify whether they have any new results that may not yet be published (Fink, 2005; Petticrew and Roberts 2005). In the case of more conventional SLRs (i.e. those investigating a specific research question or research topic), it is usually clear from initial searches whether there are specific research groups that are interested in the topic who can be approached to check that all their primary studies have been identified. In the case of tertiary studies, we could only check with the researchers we knew were interested in undertaking SLRs.

Of the other three papers, one had an embedded review that was borderline for inclusion (Torchiano and Morisio 2004); one used the term “review” but not “literature review” (Jørgensen 2004); and the final paper was a literature review of computer science papers and should probably have been omitted from the initial study (Ramesh et al. 2004).

The original manual search missed three papers from its set of 13 sources: one journal paper (Jørgensen 2005) and two conference papers (Grimstad et al. 2005; Höst et al. 2005). One of those papers would have been excluded from the original study because it did not have a defined data extraction process (Jørgensen 2005). Nonetheless, missing relevant papers suggests that the process of having only one researcher search each source was not as effective as it should have been.

Overall the broad automated search identified 11 papers that were published in the sources not searched in the original study. However, this is slightly misleading:

- Two of the studies selected (Yalaho 2006; Shaw and Clements 2006) would not have been included in the original tertiary, because they did not have a clear data extraction and aggregation process, although they did have a defined search process.
- Two of the studies that investigated empirical software engineering (Segal et al. 2005; Höfer and Tichy 2007) considered only one source (i.e. the Empirical Software Engineering journal). This was because previous literature surveys related to empirical studies in software engineering had omitted ESE from their list of sources (Tichy et al. 1995; Glass et al. 2002). Thus, whether these studies count as ancillary studies or mapping studies in their own right is problematic.

3.1.2 P1.2: Additional Primary Studies Found by a Broad Automated Search Will Change the Conclusions of the Study Even if Low Quality Papers are Removed

The results of the original study are shown in Table 7. In 11 of 17 cases, the original results were confirmed or strengthened. They were contradicted in 6 cases. We have based our comparison on results rather than on conclusions because it is easier and more objective to map individual results to one another rather than to try to compare wider conclusions. Furthermore, changed results are the underlying reason for changed conclusions.

The quality results presented in Tables 2 and 7 suggested that the additional studies found in the broad automated search were of relatively poor quality. The average quality scores for the studies are shown in Table 8. Using the Mann-Whiney rank sum test, the quality score for the additional 14 studies was significantly less than the quality scores of the initial 20 studies ($p<0.001$). Excluding the quality scores from the three studies that should have been found in the original search process makes no difference to the results. It is therefore relevant to consider what would have happened if poor quality papers were omitted from the aggregated data.

Excluding the two results related to quality (since removing low quality papers renders most results related to overall quality issues invalid) four of the original results were contradicted by the additional studies. Table 9 shows that when the poor quality studies are removed, only the researcher who was involved in most studies, after Jørgensen, still contradicts the results of the original study.

3.2 RQ2: The Importance of Grey Literature

3.2.1 P2.1: Primary Studies are of Equal Quality Irrespective of Source

The median quality scores for studies reported in different types of article are shown in Table 10. It appears that articles that we classified as grey literature (i.e. workshop papers, book chapters, and technical reports) are of lower quality than papers published in conferences and journals. This is confirmed by a Mann-Whitney rank sum test comparing grey literature with other literature ($p<0.01$). However, this result must be treated with some caution. The main concern with grey literature is that it is not properly peer-reviewed or that any peer-reviews may be less stringent resulting in lower quality studies. This is not always the case in Software Engineering workshops where some are closer to conferences in terms of the rigour of their review process. In fact one of the workshop papers scored 2.5 on the DARE scale. We do not know the review policy of each workshop so we cannot be

Table 7 Comparisons of results

Research question in original study	Results in original study	Results including all relevant papers	Impact on results
Result not related to a specific research question.	Quality is improving over time.	No relationship between quality and time.	Results contradicted
Result not related to a specific research question.	Quality not better for EBSE-related papers.	No relationship between quality and citing EBSE-related papers.	No change
4.1 How much EBSE activity has there been since 2004?	Of 20 studies, 10 cited Guidelines or EBSE paper.	Of 34 studies 15 cited Guidelines or EBSE paper.	Less EBSE-positioned research.
	Stable numbers per year: 2004 (6); 2005 (5); 2006 (6); 2007 (3) where 2007 covers only half year.	Probable increase in 2007: 2004 (6); 2005 (11); 2006 (9); 2007 (8). 2007 covers only half year.	Recent activity underestimated.
	Main publication source: IEEE SW (4 studies); TSE (4); JSS (3); IST (2).	Main publication source changes: TSE (5); IEEE SW (4); JSS (3); IST (2); Metrics (2); ICSE (2).	No major change. Changed counts due to papers missed in original search.

4.2 What topics are being addressed	Topics addressed at least twice:
	Cost estimation (7 papers);
	SE Empirical Methods (4); Testing (3)
		Cost estimation papers (11 papers);
		SE Empirical Methods (7); Testing (4);
		Requirement Engineering (2);
		Architecture (2).
			Results changed. More work on conventional SE topics.

4.3 Who is leading research	Studies per person: Jørgensen (5)	Studies per person: Jørgensen (8)	No change
	After Jørgensen, Sjöberg (3)	After Jørgensen, Shepperd (4)	Results change
	Most activity organization: Simula Laboratory (8)	Most activity organization: Simula Laboratory (11)	No change

Table 7 (continued)

Research question in original study	Results in original study	Results including all relevant papers	Impact on results
	Most studies have European authors (14 of 20).	Most studies have European authors (26 of 34).	No change
	Few studies have US authors (3 of 20).	Few studies have US authors (7 of 34).	No change
4.4 Current Limitations of SLRs	Many research trends papers (8 of 20).	Many research trends papers (19 of 34).	Original result strengthened.
	Limited number of conventional SE topics addressed.	Requirements and software architecture addressed.	More work on conventional SE topics, but general conclusion holds.
	Sample sizes for conventional SLRs relatively small (range 6–54) compared with mapping studies (range 63–1485).	Sample sizes for conventional SLRs relatively small median (21) compared with mapping studies (median 105).	No change
	Quality relatively good: Only 3 studies scored less than 2 on the DARE scale.	10 of 34 papers scored less than 2 on the DARE scale.	Original results contradicted.
	Few studies addressed the quality of primary studies. 3 fully; 5 partially; 12 not at all.	4 fully; 6 partially; 24 not at all.	No change
	Few studies provided practitioner guidelines (4 of 20).	6 of 34	No change

Table 8 Quality scores for studies found in TS1 and TS2

Data Source	Studies	Median quality score
Studies found in TS1	20	2.5
Studies found in broad automated search in sources used in original search	3	2
Studies found in broad automated search but in sources other than original ones	11	1.5

more specific with our analysis. This ambiguity also affects papers identified as book chapters. Springer-Verlag "book chapters" are often proceedings of conferences and workshops. Although we know that one of the papers came from a workshop where there was no explicit review process (i.e. the workshop on Empirical Software Engineering Issues), we do not know about the other two.

3.3 RQ3: Manual Versus Automated Searches

3.3.1 P3.1: Automated Searches Require Less Effort than Manual Searches

We do not have timesheets for the original restricted manual search, but we estimate that it took about 4 h to review a specific source and about 15 min per paper to look over the 12 disputed papers (see Table 4). This gives a total of 56 h to perform the search (although we missed 3 papers). The effort for the automated search is itemised in Table 11. Note that only the RA kept detailed timesheets, effort values for the other team members are based on post-hoc estimates.

Additional costs and time accrued because we were unable to find 12 papers online. Thus to complete our second round of screening we had to obtain the papers via inter-library loans. This took about four elapsed weeks in all (although the time period included Christmas).

Overall, our results suggest that a broad automated search requires much more effort than a restricted manual search. This result would still hold if the time for the manual search was doubled to allow two researchers to check each source. We discuss the manual and automated search in more detail below to explain this rather unexpected result.

Table 9 Effect on results of removing low quality papers

Results from original SLR	Evidence from original SLR	Evidence from all studies with a quality score of 2 or more	Impact of new evidence
Stable number of papers per year	2004 (3); 2004 (5); 2006 (6); 2007 (3)	2004 (3); 2004 (10); 2006 (7); 2007 (4)	Original result confirmed
Many studies were evidence-based SE articles	10 of 20 referenced evidence based SE articles or SLR guidelines.	13 out of 24 papers referenced evidence based SE articles or SLR guidelines.	Original result confirmed
Topics addressed at least twice	Cost Estimation (7); Empirical SE (4); Testing (2).	Cost Estimation (10); Empirical SE (5); Testing (3).	Original result confirmed
Other active researchers	Sjöberg (3)	Shepperd (4)	Original result changed

Table 10 Quality scores for study types in Ts12 and TS2

Data Source	Studies	Median quality score
Journals in original set of sources	15	2.5
Conferences in original set of sources	6	2.5
Journals not in original set of sources	2	2.5
Conferences not in original set of sources	2	3
Workshop studies (including one from original study)	5	1.5
Book Chapters	3	1.5
Technical report	1	1

The manual search involved individual researchers looking through all the papers in each of 13 specific journal and conference proceedings published between 1st January 2004 and 30th June 2007. In all cases the same researcher reviewed papers published in a specific source. Sources were searched on-line with the exception of IET Software which was searched using the printed journals. Thus, in all cases, it was simple to view the abstract and title and if necessary consult the full version of the paper. Since for each publication the search was a simple sequential task it could be stopped and started at any point without requiring any iteration. Also, since full versions of the papers were accessible, the initial inclusion/exclusion process was integrated with the search process.

In contrast, the automated search required several different stages:

1. The papers used in TS1 were associated with the digital library that indexed the journal/conference in which they appeared. Various different search strings were developed and tested on each library to find the maximum possible number of known studies. This involved many different searches and manually checking all outcomes against the relevant known papers.
2. The set of 15 search strings were applied to each of the six digital libraries and indexing systems.
3. The outcomes of all the individual searches were collated and duplicates removed.

Apart from the actual searches, none of the above tasks were automated. Furthermore, although the comparison with the set of studies found in TS1 was effort intensive, it is a normal method of validating search strings, so should be considered an integral part of an automated search process.

Table 11 Effort for selection and screening papers for the broad automated search

Activity	Effort (hours)
Search String specification and testing (Pretorius)	46.5
Search & collating papers found by searches (Pretorius)	117
Initial candidate selection (Pretorius)	161
Organising the screening process (including assigning papers to researchers and collating results) (Kitchenham)	13
Finding papers (Kitchenham, Brereton, Budgen)	11.25
1st screening (All except Pretorius)	14
2nd screening (All except Pretorius)	8.5
Total	357.25

The difficulty encountered with collating papers from different digital libraries raises the issue of whether it is better to search individual digital libraries or to use general indexing systems. In principle, automated searches of a single wide-scope indexing system such as ISI Web of Science or SCOPUS would reduce the collation problem significantly. However, the SCOPUS search found only 9 of the 20 papers included in the original tertiary study and two of the additional 14 papers found in broad automated search. One problem is that general indexing systems only allow searches that are based on title, abstract and keywords, whereas the individual digital libraries can base searches on the contents of the full paper.

4 Results of Case Study 2 (CS2)

We present the CS2 results and analysis in this section. The individual quality scores for each paper are shown in Table 12. R1 and R2 are the scores produced by one of five different researchers where the pairing of individual researchers differs for different papers. R1 refers to the scores produced by the researchers who were assigned to the paper first and R2 refers to the scores made by the researchers who were assigned second. R3, however, refers to the scores produced by a single researcher (i.e. Kitchenham). The *Initial Consensus* identifies the scores produced by researchers R1 and R2 after consultation and discussion. The *Median* is the median value of the scores obtained by R1, R2 and R3, whereas the *Final Consensus* is the score obtained by allowing R1 and R2 to compare their

Table 12 Quality Scores of SLRs obtained by different quality evaluation procedures

Id	Reference	R1	R2	Initial Consensus	R3	Median	Final consensus
1	Hannay and Jørgensen (2008)	2.5	2	2.5	2	2.5	2.5
2	Neto et al. (2008)	3	3.5	3.5	4	3.5	3.5
3	Liebchen and Shepperd (2008)	3	2	3	2.5	2.5	3
4	de Boer and Farenhorst (2008)	2.5	3	3	3	3	3
5	Beecham et al. (2008)	4	3.5	3.5	3.5	4	4
6	Freire et al. (2007)	3	2.5	2.5	3	3	3.5
7	Wicks and Dewar (2007)	1.5	0.5	1	2.5	1.5	1
8	Pino et al. (2008)	3	2.5	2.5	2.5	3	2.5
9	Kampenes et al. (2007)	2	2	2	2.5	2.5	2.5
10	Mohagheghi and Conradi (2007)	3	2.5	3	2.5	3	2.5
11	Bellini et al., (2008)	1	2.5	1.5	1.5	1.5	2
12	Staples and Niazi (2008)	3.5	3.5	3.5	3	3.5	3.5
13	Hanssen et al. (2007)	3.5	3	3.5	3.5	3.5	3.5
14	Gómez et al. (2008)	2.5	2.5	2.5	1.5	3	2.5
15	Zhang et al. (2008)	3	3.5	3	3	3	3
16	Renger et al. (2008)	1.5	1.5	1.5	2	1.5	1.5
17	Harjumaa et al. (2008)	2	1.5	1.5	2	1.5	1.5
18	Mohagheghi and Dehlen (2008)	2	2	1.5	3	2.5	2.5
19	Jefferies et al. (2008)	2	2	2	3	2	2.5
20	Bailey et al. (2007)	2.5	2.5	2.5	2.5	2.5	2.5
21	MacDonell and Shepperd (2007)	4	4	4	4	4	4

Initial Consensus with the scores given by R3 and make any final adjustments they thought appropriate.

Table 13 shows correlations among the quality scores obtained by different quality evaluation processes. The correlations among the scores for individual researchers are lower than the correlations among the aggregated scores. The scores for R1 and R2 correlate more highly with the aggregated scores than the scores for R3. This is because the scores for R1 and R2 contributed to all the aggregated scores whereas R3 did not contribute to the Initial Consensus at all, contributed only one third to the Median and contributed perhaps less than one third to the Final Consensus. Overall R1 appears to be more highly correlated with the Final Consensus and the Median than the other individual assessments (R2 and R3).

If we assume that the Final Consensus is the most rigorous we could make, it is clear that the Initial Consensus (based on agreement between two raters) and the Median assessment (based on the median of three raters without any agreement process) are both strongly correlated with the Final Consensus. However, it appears that the results for R1 are as good as the Initial Consensus and the Median.

A problem with an analysis based on the total scores is that high correlations can be caused by "accidental correctness", that is a person scoring the four questions as P,Y,P,Y respectively would obtain the same score as a person scoring the questions Y,P,Y,P. To address this issue, we also analyzed the number of scoring disagreements. Table 14 shows a count of the number of disagreements for the 4 questions used to assess quality of each SLR. Note we do not make any adjustment for the disagreement being half a point or one point. Considering half a point differences would lead to a confusion between a paper with one difference of 1 point and another paper with two differences of half a point.

Table 14 confirms the correlation analysis. Both the Median and Initial Consensus are closer to the Final Consensus than assessments provided individually by R1, R2 or R3. The Wilcoxon ranksign test confirms that Median assessment is significantly more similar to the Final Consensus assessment than R1 (p=0.034), R2 (p=0.003), or R3 (p=0.011). With respect to the Initial Consensus, the ranksign test confirms that the Initial Consensus is more similar to the Final Consensus than R1 (p=0.026) or R2 (p=0.017) but is not significantly different for R3 (p=0.061).

The results shown in Table 14 identify two papers (ids 7 and 14) where there was considerable disagreement between researcher R3 and the Final Consensus. In these cases, a fourth opinion was sought for each paper and each of the three researchers who initially assessed the papers was asked to re-extract the quality data. In the case of paper 7, the fourth opinion and R3's re-assessment confirmed the Final Consensus. In the second case (paper 14), the fourth opinion and the R3's opinions coincided and both disagreed with the Final Consensus agreed by R1 and R2. The fourth researcher then chaired a meeting with researchers R1 and R2 and they finally agreed a score of 1.5 for the paper.

Table 13 Spearman rank correlation among quality scores (all correlations are significant P<0.001)

Variable	R1	R2	R3	Initial consensus	Median	Final consensus
R1	1					
R2	0.78	1				
R3	0.66	0.66	1			
Initial Consensus	0.93	0.86	0.68	1		
Median	0.92	0.89	0.72	0.92	1	
Final Consensus	0.89	0.83	0.81	0.90	0.90	1

Table 14 Between score disagreement identifying the number of times the answers to individual questions differed

Id	R1 v R2	R1 v. Initial Consensus	R1 v. Median	R1 v. Final Consensus	R2 v. Initial Consensus	R2 v. Median	R2 v. Final Consensus	R3 v Initial consensus	R3v. Median	R3 v. Final consensus	Median v Final consensus	Initial v Final consensus
1	1	0	0	0	1	1	1	1	1	1	0	0
2	1	1	1	1	0	0	0	1	1	1	0	0
3	2	0	1	0	2	2	2	1	0	1	0	0
4	1	1	1	1	0	0	0	0	0	0	0	0
5	1	1	1	1	0	0	0	2	1	1	0	1
6	1	1	0	1	0	1	1	1	1	0	1	1
7	2	1	0	1	1	2	1	3	2	3	1	0
8	1	1	0	1	0	1	1	2	1	2	1	0
9	2	0	1	1	2	1	3	3	2	2	2	1
10	3	0	0	1	3	3	3	1	1	0	1	1
11	3	1	1	2	2	1	1	2	0	1	1	2
12	0	0	0	0	0	0	0	1	1	1	0	0
13	0	0	0	0	1	1	1	0	0	0	0	0
14	2	0	1	1	2	2	2	4	3	3	1	1
15	2	1	1	1	1	1	1	0	0	0	0	0
16	0	0	0	0	0	0	0	1	1	1	0	0
17	1	0	1	1	1	2	2	2	1	1	0	1
18	2	1	1	1	1	1	1	3	1	1	0	1
19	0	0	0	1	0	0	1	2	2	1	1	1
20	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0
Mean	1.19	0.43	0.48	0.71	0.81	0.90	1	1.43	0.90	0.95	0.43	0.48

4.1 P4.1: Aggregating Evaluations from Researchers Improves Accuracy

Overall R1 appears to be closer to the Final Consensus and the Median than the other individual assessments (R2 and R3). However R1 corresponds to a group of 5 different researchers and the same five researchers contributed to the R2 assessment. This implies that although individual researchers can sometimes perform as well as a consensus approach, this cannot be guaranteed. Thus, P1 is partially supported.

4.2 P4.2: Numerical Aggregation is Worse than Aggregation Based on Discussion

Tables 14 and 13 confirm that the accuracy of the Median and the Initial Consensus are very similar. Therefore, in this case, P2 is not supported.

4.3 P4.3: The More Rigorous the Evaluation Method the More Time Consuming it Will be

The average time spent on data extraction (including quality and other data) per SLR was almost exactly the same for TS2 and for this study: 0.41 h per SLR for TS2 based on 42 extractions and 0.42 h per SLR for TS3 based on 75 extractions. Each researcher apart from Kitchenham extracted less data in TS3 than in TS2 but the omitted data was objective data about the authors, author affiliations and study publication details, so was not very time consuming. Therefore, in this case, P3 is not supported.

5 Discussion and Conclusion

5.1 Case Study 1 (CS1)

Overall CS1 indicates that:

- A broad automated search finds more relevant studies than a restricted manual search.
- Additional papers will cause some results to be revised. In this case, 6 of 17 results were revised.
- Removing poor quality papers may reduce the number of revised results. In this case, three fewer results were revised (i.e. only 1 of 15 non-quality related results).
- Book chapter and workshop papers may be of relatively poor quality, so excluding them will be equivalent to excluding low quality papers. The additional 14 papers found by the broad automated search, included 8 such literature studies of which only two scored two or more on the DARE quality scale.
- Broad automated searches take more time and effort than restricted manual searches.

The broad automated search found seven good quality studies that were not detected by the manual search (although two of those studies should have been found by the original search). However, with respect to this case study, the impact of the broad automated search on the study results (other than completeness) was rather limited once low quality papers were removed.

5.1.1 Implications of CS1 Results

Overall these results suggest that researchers would be justified in adopting a restricted manual search if they are intending to exclude low quality studies from their results.

However, we note that any restricted search must be targeted to an appropriate set of sources.

Clearly this conclusion has limitations related to the nature of the case used in this case study. Our case study is based on a tertiary study investigating research trends of a general research methodology. In such a study, *publication bias* (i.e. the problem that papers that find no statistically significant results are less likely to be published than papers that find significant results) is unlikely to be a problem. In contrast, for a conventional SLR looking at a specific research question such as whether one technology is better than another, publication bias is a potential problem. In such a case, the grey literature is likely to be of much greater importance. Thus, restricted manual searches are more justifiable for studies of research trends than for studies of competing SE methodologies.

Research trend studies are usually mapping studies. Jørgensen and Shepperd (2007) report results from a broad manual search used for a mapping study of a specific software engineering topic (i.e. cost estimation) and warn against restricted searches because important studies may be omitted. Thus, another issue when considering the use of restricted manual searches relates to the importance of completeness. For studies investigating general research trends (e.g. the extent of empirical validation), or research trends related to a research methodology (such as the use of formal experiments) a restricted manual search may be appropriate, but in order to identify all relevant research on a specific SE topic a broad search strategy is likely to be required.

TS1 found two additional papers by approaching researchers and TS2 found an extra paper by reviewing the references of an excluded paper. This supports the recommendations that *any* basic search strategy aiming at completeness, whether manual or automated, should also include searching primary study references and contacting individual researchers (Fink 2005; Petticrew and Roberts 2005).

Finally, we found that targeted manual searches took less effort than automated searches primarily because so much of the search and selection process was not, in fact, automated. We have commented before on the problems of software engineering search engines (Brereton et al. 2007), but it is seems clear that we need better tools to support the overall systematic review process.

5.1.2 CS1 Limitations

Our search strings were designed to find the maximum number of known studies. They included some obvious terms such as “literature review”, “literature survey” and “systematic review” but also included terms used by the known SLRs to describe themselves. It is possible that we missed some SLRs that used other terminology.

Many of our results rely on being able to assess the quality of the primary studies. We used the DARE criteria because they are relatively straightforward (having only four main questions). Nonetheless there are other suggestions for evaluating SLRs (e.g. Greenhalgh 2000) and we cannot be sure that the results would be the same if we had used other quality criteria. In addition, there was considerable disagreement among researchers with respect to answering the individual quality questions; there was only one case in which all three researchers assessed each of the four quality questions identically.

We cannot make any excessive claims for the completeness of this study. We have excluded non-English papers and made no attempt to look for PhD theses that include systematic literature reviews. However, the search process was comparable with the most extensive automated search processes found in the SLRs identified in TS1.

In addition, there were subtle differences between TS1 and TS2. For example, the inclusion criteria were more stringent in TS1, but other aspects of the search process were more rigorous in TS2, particularly the screening method and data extraction processes. We have pointed out specific issues in Table 4 and believe that the effects on the case study propositions, with the exception of primary study counts, are relatively minor, but when undertaking a participant-observer case study there is always a danger that personal opinions and preferences might cloud our judgment.

Another possible cause of bias is the fact that we ourselves analyzed the impact of the additional tertiary studies. In particular, our interpretation of the differences made by the additional studies could have been influenced by a desire to demonstrate that our previously published results were reliable. To address this issue, we have presented our analysis of the changes to the results in some detail in order to make the analysis as transparent as possible.

Finally, there are several problems with our choice of "case". We pointed out in Table 3 the limitations attached to our propositions given the specific case, and in Section 2.1.4 we noted the differences between our "case" and a typical SLR. However, there are seldom perfect cases readily available to researchers, so it is necessary to balance the value of the research questions against the representativeness of the case. The main advantage of our "case" is that it is not an artificial example, i.e. we performed the second tertiary study as a research project in its own right. Another advantage of the study is that it addresses a topic of current concern, because individual researchers are making different decisions about how best to organize their searches. In spite of the limitations, the case study has allowed us to gain some insight into the issues associated with broad and restricted searches that might help other researchers make better-informed choices about their search strategies for SLRs.

5.2 Case Study 2 (CS2)

Overall the CS2 confirms the need for multiple researchers to evaluate subjective issues such as the quality of primary studies. It is perhaps more surprising that a simple numerical aggregate of subjective assessments proved as accurate as a consensus-making process based on two different independent assessments. However, the median might not have been so good if researchers had not been asked to record the reason for their score for each question. A sobering point, however, is that for 9 of the 21 papers the Initial Consensus was revised as a result of the third researcher's assessment. Furthermore two papers required the opinion of a fourth researcher before a reliable consensus was achieved. This confirms how difficult it is to evaluate subjective issues such as quality.

5.2.1 Implications of CS2 Results

Evaluation of the quality of primary studies is a critical part of EBSE and SLRs, but it is difficult. Every effort must be made to ensure such evaluations are as rigorous as possible. For most SLRs a consensus based on two independent assessments should be sufficient. For SLRs where primary study quality is one of the research questions, even more rigorous processes may be necessary such as the "consensus and minority report" approach we used in TS3. Using such a process, papers that cause particular difficulties can be identified and subjected to additional assessment.

5.2.2 Limitations of CS2

As noted in Section 2.2.3, the DARE criteria involve only 4 quality questions, whereas quality criteria for other types of primary study may include far more questions. It is likely, however, that the impact of a larger set of questions would have resulted in larger disagreements among individual researchers and would have provided more support for multiple independent assessments. Furthermore, we cannot be sure whether a simple median of three scores would perform as well with a larger set of quality questions.

5.3 Overall Conclusions and Future Research

In this paper we report the results of investigating two SLR procedures (searching the literature and quality evaluation). Our EPIC research program is investigating how to adapt evidence-based methods (in particular SLRs) to the SE domain (Brereton and Kitchenham 2007). We plan to continue investigating SLR procedures to address step 5 of the EBSE process. In particular, we are undertaking another case study aimed at replicating a previous SLR but undertaking a broader search process. We are also planning a series of studies comparing formal and informal literature reviews.

Our approach to investigating the SLR process has been to adopt a participant-observer case study methodology. Although our research is very specialized, we think the basic methodology has a wider application. Many software engineering processes and technologies are large scale and cannot easily be studied empirically using formal experiments. In such cases researchers often undertake a trial use of a process or technology to assess its value. Such evaluations are a kind of informal case study. We suggest formally adopting the discipline of a participant-observer case study would improve the rigour of such evaluations. However, the methodology is neither simple to undertake not straightforward to report. The main problem we found in CS1 was to ensure that the data needed for the case study were not forgotten when the researchers' main concern was to perform the tasks needed for the tertiary study itself.

Acknowledgment This study was funded by the UK Engineering and Physical Sciences Research Council project EPIC/E046983/1. We thank the anonymous reviewers for their helpful comments.

References

Bailey J, Budgen D, Turner M, Kitchenham B, Brereton P, Linkman S (2007) Evidence relating to Object-Oriented software design: A survey, Proceedings of Empirical Software Engineering and Measurement, IEEE Computer Society Press pp. 482–484

Barcelos RF, Travassos GH (2006) Evaluation approaches for Software Architectural Documents: A systematic Review, Ibero-American Workshop on Requirements Engineering and Software Environments (IDEAS). La Plata, Argentina

Beecham S, Baddoo N, Hall T, Robinson H, Sharp H (2008) Motivation in Software Engineering: A systematic literature review. IST 50:860–878

Bellini CG, Pereira RDCDF, Becker JL (2008) Measurement in Software Engineering from the Roadmap to the Crossroads. Int J Softw Eng Knowl 18(1):37–64

Bornmann L, Mutz R, Daniel D-D (2010)A reliability-generalization study of journal peer reviews—a multi-level analysis of inter-rater reliability and its determinants. (Submitted)

Brereton OP, Kitchenham BA (2007) The Scope of EPIC Case Studies. EPIC technical Report EPIC-2007–04

Brereton OP, Kitchenham BA, Budgen D, Turner M, Khalil MA (2007) Lessons from applying the Systematic Literature Review process within the Software Engineering domain. J Syst & Softw 80 (4):571–583

Brereton P, Kitchenham B, Budgen D, Li Z (2008) Using a Protocol Template for Case Study Planning. Proceedings of EASE 2008, BCS-eWIC

Centre for Reviews and Dissemination (2007) What are the criteria for the inclusion of reviews on DARE? Available at http://www.york.ac.uk/inst/crd/faq4.htm (accessed 24 July 2007)

Davis A, Dieste O Hickey A, Juristo N, Moreno AM (2006) Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review, 14th IEEE International Requirements Engineering Conference (RE'06), pp. 179–188

Davis A, Hickey A, Dieste O, Juristo N, Moreno AM (2007) A Quantitative Assessment of Requirements Engineering Publications–1963–2006, LNCS 4542/2007. Requirements Engineering, Foundation for Software Quality, pp 129–143

de Boer RC, Farenhorst R (2008) In search of 'Architectural Knowledge', SHARK '08: Proceedings of the 3 rd international workshop on Sharing and reusing architectural knowledge, May, pp 71–78

Dybå T, Dingsøyr T (2008) Empirical studies of agile software development: A systematic review. IST 50:833–859

Dybå T, Kitchenham B, Jørgensen M (2005) Evidence-based Software Engineering for Practitioners. IEEE Softw 22(1):58–65

Feller J, Finnegan P, Kelly D, MacNamara M (2006) Developing Open Source Software: A Community-Based Analysis of Research, IFIP International Federation for Information Processing, 208/2006, Social Inclusion: Societal and Organizational Implications for Information Systems, pp 261–278

Fink A (2005) Conducting Research Literature Reviews. Conducting Research Literature Reviews. From the Internet to Paper. 2nd Edition Sage Publications Ltd.

Freire AP, Goularte R, Fortes RPM (2007) Techniques for developing more accessible web applications: a survey towards a process classification, SIGDOC '07: Proceedings of the 25th annual ACM international conference on Design of communication, October, pp 162–169

Glass RL, Vessey I, Ramesh V (2002) Research in software engineering: an analysis of the literature. Inf Softw Technol 44:491–506

Gómez O, Oktaba H, Piattini M, García F (2008) A Systematic Review Measurement in Software Engineering: State-of-the-Art in Measures ICSOFT 2006, CCIS 10. Lect Notes Comput Sci 5007:165–176

Greenhalgh Trisha (2000) How to read a paper: The Basics of Evidence-Based Medicine. BMJ Books

Grimstad S, Jorgensen M, Møløkken-Østvold K (2005) The Clients' Impact on Effort Estimation Accuracy in Software Development Projects, 11th IEEE International Software Metrics Symposium (METRICS'05), pp 3

Hannay J, Jørgensen M (2008) The Role of Deliberate Artificial Design Elements in Software Engineering Experiments. IEEE Trans Softw Eng 34(2):242–259

Hanssen GK, Bjørnson FO, Westerheim H (2007) Tailoring and Introduction of the Rational Unified Process. EuroSPI 2007, LNCS 4764, pp 7–18

Harjumaa L, Markkula J, Oivo M (2008) How does a Measurement Programme Evolve in Software Organizations? PROFES 2008. LNCS 5089:230–243

Höfer A, Tichy WF (2007) Status of Empirical Research in Software Engineering, in V. Basili et al., (Eds) Empirical Software Engineering Issues, Springer-Verlag LNCS 4336, pp 10–19

Hosbond JH, Nielsen PA (2005) Mobile Systems Development—A literature review, Proceedings of IFIP 8.2 Annual Conference

Höst M, Wohlin C, Thelin T (2005) Experimental context classification: incentives and experience of subjects, ICSE'05, Proceedings of the 27th international conference on Software engineering, ACM

Jefferies C, Brereton P, Turner M (2008) A Systematic Literature review to investigate Reengineering Existing Systems for Multi-Chanel Access, Conference on Software Maintenance and Reengineering (CSMR). April, Athens, pp 258–262

Jørgensen M (2004) A review of studies on expert estimation of software development effort. J Syst Softw 70(1–2):37–60

Jørgensen M (2005) Evidence-Based Guidelines for Assessment of Software Development Cost Uncertainty. IEEE Trans Softw Eng 2005:942–954

Jørgensen M (2007) Estimation of Software Development Work Effort: Evidence on Expert Judgement and Formal Models. Int J Forecasting 3(3):449–462

Jørgensen M, Shepperd M (2007) A Systematic Review of Software Development Cost Estimation Studies. IEEE Trans SE 33(1):33–53

Jørgensen M, Dybå T, Kitchenham BA (2005) Teaching Evidence-Based Software Engineering to University Students, 11th IEEE International Software Metrics Symposium (METRICS'05), p. 24

Juristo N, Moreno AM, Vegas S, Solari M (2006) In Search of What We Experimentally Know about Unit Testing. IEEE Softw 23(6):72–80

Kagdi H, Collard ML, Maletic JI (2006) A survey and taxonomy of approaches for mining software repositories in the context of software evolution. J Softw Maintenance Evol Res Pract 19(2):77–131

Kampenes VB, Dybå T, Hannay JE, Sjøberg DIK (2007) A Systematic Review of Effect Size in Software Engineering Experiments. Inf Softw Technol 49(11–12):1073–1086

Khan Khalid S, Kunz Regina, Kleijnen Jos, Antes Gerd (2003) Systematic Reviews to Support Evidence-based Medicine, The Royal Society of Medicine Press Ltd

Kitchenham BA (2004) Procedures for Undertaking Systematic Reviews, Joint Technical Report, Computer Science Department, Keele University (TR/SE-0401) and National ICT Australia Ltd (0400011 T.1)

Kitchenham BA, Charters S (2007) Guidelines for performing Systematic Literature Reviews in Software Engineering Technical Report EBSE/EPIC-2007-01, 2007

Kitchenham B, Dybå T, Jørgensen M (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

Kitchenham B, Mendes E, Travassos GH (2007) A Systematic Review of Cross- vs. Within-Company Cost Estimation Studies. IEEE Trans SE 33(5):316–329

Kitchenham BA, Brereton OP, Turner M (2008a) EPIC Case Study 2—Extension of a Tertiary Study. EPIC Technical Report, EPIC-2009-007

Kitchenham BA, Brereton OP, Budgen D (2008b) Protocol for extending a Tertiary Study of Systematic Literature Reviews in Software Engineering. EPIC Technical Report, EBSE-2008-006, June

Kitchenham BA, Brereton OP, Budgen D, Turner M, Bailey J, Linkman SG (2009a) Systematic Literature reviews in Software Engineering—A Systematic Literature review. Inf Softw Technol 51:7–15

Kitchenham BA, Brereton P, Turner M, Niazi M, Linkman S, Pretorius R, Budgen D (2009b) The Impact of Limited Search Procedures for Systematic Literature Reviews—A Participant-Observer Case Study, Proceedings of the Third Symposium on Empirical Software Engineering and Measurement, ESEM'09, pp 336–345

Kitchenham B, Brereton P, Budgen D, Li Z (2009c). An Evaluation of Quality Checklist Proposals—A participant-observer case study, in Proceedings of EASE 2009, BCS-eWiC

Kitchenham B, Pretorius R, Budgen D, Brereton OP, Turner M, Niazi M, Linkman S (2010a) Systematic Literature Reviews—A Tertiary Study, Information and Software Technology, accepted for publication.

Kitchenham BA, Budgen D, Brereton P (2010b) The value of mapping studies—An observer-participant case study. EASE 2010

Liebchen GA, Shepperd M (2008) Data sets and Data Quality in Software Engineering. PROMISE '08: Proceedings of the 4th international workshop on Predictor models in software engineering, May 2008, pp 39–44.

MacDonell S, Shepperd M (2007) Comparing local and global effort estimation models reflections on a systematic reviews. Proceedings of Empirical Software Engineering and Measurement, IEEE Computer Society Press

Mair M, Shepperd M, Jørgensen M (2005) An analysis of data sets used to train and validate cost prediction systems, PROMISE'05 Workshop

Martin BR (1996) The use of multiple indicators in the assessment of basic research. Scientometrics 36 (3):343–362

Martin BR, Irvine J (1983) Assessing Basic Research. Some partial indicators of scientific progress in radio astronomy. Res Policy 12:61–90

Mohagheghi P, Conradi R (2007) Quality, productivity and economic benefits of software reuse: a review of industrial studies. Empirical Softw Eng 12:471–516

Mohagheghi P, Dehlen V (2008) Where Is the Proof?—A Review of Experiences from Applying MDE in Industry. ECMDA-FA 2008, LNCS 5095, pp. 432–443

Neto AD, Subramanyan R, Viera M, Travassos GH Shull F (2008) Improving Evidence about Software Technologies. A Look at model-based testing. IEEE Softw 25(6):242–249

Petticrew Mark, Helen Roberts (2005) Systematic Reviews in the Social Sciences: A Practical Guide, Blackwell Publishing

Pino FJ, García F, Piattini M (2008) Software process improvement in small and medium enterprises: a review. Softw Qual J 16:237–261

Ramesh V, Glass RL, Vessey I (2004) Research in computer science: an empirical study. J Syst Softw 70(1–2):165–176

Renger M, Kolfschoten GL, de Vreede G-J (2008) Challenges in Collaborative Modeling: A Literature Review, CIAO! 2008 and EOMAS 2008, LNBIP 10, 2008, pp 61–77

Segal J, Grinyer A, Sharp H (2005) The type of evidence produced by empirical software engineers. REBSE’05

Shaw M, Clements P (2006) The Golden Age of Software Architecture: A Comprehensive Survey. Technical Report CMU-ISRI-06-101, Software Engineering Institute, Carnegie Mellon University

Shepperd M (2007) Software project economics: a roadmap, FOSE’07.

Sjøberg DIK, Hannay JE, Hansen O, Kampenes VB, Karahasanovic A, Liborg NK, Rekdal AC (2005) A survey of controlled experiments in software engineering. IEEE Trans SE 31(9):733–753

Staples M, Niazi M (2008) Systematic review of organizational motivation for adopting CMM-based SPI. Inf Softw Technol 50(7–8):605–620

Tichy W, Lukowicz P, Prechelt L, Heinze E (1995) Experimental Evaluation in Computer Science: A Quantitative Study. J Syst Softw 28(9):9–18

Torchiano M, Morisio M (2004) Overlooked Aspects of COTS-Based Development. IEEE Software, pp 88–93

Turner M, Kitchenham B, Budgen D, Brereton P (2008) Lessons learnt Undertaking a Large-scale Systematic Literature Review, in Proceedings of EASE 2008, BCS-eWiC

Weller AC (2001) Editorial Peer Review. Its Strengths and Weaknesses, Assist Monograph Series, Nre Jersey, USA

Wicks MN, Dewar RG (2007) A new research agenda for tool integration, J. Syst Softw 80:1567–1585

Yalaho A (2006) A Conceptual Model of ICT-Supported Unified Process of International Outsourcing of Software Production, 10th International Enterprise Distributed Object Computing Conference Workshops (EDOCW’06)

Yin Robert K (2003) Case Study Research: Design and Methods, 3 rd Edition, Sage Publications

Zhang H, Kitchenham B, Pfahl D (2008) Reflections on 10 years of Software Process Simulation Modeling: A systematic Review. International Workshop on Software Process Simulation Modeling. LNCS 5007:345–356

Barbara Kitchenham is Professor of Quantitative Software Engineering at Keele University in the UK. She has worked in software engineering for over 30 years both in industry and academia. Her main research interest is software measurement and its application to project management, quality control, risk management and evaluation of software technologies. Her most recent research has focused on the application of evidence-based practice to software engineering. She is a Chartered Mathematician and Fellow of the Institute of Mathematics and Its Applications, a Fellow of the Royal Statistical Society and a member of the IEEE Computer Society.

Pearl Brereton is Professor of Software Engineering in the School of Computing and Mathematics at Keele University. She was awarded a BSc degree in Applied Mathematics and Computer Science from Sheffield University (1970) and a PhD in Computer Science from Keele University (1977). After a period in industry and working for the UK's Science and Engineering Research Council she moved to a research post at Keele University in 1980. She was awarded a personal chair in 2003. Her research focuses on evidence-based software engineering and component-based/service-oriented systems. She is a member of the EPSRC computing college, IEEE Computer Society, the ACM, and the British Computer Society.

Mark Turner is a Lecturer in the School of Computing and Mathematics at Keele University, UK. His research interests include evidence-based software engineering, service-based software engineering and dynamic access control. Turner received a PhD in computer science from Keele University. He is a member of the IEEE Computer Society and the British Computer Society.

Mahmood Niazi is a Lecturer in the School of Computing and Mathematics at Keele University. He has spent more than a decade with leading technology firms and universities as a process analyst, senior systems analyst, project manager, and lecturer. His research interests include global software engineering, software process improvement and requirements engineering. He holds a Ph.D. from the Faculty of IT, University of Technology Sydney Australia.

Stephen Linkman is a Senior Lecturer in the School of Computing and Mathematics at Keele University and holds an MSc from the University of Leicester. His main research interests lie in the fields of software metrics and their application to project management, quality control, risk management and the evaluation of software systems and process. He is a visiting Professor at the University of Sao Paulo in Brazil.

Rialette Pretorius graduated from the University of Durham with an M.Sci(Hons) degree in Natural Sciences (Physics with e-Science). She is currently studying for a PGCE in Secondary Science.

David Budgen is a Professor of Software Engineering at Durham University in the UK. His research interests include software design, design environments, healthcare computing and evidence-based software engineering. He was awarded a BSc(Hons) in Physics and a PhD in Theoretical Physics from Durham University, following which he worked as a research scientist for the Admiralty and then held academic positions at Stirling University and Keele University before moving to his present post at Durham University in 2005. He is a member of the IEEE Computer Society, the ACM and the Institution of Engineering & Technology (IET).