Information and Software Technology 51 (2009) 7–15

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

Systematic literature reviews in software engineering – A systematic literature review

Barbara Kitchenham [a,*], O. Pearl Brereton [a], David Budgen [b], Mark Turner [a], John Bailey [b], Stephen Linkman [a]

[a] *Software Engineering Group, School of Computer Science and Mathematics, Keele University, Keele Village, Keele, Staffs, ST5 5BG, UK* [b] *Department of Computer Science, Durham University, Durham, UK*

ARTICLE INFO

ABSTRACT

Available online 12 November 2008

Background: In 2004 the concept of evidence-based software engineering

(EBSE) was introduced at the ICSE04 conference.

Aims: This study assesses the impact of systematic literature reviews (SLRs) which are the recommended EBSE method for aggregating evidence.

Method: We used the standard systematic literature review method employing a manual search of 10 journals and 4 conference proceedings.

Results: Of 20 relevant studies, eight addressed research trends rather than technique evaluation. Seven SLRs addressed cost estimation. The quality of SLRs was fair with only three scoring less than 2 out of 4.

Conclusions: Currently, the topic areas covered by SLRs are limited. European researchers, particularly those at the Simula Laboratory appear to be the leading exponents of systematic literature reviews. The series of cost estimation SLRs demonstrate the potential value of EBSE for synthesising evidence and making it available to practitioners.

Keywords: Systematic literature review
Evidence-based software engineering
Tertiary study
Systematic review quality
Cost estimation

Contents

* Corresponding author. Tel.: +44 1622 820484; fax: +44 1622 820176.
E-mail address: barbara@kitchenham.me.uk (B. Kitchenham).

0950-5849/$ - see front matter
doi:10.1016/j.infsof.2008.09.009

1. Introduction

At ICSE04, Kitchenham et al. [23] suggested software engineering researchers should adopt "Evidence-based Software Engineering" (EBSE). EBSE aims to apply an evidence-based approach to software engineering research and practice. The ICSE paper was followed-up by an article in IEEE Software [5] and a paper at Metrics05 [17].

Evidence-based research and practice was developed initially in medicine because research indicated that expert opinion based medical advice was not as reliable as advice based on the accumulation of results from scientific experiments. Since then many domains have adopted this approach, e.g. Criminology, Social policy, Economics, Nursing etc. Based on Evidence-based medicine, the goal of Evidence-based Software Engineering is:

> "To provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision making process regarding the development and maintenance of software" [5].

In this context, *evidence* is defined as a synthesis of best quality scientific studies on a specific topic or research question. The main method of synthesis is a *systematic literature review* (SLR). In contrast to an expert review using ad hoc literature selection, an SLR is a methodologically rigorous review of research results. The aim of an SLR is not just to aggregate all existing evidence on a research question; it is also intended to support the development of evidence-based guidelines for practitioners. The end point of EBSE

is for practitioners to use the guidelines to provide appropriate software engineering solutions in a specific context.

The purpose of this study is to review the current status of EBSE since 2004 using a tertiary study to review articles related to EBSE and, in particular, we concentrate on articles describing systematic literature reviews (SLRs). Although SLRs are not synonymous with EBSE, the aggregation of research results is an important part of the EBSE process and, furthermore, is the part of the EBSE process that can be readily observed in the scientific literature. We describe our methodology in Section 2 and present our results in Section 3. In Section 4 we answer our 4 major research questions. We present our conclusions in Section 5.

2. Method

This study has been undertaken as a systematic literature review based on the original guidelines as proposed by Kitchenham [22]. In this case the goal of the review is to assess systematic literature reviews (which are referred to as secondary studies), so this study is categorised as a tertiary literature review. The steps in the systematic literature review method are documented below.

2.1. Research questions

The research questions addressed by this study are:

RQ1. How much SLR activity has there been since 2004?
RQ2. What research topics are being addressed?
RQ3. Who is leading SLR research?
RQ4. What are the limitations of current research?

With respect to RQ1, it may be a concern that we started our

search at the start of 2004. We recognise that the term "systematic literature review" was not in common usage in the time period during which literature reviews published in 2004 were conducted. However, there were examples both of rigours literature reviews and of meta-analysis studies prior to 2004 [37,41,42,10,33,29,30,13]. Furthermore, the concepts of evidence-based software engineering had been discussed by research groups in Europe for some time before 2004 as part of some (unsuccessful) European Commission Research proposals. Thus, although we would not expect papers published in 2004 to have been directly influenced by the EBSE papers [23,5] or the guidelines for systematic reviews [22], we thought it was important to have some idea of the extent of systematic approaches to literature reviews before the guidelines were made generally available.

To address RQ1, we identified the number of SLRs published per year, the journal/conferences that published them and whether or not they referenced the EBSE papers [23,5] or Guidelines paper [22].

With respect to RQ2, we considered the scope of the study (i.e. whether it looked at research trends, or whether it addressed a technology-centred research question) and the software engineering topic area. With respect to RQ3, we considered individual researchers, the organisation to which researchers were affiliated and the country in which the organisation is situated.

With respect to limitations of SLRs (RQ4) we considered a number of issues:

RQ4.1. Were the research topics limited?
RQ4.2. Is there evidence that the use of SLRs is limited due to lack of primary studies?
RQ4.3. Is the quality of SLRs appropriate, if not, is it improving?
RQ4.4. Are SLRs contributing to practice by defining practice guidelines?

2.2. Search process

The search process was a manual search of specific conference proceedings and journal papers since 2004. The selected journals and conferences are shown in Table 1. The journals were selected because they were known to include either empirical studies or literature surveys, and to have been used as sources for other systematic literature reviews related to software engineering (e.g. [10 and 36]).

Each journal and conference proceedings was reviewed by one of four different researchers (i.e. Kitchenham, Brereton, Budgen and Linkman) and the papers that addressed literature surveys of any type were identified as potentially relevant. Kitchenham coordinated the allocation of researchers to tasks based on the availability of each researcher and their ability to access the specific journals and conference proceedings. The researcher responsible for searching the specific journal or conference applied the detailed inclusion and exclusion criteria to the relevant papers (see Section 2.3). Another researcher checked any papers included and excluded at this stage.

Table 1
Selected journals and conference proceedings.

Source	Acronym
Information and Software Technology	IST
Journal of Systems and Software	JSS
IEEE Transactions on Software Engineering	TSE
IEEE Software	IEEE SW
Communications of the ACM	CACM
ACM Computer Surveys	ACM Sur
ACM Transactions on Software Engineering Methodologies	TOSEM
Software Practice and Experience	SPE

Empirical Software Engineering Journal	EMSE
IEE Proceedings Software (now IET Software)	IET SW
Proceedings International Conference on Software Engineering	ICSE
Proceedings International Symposium of Software Metrics	Metrics
Proceedings International Symposium on Empirical Software Engineering	ISESE

In addition, we contacted Professor Guilherme Travassos directly and Professor Magne Jørgensen indirectly by reviewing the references in his web page. We did this because Professor Travassos had reported to one of us that his research group was attempting to adopt the SLR process and because Professor Jørgensen was known to be the author of a substantial number of SLRs.

2.3. Inclusion and exclusion criteria

Peer-reviewed articles on the following topics, published between Jan 1st 2004 and June 30th 2007, were included:

- Systematic Literature Reviews (SLRs) i.e. literature surveys with defined research questions, search process, data extraction and data presentation, whether or not the researchers referred to their study as a *systematic* literature review.
- Meta-analyses (MA).

Note, we included articles where the literature review was only one element of the articles as well as articles for which the literature review was the main purpose of the article.

Articles on the following topics were excluded

- Informal literature surveys (no defined research questions; no defined search process; no defined data extraction process).
- Papers discussing the procedures used for EBSE or SLRs.
- Duplicate reports of the same study (when several reports of a study exist in different journals the most complete version of

the study was included in the review).

2.4. Quality assessment

Each SLR was evaluated using the York University, Centre for Reviews and Dissemination (CDR) Database of Abstracts of Reviews of Effects (DARE) criteria [3]. The criteria are based on four quality assessment (QA) questions:

QA1. Are the review's inclusion and exclusion criteria described and appropriate?
QA2. Is the literature search likely to have covered all relevant studies?
QA3. Did the reviewers assess the quality/validity of the included studies?
QA4. Were the basic data/studies adequately described?

The questions were scored as follows:

- QA1: Y (yes), the inclusion criteria are explicitly defined in the study, P (Partly), the inclusion criteria are implicit; N (no), the inclusion criteria are not defined and cannot be readily inferred.
- QA2: Y, the authors have either searched 4 or more digital libraries and included additional search strategies or identified and referenced all journals addressing the topic of interest; P, the authors have searched 3 or 4 digital libraries with no extra search strategies, or searched a defined but restricted set of journals and conference proceedings; N, the authors have search up to 2 digital libraries or an extremely restricted set of journals.
- QA3: 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 primary studies has been attempted.
- QA4: Y Information is presented about each study; P only summary information about primary studies is presented; N the results of the individual primary studies are not specified.

The scoring procedure was Y = 1, P = 0.5, N = 0, or Unknown (i.e. the information is not specified). Kitchenham coordinated the quality evaluation extraction process. Kitchenham assessed every paper, and allocated 4 papers to each of the other authors of this study to assess independently. When there was a disagreement, we discussed the issues until we reached agreement. When a question was scored as unknown we e-mailed the authors of the paper and asked them to provide the relevant information and the question re-scored appropriately.

2.5. Data collection

The data extracted from each study were:

- The source (journal or conference) and full reference.
- Classification of the study Type (SLR, Meta-Analysis MA); Scope (Research trends or specific technology evaluation question).
- Main topic area.
- The author(s) and their institution and the country where it is situated.
- Summary of the study including the main research questions and the answers.
- Research question/issue.
- Quality evaluation.
- Whether the study referenced the EBSE papers [23,5] or the SLR Guidelines [22].
- Whether the study proposed practitioner-based guidelines.
- How many primary studies were used in the SLR.

One researcher extracted the data and another checked the extraction. The procedure of having one extractor and one checker is not consistent with the medical standards summarized in Kitchenham's guidelines [22], but is a procedure we had found useful in practice [2]. Kitchenham coordinated the data extraction and checking tasks, which involved all of the authors of this paper. Allocation was not randomized, it was based on the time availability of the individual researchers. When there was a disagreement, we discussed the issues until we reached agreement.

2.6. Data analysis

The data was tabulated to show:

- The number of SLRs published per year and their source (addressing RQ1).
- Whether the SLR referenced the EBSE papers or the SLR guidelines (addressing RQ1).
- The number of studies in each major category i.e. research trends or technology questions (addressing RQ2 and RQ4.1).
- The topics studied by the SLRs and their scope (addressing RQ2 and RQ4.1).
- The affiliations of the authors and their institutions (addressing RQ3).
- The number of primary studies in each SLR (addressing RQ4.2).
- The quality score for each SLR (addressing RQ4.3).
- Whether the SLR proposed practitioner-oriented guidelines (addressing RQ4.4).

2.7. Deviations from protocol

As a result of an anonymous review of an earlier version of this paper, we made some changes to our original experimental protocol (see [24] Appendix 1):

- We explained our concentration on SLRs as part of EBSE.

- We extended the description of our research questions.
- We asked the authors of studies for which the answers to certain quality questions were unknown to provide the information.
- We clarified the link between the research questions and the data collection and analysis procedures

3. Results

This section summarizes the results of the study.

3.1. Search results

Table A1 (in Appendix 1) shows the results of the search procedure. Although we identified 19 articles by this search process, one of the articles [19] is a short version of another article [18]. Thus we identified 18 unique studies. In addition, we found another two other studies that had been subject to peer review: one by asking researchers about their current work [1] and the other by searching the Simula Research Laboratory website [14]. Other potentially relevant studies that were excluded as a result of applying the detailed inclusion and exclusion criteria are listed in Table A2 in Appendix 1. One of the excluded papers positioned itself as an EBSE paper but did not specify how it applied the EBSE principles [26].

Two studies were published in conference proceedings as well as in journals: Galin and Avrahami [7] is a conference version of

Table 2

Systematic review studies.

Galin and Avrahami [8] and Kitchenham et al. [20] is a conference version of Kitchenham et al. [21].

The data extracted from each study are shown in Tables A2 and A3 (in Appendix 1). Summaries of the studies can be found in [24], Appendix 3.

3.2. Quality evaluation of SLRs

We assessed the studies for quality using the DARE criteria (see Section 2.4). The score for each study is shown in Table 3. The fields marked with an asterisk in Table 3 were originally marked as unknown and were re-assigned after communicating with the study authors.

The last column in Table 5 shows the number of questions where the researchers were in agreement. All disagreements were discussed and resolved.

The results of the quality analysis show that all studies scored 1 or more on the DARE scale and only three studies scored less than 2. Two studies scored 4 ([15 and 21]) and two studies scored 3.5 ([14 and 40]).

3.3. Quality factors

We investigated the relationship between the quality score for an SLR and both the date when the article was published, and the use or not of the guidelines for SLRs [22]. The average quality scores for studies each year is shown in Table 4. Note, for this anal-

ID	Author	Date	Topic type	Topic area	Article	Refs.	Include	Num.

	type	practitioner	primary guidelines studies
S1	Barcelos and Travassos [1]	2006	Technology evaluation
S2	Dyba et al. [4]	2006	Research trends
S3	Galin and Avrahami [7,8]	2005 & 2006	Technology evaluation
S4	Glass et al. [9]	2004	Research trends
S5	Grimstad et al. [11]	2006	Technology evaluation
S6	Hannay et al. [12]	2007	Research trends
S7	Jørgensen [15]	2004	Technology evaluation
S8	Jørgensen [14]	2007	Technology evaluation
S9	Jørgensen and Shepperd [16]	2007	Research trends
S10	Juristo et al. [18,19]	2004 & 2006	Technology evaluation
S11	Kitchenham et al. [20,21]	2006 & 2007	Technology evaluation
S12	Mair and Shepperd [27]	2005	Technology evaluation
S13	Mendes [28]	2005	Research trends
S14	Moløkken-Østvold et al. [31]	2005	Technology evaluation
S15	Petersson et al. [32]	2004	Technology evaluation

S16	Ramesh et al. [34]	2004	Research trends	Software architecture evaluation methods	SLR	Guideline TR	No	54
S17	Runeson et al.[35]	2006	Technology evaluation	Power in SE experiments	SLR	Guideline TR	No	103
S18	Torchiano and Morisio [38]	2004	Technology evaluation	CMM	MA	No	No	19
S19	Sjøberg et al. [36]	2005	Research trends	Comparative trends in CS, IS and SE	SLR	No	No	1485
S20	Zannier et al. [40]	2006	Research trends	Cost estimation	SLR	Guideline TR	Yes	32
				Theory in SE experiments	SLR	Guideline TR	No	103
				Cost estimation	SLR	No	Yes	15
				Cost estimation	SLR	No	Yes	16
				Cost estimation	SLR	GuidelineTR	No	304
				Unit testing	SLR	EBSE paper	No	24
				Cost estimation	SLR	Guideline TR	Yes	10
				Cost estimation	SLR	No	No	20
				Web research	SLR	Guideline TR	No	173
				Cost estimation	SLR	No	No	6
				Capture–recapture in inspections	SLR	No	No	29
				Computer science research	SLR	No	No	628

Testing methods	SLR	EBSE paper	No [a]	12
COTS development	SLR	No	No	21
SE experiments	SLR	Guideline	No	103
Empirical studies in ICSE	SLR	TR No	No	63

[a] Runeson et al. suggest how practitioners can use their results but do not explicitly define guidelines.

Table 3
Quality evaluation of SLRs.

Study	Article type	QA1	QA2	QA3	QA4	Total score	Initial rater agreement
S1	SLR	Y	P	N	Y	2.5	4
S2	SLR	Y	P	P	P	2.5	4
S3	MA	Y	P*	P	P	2.5	4
S4	SLR	Y	P	N	P	2	4
S5	SLR	Y	Y	N	Y	3	4
S6	SLR	Y	P	N	Y	2.5	4
S7	SLR	Y	Y*	Y	Y	4	4
S8	SLR	Y	Y	P	Y	3.5	4
S9	SLR	Y	Y	N	Y	3	4
S10	SLR	P	N	P	P	1.5	4
S11	SLR	Y	Y	Y	Y	4	4
S12	SLR	Y	P*	N	Y	2.5	4
S13	SLR	Y	N	P	P	2	4
S14	SLR	Y	Y*	N	Y	3	4
S15	SLR	P	Y	N	Y	2.5	3
S16	SLR	P	P	N	P	1.5	3
S17	SLR	Y	N	N	Y	2	2
S18	SLR	Y	N	N	N	1	4
S19	SLR	Y	P	N	P	2	3
S20	SLR	Y	Y	Y	P	3.5	3

Table 4
Average quality scores for studies by publication date.

	Year			
	2004	2005	2006	2007
Number of studies	6	5	6	3

Mean quality score	2.08	2.4	2.92	3
Standard deviation of quality score	1.068	0.418	0.736	0.50

ysis we used the first publication date for any duplicated study. Table 4 indicates that the number of studies published per year has been quite stable. The average quality score appears to be increasing, the Spearman correlation between year and score was 0.51 ($p < 0.023$)

The average quality scores for studies that did or did not reference the SLR guidelines are shown in Table 5. A one way analysis of variance showed that the mean quality score of studies that referenced the SLR guidelines [22] compared with those that did not, was not significant ($F = 0.37$, $p = 0.55$). Thus, it appears that the quality of SLRs is improving but the improvement cannot be attributed to the guidelines.

4. Discussion

In this section, we discuss the answers to our research questions.

4.1. How much EBSE Activity has there been since 2004?

Overall, we identified 20 relevant studies in the sources that we searched, as shown in Table 2. 19 studies were classified as SLRs and one study was classified as a meta-analysis [8]. Twelve studies addressed technology evaluation issues and 8 addressed research trends. We found that 8 studies referenced Kitchenham's guide-

Table 5
Average quality score for studies according to use of guidelines.

	Referenced SLR	Did not reference

	guidelines	SLR guidelines
Number of studies	8	12
Mean quality score	2.69	2.46

lines [22] and two referenced the EBSE paper [5]. Thus, half the studies directly positioned themselves as related to Evidence-based Software Engineering.

With respect to where SLRs are published, IEEE Software and IEEE TSE each published 4 studies, JSS published 3 and IST published 2. Thus, it appeared that IST's attempts to encourage the publication of SLRs, was unsuccessful [6]. However, a further check of IST publications (on September 17th 2008 using the search string *systematic AND review*) found seven more SLRs, whereas similar searches of TSE and JSS found no new SLRs.

Initially, we were surprised that ACM Computer Surveys did not include any relevant software engineering studies, although the journal published a systematic literature review on the topic of education [25]. An automated search of ACM Computer Surveys using the ACM digital library on September 20th 2008, found no software-related surveys that used the systematic review methodology. However, the apparent lack of software SLRs in ACM Computer Surveys may be because, with a maximum of four issues per year, the journal is likely to have a significant publication lag.

4.2. What research topics are being addressed?

With respect to the topic of the articles, eight were related to research trends rather than specific research questions. In terms of the software engineering topic area addressed by the SLRs:

- 7 related to software cost estimation (one of those covered research trends), in addition, the four studies that included evi-

dence-based guidelines all related to cost estimation.
- 3 articles related to software engineering experiments (all investigated research trends).
- 3 articles related to test methods.

In the area of cost estimation, researchers are addressing specific research questions including:

- Are mathematical estimating models more accurate than expert opinion based estimates?
  - No. [15].

- What is the level of overrun of software projects and is it changing over time?
  - 30% and unchanging [31].

- Are regression-based estimation models more accurate than analogy-based models?
  - No. [27].

- Should you use a benchmarking data base to construct an estimating model for a particular company if you have no data of your own?
  - Not if you work for a small company doing niche applications [21].

- Do researchers use cost estimation terms consistently and appropriately?
  - No they confuse prices, estimates, and budgets [11].

- When should you use expert opinion estimates?
  - When you don't have a calibrated model, or important contextual information is not incorporated into your model [14].

The testing studies have investigated:

- Whether testing is better than inspections.

- Yes for design documents, No for code.[35].

- Different capture–recapture methods used to predict the defects remaining after inspections.
  - Most studies recommend the Mh-JK model. Only one of 29 studies was an application study [32].

- Empirical studies in unit testing.
  - Empirical studies in unit testing are mapped to a framework and summarized [18].

4.3. Who is leading EBSE research?

Overall, the set of studies are dominated by European researchers who have been involved in 14 of the studies, in particular the Simula Research Laboratory in Norway which has been involved in 8 of the studies. The two researchers who contributed to more than two SLRs, Jørgensen (5) and Sjøberg (3), are both affiliated to the Simula Research Laboratory. Only four studies had North American authors.

The success of the Simula Research Laboratory in applying the principles of EBSE and performing high quality SLRs is supported by the strategy of constructing databases of primary studies related to specific topic areas and using those databases to address specific research questions. A database of cost estimation papers from over 70 journals [16] has been the basis of many of the detailed cost estimation studies authored or co-authored by Jørgensen and the database of 103 software experiments [36] has

allowed researchers to assess a number of specific research trends in software experimentation.

4.4. What are the limitations of current research?

With respect to whether research topics addressed by SLRs are somewhat limited (RQ4.1), a relatively large number of studies relate to research practice rather than questions concerning specific software engineering practices and techniques. This is disappointing since this type of study benefits researchers rather than practitioners, and evidence-based software engineering is meant to be of benefit to practitioners. However, three of the research trend studies addressed the quality of current experimental studies and identified areas for improvement, and improved empirical methods might be expected to benefit practitioners in the longer term. Furthermore, the Jørgensen and Shepperd study [16], although classified as a research trends study, is also an example of a mapping study (i.e. a study that aims to identify and categorise the research in a fairly broad topic area). The availability of high quality mapping studies has the potential to radically change the nature of software engineering research. Mapping studies can highlight areas where there is a large amount of research that would benefit from more detailed SLRs and areas where there is little research that require more theoretical and empirical research. Thus, instead of every researcher undertaking their own research from scratch, a broad mapping study provides a common starting point for many researchers and many research initiatives. On September 17, 2008, the SCOPUS search engine found already 23 citations of this paper of which only four were self-citations. This suggests that the research community has already recognised the value of a good mapping study.

For studies that investigated technology questions, the majority have been in the cost estimation field. Of the conventional software

engineering lifecycle, only testing, with three studies, has been addressed.

Juristo et al. [18,19] found only 24 studies comparing unit testing techniques. This is extremely surprising given that unit testing is a software activity that is relatively easily studied using experiments since tasks are relatively small and can be treated in isolation. We found this particularly curious in the light of 29 experiments that compared test–retest methods of predicting remaining defects after inspections [32] which is a far less central element of software engineering practice than unit testing. Juristo et al.'s study was based on a search of only the ACM and IEEE electronic databases, so this may be an example of area where a broader search strategy would be useful.

Looking at the number of primary studies in each SLR (RQ4.2), unsurprisingly, the research trends studies were based on a larger number of primary studies (i.e. 63–1485) than the technology evaluation studies (i.e. 6–54). However, the results confirm that some topics have attracted sufficient primary studies to permit SLRs to address detailed research questions, although, as yet, only a limited number of topics are addressed.

With respect to the quality of SLRs (RQ4.3), the results of the quality analysis show that all studies scored 1 or more on the DARE scale and only three studies scored less than 2. However, relatively few SLRs have assessed the quality of the primary studies included in the review. This is acceptable in the context of studies of research trends but is more problematic for reviews that attempt to evaluate technologies.

With respect to the contribution of SLRs to software engineering practice (RQ4.4), of the 12 SLRs that addressed research ques-

Table A1
Sources searched for years 2004–2007 (including articles up to June 30 2007).

Year	2004	2005	2006	2007	Total

IST (Total)	85	95	72	47	299
IST (Relevant)	0	2	2	0	4
IST (Selected)	0	0	2	0	2
JSS (Total)	139	122	124	43	428
JSS (Relevant)	4	0	0	0	4
JSS (Selected)	3	0	0	0	3
IEEE SW (Total)	51	52	48	24	175
IEEE SW (Relevant)	1	0	5	2	9
IEEE SW (Selected)	1	0	3	0	4
TSE (Total)	69	66	56	25	216
TSE (Relevant)	2	1	0	3	7
TSE (Selected)	0	1	0	3	4
CACM (Total)	148	141	158	64	511
CACM (Relevant)	1	0	0	0	1
CACM (Selected)	1	0	0	0	1
ACM Sur (Total)	12	11	13	3	39
ACM Sur (Relevant)	0	0	1	0	1
ACM Sur (Selected)	0	0	0	0	0
TOSEM (Total)	10	12	12	6	40
TOSEM (Relevant)	0	2	0	0	2
TOSEM (Selected)	0	0	0	0	0
SPE (Total)	64	59	68	29	220
SPE (Relevant)	0	0	0	0	0
SPE (Selected)	0	0	0	0	0
ICSE (Total)	58	58	36	64	216
ICSE (Relevant)	0	0	1	0	1
ICSE (Selected)	0	0	1	0	1
ISESE (Total)	26	50	56	n/a	132
ISESE (Relevant)	0	2	1	n/a	3
ISESE (Selected)	0	2	0	n/a	2
IET SW (Total)	22	28	22	9	81
IET SW (Relevant)	0	0	0	1	1
IET SW (Selected)	0	0	0	0	0
EMSE (Total)	14	19	20	12	61
EMSE (Relevant)	1	0	0	0	1
EMSE (Selected)	1	0	0	0	1
Metrics (Total)	36	48	n/a	n/a	
Metrics (Relevant)	1	0	n/a	n/a	1
Metrics (Selected)	1	0	n/a	n/a	1
Total	734	761	685	326	2506

Total relevant	10	7	10	6	33
Total selected	7	3	6	3	19

tions only four offered advice to practitioners. This is an issue where there needs to be improvement, since Evidence-based Software Engineering is meant to impact practice not just academia.

4.5. Limitations of this study

The procedures used in this study have deviated from the advice presented in Kitchenham's 2004 guidelines [22] in several ways:

- The search was organised as a manual search process of a specific set of journals and conference proceedings not an automated search process. This was consistent with the practices of other researchers looking at research trends as opposed to software technology evaluation.
- A single researcher selected the candidate studies, although the studies included and excluded were checked by another researcher.
- A single researcher extracted the data and another researcher checked the data extraction, as suggested by Brereton et al. [2].

The first point above implies that we may have missed some relevant studies, and thus underestimate the extent of EBSE-related research. In particular, we will have missed articles published in national journals and conferences. We will also have missed articles in conferences aimed at specific software engineering topics which are more likely to have addressed research questions rather than research trends. Thus, our results must be qualified

as applying only to systematic literature reviews published in the major international software engineering journals, and the major general and empirical software engineering conferences.

With respect to the second point, given our interest in systematic literature reviews, we are likely to have erred on the side of

Table A2
Candidate articles not selected.

including studies that were not very systematic, rather than omitting any relevant studies. For example, the literature review in the primary study, that was assigned the lowest quality score [38], was only a minor part of the article.

The third point means that some of the data we collected may be erroneous. A detailed review of one of our own systematic literature reviews has suggested that the extractor/checker mode of working can lead to data extraction and aggregation problems when there are a large number of primary studies or the data is complex [39]. However, in this tertiary study, there were relatively few primary studies and the data extracted from the selected articles were relatively objective, so we do not expect many data extraction errors. The quality assessment criteria proved the most difficult data to extract because the DARE criteria are somewhat subjective. However quality criteria were evaluated independently by two researchers, hopefully reducing the likelihood of erroneous results.

5. Conclusions

Although 10 of the SLR studies in this review cited one of the EBSE papers [5] or the SLR Guidelines [22], the number of SLRs has remained extremely stable in the 3.5 years included in this

study. Furthermore, Table A2 (see Appendix 1) also makes it clear that many researchers still prefer to undertake informal literature surveys. However, we have found that the quality of SLRs is improving, suggesting that researchers who are interested in the EBSE approach are becoming more competent in the SLR methodology.

The spread of topics covered by current SLRs is fairly limited. Furthermore main stream software engineering topics are not well represented. However, even if these areas are unsuitable for SLRs aimed at empirical assessments of software technology, we believe

Source	Authors	Reference
Year Title		Reason for rejection
TSE	T. Mens and T. Tourwé	30(2), pp 126–139
TSE	S. Balsamo, A. Di Marco, P. Inverardi	30(5), pp. 295–309
IET Software	S. Mahmood, R. Lai and Y.S. Kim	1(2), pp 57–66
IEEE Software	D.C. Gumm	23(5) pp. 45–51
IEEE Software	M. Shaw and P Clements	23(2) pp. 31–39
IEEE Software	M. Aberdour	24(1), pp. 58–64
IEEE Software	D. Damian	24(2), pp. 21–27
JSS	E. Folmer and J. Bosch	70, pp. 61–78
IST	Hochstein and Lindvall	47, pp. 643–656
IST	S. Mahmood, R. Lai, Y.S. Kim, J.H. Kim, S.C. Park, H.S. h	47, pp. 693–707
TOSEM	J. Estublier, D. Leblang, A. van der Hoek, R. Conradi, G. Clemm, W. Tichy, D. Wiborg-Weber	pp. 383–430
TOSEM	Barbara G. Ryder, Mary Lou Soffa, Margaret Burnett	pp. 431–477
ACM Surv	J. Ma and J. V. Nickerson	38(3), pp. 1–24
ISESE	S. Wagner	
2004	A survey of software refactoring	Informal literature survey
2004	Model-based performance prediction in software development	Informal literature survey

2007	Survey of component-based software development	Informal literature survey
2006	Distribution dimensions in software development	Literature survey referenced but not described in article
2006	The golden age of software Architecture	Informal literature survey
2007	Achieving quality in open source software	Informal literature survey
2007	Stakeholders in global requirements engineering: lessons learnt from practice	Informal literature survey
2004	Architecting for usability: a survey	Informal literature survey
2005	Combating architectural degeneration: a survey	Informal literature survey
2005	A survey of component-based system quality assurance and assessment	Informal literature survey
2005	Impact of software engineering research on the practice of software configuration management	Informal literature survey
2005	The impact of software engineering research on modern programming languages	Informal literature survey. No clear search criteria, no data extraction process.
2006	Hands-on, simulated and remote laboratories: a comparative literature review	Not a software engineering topic
2006	A literature survey of the quality economics of defect-detection techniques	Informal literature survey although quantitative data tabulated for different testing techniques.

it would be possible, and extremely valuable, for leading software In the area of cost estimation there have been a series of engineering researchers to undertake mapping studies of their do- systematic

literature reviews. This accumulation of evidence in a main similar to that provided by Jørgensen and Shepperd study specific topic area is starting to demonstrate the value of evi[16] for cost estimation research.
dence-based software engineering. For example, the evidence

Table A3
Author affiliation details.

ID Institution Country of institution	Authors
S1	Barcelos Travassos
S2	Dybå Kampenes Sjøberg
S3	Gavin Avrahami
S4	Glass Ramesh Vessey
S5	Grimstad

	Jørgensen
	Moløkken-Østvold
S6	Hannay
	Sjøberg
	Dybå
S7	Jørgensen
S8	Jørgensen
S9	Jørgensen
	Shepperd
S10	Juristo
	Moreno
	Vegas
S11	Kitchenham
	Mendes
	Travassos
S12	Mair
	Shepperd
S13	Mendes
S14	Moløkken-Østvold
	Jørgensen

	Tanilkan
	Gallis
	Lien
	Hove
S15	Petersson
	Thelin
	Runeson
	Wohlin
S16	Ramesh
	Glass
	Vessey
S17	Runeson
	Andersson
	Thelin
	Andrews
	Berling
S18	Sjøberg
	Hannay
	Hansen
	Kampenes
	Karahasanović

		Liborg	
		Rakdal	
S19		Torchiano	
		Morisio	
S20		Zannier	
		Melnik	
		Maurer	
Federal University of Rio de Janeiro			Brazil
Federal University of Rio de Janeiro			Brazil
SINTEF & Simula Laboratory			Norway
Simula Laboratory			Norway
Simula Laboratory			Norway
Ruppin Academic Center			Israel
Lipman Electronic Engineering			Israel
Computing Trends			USA
Kelley Business School, Indiana University			USA
Kelley Business School, Indiana University			USA
Simula Research Laboratory			Norway
Simula Research Laboratory			Norway
Simula Research Laboratory			Norway
Simula Research Laboratory			Norway
Simula Research Laboratory			Norway
SINTEF & Simula Research Laboratory			Norway
Simula Research Laboratory			Norway
Simula Research Laboratory			Norway
Simula Research Laboratory			Norway
Brunel University			UK
Univsidad Politéncia de Madrid			Spain
Univsidad Politéncia de Madrid			Spain
Univsidad Politéncia de Madrid			Spain
Keele University & NICTA			UK & Australia
University of Auckland			New Zealand

Federal University of Rio de Janeiro	Brazil
Brunel University	UK
Brunel University	UK
University of Auckland	New Zealand
Simula Research Laboratory & OSLO University	Norway
Simula Research Laboratory	Norway
OSLO University	Norway
Simula Research Laboratory & OSLO University	Norway
Simula Research Laboratory	Norway
Simula Research Laboratory	Norway
Lund University	Sweden
Lund University	Sweden
Lund University	Sweden
Bleking Institute of Technology	Sweden
Kelley School of Business, Indiana University	USA
Computing Trends	USA
Kelley School of Business, Indiana University	USA
Lund University	Sweden
Lund University	Sweden
Lund University	Sweden
University of Denver	USA
Lund University	Sweden
Simula Research Laboratory	Norway
Simula Research Laboratory	Norway
Simula Research Laboratory	Norway
Simula Research Laboratory	Norway
Simula Research Laboratory	Norway
BNP Paribas	Norway
Unified Consulting	Norway
Norwegian University of Science and technology	Norway
Politecnico de Torino	Italy
University of Calgary	Canada
University of Calgary	Canada
University of Calgary	Canada

gathered by means of the SLRs has overturned existing "common knowledge" about the efficacy of models compared with expert opinion and the size of project overruns. Furthermore in this area we are beginning to see the publication of evidence-based guidelines aimed at practitioners, which is a specific goal of evidence-based software engineering.

This review suggests that the Simula Research Laboratory, Norway is currently the leading software engineering institution in terms of undertaking SLRs. The research group has benefited from developing extremely effective research procedures to support their secondary studies. We recommend other research groups adopt similar research procedures, allowing the results of their own literature reviews to build up into a data base of categorised research papers that is available to initiate research programmes and provide the references needed for research articles.

The results in this study suggest that the current output of EBSE articles is strongly supported by European researchers. However, if EBSE is to have a serious impact on software engineering research and practice, it is important that researchers in other areas of the world take an increased interest in a formal approach to literature reviews, particularly, the US, because of its leadership in software engineering research.

This study suffers from a number of limitations; in particular, we have restricted ourselves to a manual search of international journals and conferences. We plan to extend this study by undertaking a broader automated search for other SLRs over the same time period. This has the joint aim of extending the generality of this study and investigating a number of issues associated with systematic literature reviews in software engineering i.e. whether

we should use manual or automated searchers, and whether restricted searches provide reliable results. We also plan to repeat this study at the end of 2009 to track the progress of SLRs and evidence-based software engineering.

Acknowledgements

This research was funded by The Engineering and Physical Sciences Research Council (EPSRC) EBSE Project (EP/C51839X/1). Short, preliminary versions of this study were presented at the REBSE2 workshop at ICSE07 and the EASE07 Conference at Keele University.

Appendix 1. Tables of the systematic review results.

See Tables A1–A3.

References

[1] R.F. Barcelos, G.H. Travassos, Evaluation approaches for software architectural documents: a systematic review, in: Ibero-American Workshop on Requirements Engineering and Software Environments (IDEAS), La Plata, Argentina, 2006.
[2] O.P. Brereton, B.A. Kitchenham, D. Turner Budgen, M. Khalil, Lessons from applying the systematic literature review process within the software engineering domain, Journal of Systems and Software 80 (4) (2007) 571–583.
[3] Centre for Reviews and Dissemination, What are the criteria for the inclusion of reviews on DARE? 2007. Available at <http://www.york.ac.uk/inst/crd/faq4.htm, 2007 (accessed 24.07.07>).
[4] T. Dyba, V.B. Kampenes, D.I.K. Sjøberg, A systematic review of statistical power in software engineering experiments, Information and Software Technology 48 (8) (2006) 745–755.
[5] T. Dybå, B.A. Kitchenham, M. Jørgensen, Evidence-based software engineering for practitioners, IEEE Software 22 (1) (2005) 58–65.
[6] M. Dyer, M. Shepperd, C. Wohlin, Systematic Reviews in Evidence-Based Software Technology and Software Engineering 47 (1) (2005) 1.

[7] D. Galin, M. Avrahami, Do SQA programs work – CMM works. A meta analysis, IEEE International Conference on Software – Science, Technology and Engineering (2005).
[8] D. Galin, M. Avrahami, Are CMM program investments beneficial? Analyzing past studies, IEEE Software 23 (6) (2006) 81–87.
[9] R.L. Glass, V. Ramesh, I. Vessey, An analysis of research in computing disciplines, CACM 47 (6) (2004) 89–94.
[10] R.L. Glass, I. Vessey, V. Ramesh, Research in software engineering: an analysis of the literature, Information and Software technology 44 (8) (2002) 491–506.
[11] S. Grimstad, M. Jorgensen, K. Molokken-Ostvold, Software effort estimation terminology: the tower of Babel, Information and Software Technology 48 (4) (2006) 302–310.
[12] J.E. Hannay, D.I.K. Sjøberg, T. Dybå, A systematic review of theory use in software engineering experiments, IEEE Transactions on SE 33 (2) (2007) 87–107.
[13] W. Hayes, Research synthesis in software engineering: the case for meta-analysis, Proceedings 6th International Software Metrics Symposium, IEEE Computer Press, 1999. pp. 143–151.
[14] M. Jørgensen, Estimation of software development work effort: evidence on expert judgement and formal models, International Journal of Forecasting 3 (3) (2007) 449–462.
[15] M. Jørgensen, A review of studies on expert estimation of software development effort, Journal of Systems and Software 70 (1–2) (2004) 37–60.
[16] M. Jørgensen, M. Shepperd, A systematic review of software development cost estimation studies, IEEE Transactions on SE 33 (1) (2007) 33–53.
[17] M. Jørgensen, T. Dybå, B.A. Kitchenham, Teaching evidence-based software engineering to university students, in: 11th IEEE International Software Metrics Symposium (METRICS'05), 2005, p. 24.
[18] N. Juristo, A.M. Moreno, S. Vegas, Reviewing 25 years of testing technique experiments, Empirical Software Engineering Journal (1–2) (2004) 7–44.
[19] N. Juristo, A.M. Moreno, S. Vegas, M. Solari, In search of what we experimentally know about unit testing, IEEE Software 23 (6) (2006) 72–80.
[20] B. Kitchenham, E. Mendes, G.H. Travassos, A systematic review of cross-company vs. within-company cost estimation studies, Proceedings of EASE06, BSC (2006) 89–98.
[21] B. Kitchenham, E. Mendes, G.H. Travassos, A systematic review of cross- vs. within-company cost estimation studies, IEEE Transactions on SE 33 (5) (2007) 316–329.
[22] B.A. Kitchenham, Procedures for Undertaking Systematic Reviews, Joint Technical Report, Computer Science Department, Keele University (TR/SE-0401) and National ICT Australia Ltd. (0400011T.1), 2004.
[23] B.A. Kitchenham, T. Dybå, M. Jørgensen, Evidence-based software engineering, in: Proceedings of the 26th International Conference on Software Engineering, (ICSE'04), IEEE Computer Society, Washington DC, USA, 2004, pp. 273–281.

[24] B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, A Systematic Literature Review of Evidence-based Software Engineering, EBSE Technical Report, EBSE-2007-03, 2007.
[25] J. Ma, J.V. Nickerson, Hands-on, simulated and remote laboratories: a comparative literature review, ACM Surveys 38 (3) (2006) 1–24.
[26] S. Mahmood, R. La, Y.S. Kim, A survey of component-based system quality assurance and assessment, IET Software 1 (2) (2005) 57–66.
[27] C. Mair, M. Shepperd, The consistency of empirical comparisons of regression and analogy-based software project cost prediction, International Symposium on Empirical Software Engineering (2005) 509–518.
[28] E. Mendes, A systematic review of Web engineering research, International Symposium on Empirical Software Engineering (2005) 498–507.
[29] J. Miller, Can results from software engineering experiments be safely combined?, in: Proceedings 6th International Software Metrics Symposium, IEEE Computer Press, 1999, pp 152–158.
[30] J. Miller, Applying meta-analytical procedures to software engineering experiments, JSS 54 (1) (2000) 29–39.
[31] K.J. Moløkken-Østvold, M. Jørgensen, S.S. Tanilkan, H. Gallis, A.C. Lien, S.E. Hove, A survey on software estimation in the Norwegian industry, Proceedings Software Metrics Symposium (2004) 208–219.
[32] H. Petersson, T. Thelin, P. Runeson, C. Wohlin, Capture–recapture in software inspections after 10 years research – theory, evaluation and application, Journal of Systems and Software 72 (2004) 249–264.
[33] L.M. Pickard, B.A. Kitchenham, P. Jones, Combining empirical results in software engineering, Information and Software Technology 40 (14) (1998) 811–821.
[34] V. Ramesh, R.L. Glass, I. Vessey, Research in computer science: an empirical study, Journal of Systems and Software 70 (1–2) (2004) 165–176.
[35] P. Runeson, C. Andersson, T. Thelin, A. Andrews, T. Berling, What do we know about defect detection methods?, IEEE Software 23 (3) (2006) 82–86.
[36] D.I.K. Sjøberg, J.E. Hannay, O. Hansen, V.B. Kampenes, A. Karahasanovic, N.K. Liborg, A.C. Rekdal, A survey of controlled experiments in software engineering, IEEE Transactions on SE 31 (9) (2005) 733–753.
[37] W.F. Tichy, P. Lukowicz, L. Prechelt, E.A. Heinz, Experimental evaluation in computer science: a quantitative study, Journal of Systems and Software 28 (1) (1995) 9–18.
[38] M. Torchiano, M. Morisio, Overlooked aspects of COTS-based development, IEEE Software 21 (2) (2004) 88–93.
[39] M. Turner, B. Kitchenham, D. Budgen, O.P. Brereton, Lessons learnt undertaking a large-scale systematic literature review, in: Proceedings of EASE'08, British Computer Society, 2008.
[40] C. Zannier, G. Melnik, F. Maurer, On the success of empirical studies in the international conference on software engineering, ICSE06 (2006) 341–350.
[41] M. Zelkowitz, D. Wallace, Experimental validation in software engineering,

Information and Software Technology 39 (1997) 735–743.
[42] M. Zelkowitz, D. Wallace, Experimental models for validating computer technology, IEEE Computer 31 (5) (1998) 23–31.