Information and Software Technology 55 (2013) 2049–2075

Contents lists available at ScienceDirect

Information and Software Technology

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

A systematic review of systematic review process research in software engineering

Barbara Kitchenham *, Pearl Brereton

School of Computing and Mathematics, Keele University, Staffordshire ST5 5BG, United Kingdom

ARTICLE INFO

Article history:
Received 20 February 2013
Received in revised form 25 July 2013
Accepted 26 July 2013
Available online 3 August 2013

Keywords:
Systematic review
Systematic literature review
Systematic review methodology
Mapping study

ABSTRACT

Context: Many researchers adopting systematic reviews (SRs) have also published papers discussing problems with the SR methodology and suggestions for improving it. Since guidelines for SRs in software engineering (SE) were last updated in 2007, we believe it is time to investigate whether the guidelines need to be amended in the light of recent research.
Objective: To identify, evaluate and synthesize research published by software engineering researchers concerning their experiences of performing SRs and their proposals for improving the SR process.
Method: We undertook a systematic review of papers reporting experiences of undertaking SRs and/or discussing techniques that could be used to improve the SR process. Studies were classified with respect to the stage in the SR process they addressed, whether they related to education or problems faced by novices and whether they proposed the use of textual analysis tools.
Results: We identified 68 papers reporting 63 unique studies published in SE conferences and journals between 2005 and mid-2012. The most common criticisms of SRs were that they take a long time, that SE digital libraries are not appropriate for broad literature searches and that assessing the quality of empirical studies of different types is difficult.
Conclusion: We recommend removing advice to use structured questions to construct search strings and including advice to use a quasi-gold standard based on a limited manual search to assist the construction of search stings and evaluation of the search process. Textual analysis tools are likely to be useful for inclusion/exclusion decisions and search string construction but require more stringent evaluation. SE researchers would benefit from tools to manage the SR process but existing tools need independent validation. Quality assessment of studies using a variety of empirical methods remains a major problem.

Contents

* Corresponding author. Tel.: +44 1782 733979; fax: +44 1782 734268.
E-mail addresses: b.a.kitchenham@keele.ac.uk (B. Kitchenham), o.p.brereton@keele.ac.uk (P. Brereton).

0950-5849/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.infsof.2013.07.010

1. Introduction

In 2004 and 2005, Kitchenham, Dybå and Jørgensen proposed the adoption of evidence-based software engineering (EBSE) and the use of systematic reviews of the software engineering literature to support EBSE [18,7]. Since then, systematic reviews (SRs) have become increasingly popular in empirical software engineering as demonstrated by three tertiary studies reporting the numbers of such studies [15,12,4]. Many of these studies adopted the guidelines for undertaking systematic review, based on medical standards, proposed by Kitchenham [17], and revised first by Biolchini et al. [2] to take into account practical problems associated with using the guidelines and later by Kitchenham and Charters [16] who incorporated approaches to systematic reviews proposed by sociologists.

As software engineers began to use the SR technology, many researchers also began to comment on the SR process itself. Brereton et al. [1] wrote one of the first papers that commented on issues connected with performing SRs and many such papers have followed since, for example:

- Staples and Niazi [35,34] discussed the issues they faced extracting and aggregating qualitative information.
- Budgen et al. [3] and Petersen et al. [26] identified the difference between mapping studies and conventional systematic reviews.
- Kitchenham et al. [14] considered the use of SRs and mapping studies in an educational setting.
- MacDonell et al. [21] and Kitchenham et al. [11] studied the claims of the SR technology with respect to reliability/consistency.
- Dieste and Padua [5] and Skoglund and Runeson [32] investigated how to improve the search process.
- Kitchenham et al. [13] investigated how best to evaluate the quality of primary studies (i.e. the empirical studies found by the systematic review search and selection process).

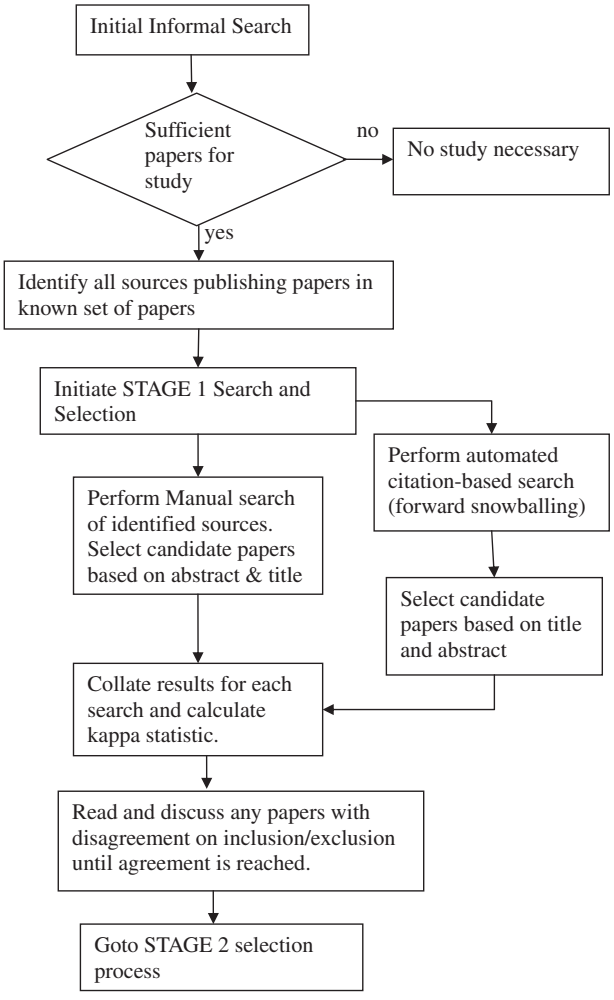

Fig. 1. Initial search and Stage 1 search and selection process.

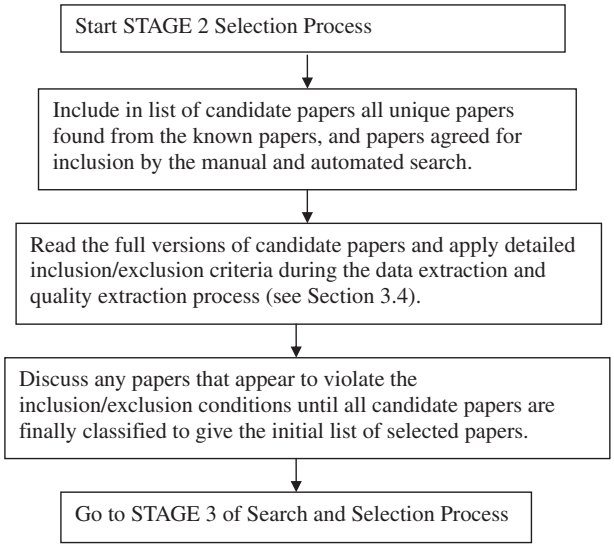

Fig. 2. Stage 2 selection process.

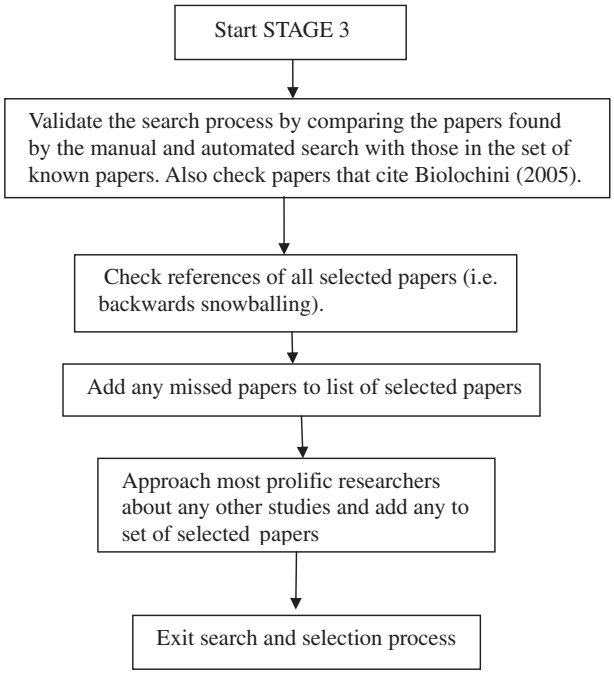

Fig. 3. Stage 3 search and selection process – validation & snowballing & contacting researchers.

It therefore seems appropriate to identify the current status of such studies in software engineering, and identify whether there is evidence for revising and/or extending the guidelines for performing systematic reviews in software engineering. To that end we undertook a systematic review of papers that discuss problems with the current SR guidelines and/or propose methods to address those problems.

Section 2 discusses the aims of our research, reports related research and identifies the specific research questions we address. Section 3 reports the search and paper selection process we adopted and reports the basic limitations of our approach. Section 4 reports the outcome of our search and selection process and its validity. We also report the reliability of our data extraction and quality assessment process. Section 5 presents our aggregation and synthesis of information from the papers we included in the study. Section 6 discusses our results and the limitations that arose during our study. We present our conclusions in Section 7.

2. Aims and research questions

Our aim is to assess whether our guidelines for performing systematic reviews in software engineering need to be amended to reflect the results of methodological investigations of SRs undertaken by software engineering researchers. In order to do this we undertook a systematic review of papers reporting experiences of using the SR methodology and/or investigating the SR process in software engineering (SE). We use this information to assess whether SRs have delivered the expected benefits to SE, to identify problems found by software engineering researchers when undertaking SRs, and to identify and assess proposals aimed at addressing perceived problems with the SR methodology.

There have been two mapping studies that address methods for supporting SRs. Felizardo et al. [8] report a mapping study of the use of visual data mining (VDM) techniques to support SRs. Their mapping study concentrated on a specific technique and was not restricted to SE studies. In contrast, our SR considers a broader range of techniques but is restricted to studies in the SE domain. Marshall and Brereton [22] have undertaken a mapping study of tools to support SRs in SE. Compared with our study:

- Their mapping study focused specifically on tools for SRs in the SE.
- They used a search string-based automated search process, using papers identified in this study as a set of known studies to refine their search strings.
- The time period of their search was longer, going from 2005 to the end of 2012.

Thus the value of this study is that it addresses a wider range of technologies than either of the mapping studies, and as an SR provides a more in-depth aggregation of the results of the identified primary studies.

Our SR addresses the following research questions:

RQ1. What papers report experiences of using the SR methodology and/or investigate the SR process in software engineering between the years 2005 and 2012 (to June)?
RQ2. To what extent has research confirmed the claims of the SR methodology?
RQ3. What problems have been observed by SE researchers when undertaking SRs?
RQ4. What advice and/or techniques related to performing SR tasks have been proposed and what is the strength of evidence supporting them?

3. Search and selection process

Before starting our SR, we produced a review protocol which is summarized in this section. Figs. 1–3 give an overview of the search and selection process which are described in more detail below.

3.1. Initial search process

Kitchenham undertook an initial informal search of two conference proceedings (Evaluation and Assessment in Software engineering and Empirical Software Engineering and Measurement) from 2005 to mid 2012 which together with personal knowledge identified 55 papers related to methods for performing systematic reviews and mapping studies in SE. This initial search confirmed

that there are a substantial number of papers on the topic and that a systematic review would be appropriate. It also provided the information needed to guide the manual search process.

3.2. Search and selection process

3.2.1. Stage 1 manual search and selection

The 55 known papers identified the main sources of papers on methodology to be:

- Evaluation and Assessment in Software Engineering (EASE): 21 papers.
- Empirical Software Engineering and Measurement (ESEM): 18 papers.
- Information and Software Technology (IST): 6 papers.
- Empirical Software Engineering Journal (ESE): 2 papers.
- Journal of Systems and Software (JSS): 2 papers.
- International Conference on Software Engineering (ICSE): 2 papers.

Five other sources each published a single SR methodology paper:

- Empirical Assessment for Software Technologies (EAST).
- Advanced Engineering Informatics.
- IEEE Transactions of Software Engineering (TSE).
- Lecture notes on Computer Science Volume 5089.
- Proceedings of Psychology of Programming Special Interest Group (PPIG) '08.

Of these sources only EAST which is targeted at evidence-based software engineering and systematic reviews was both relevant and unlikely to be found by an automated search. Kitchenham attended EAST 2012 and identified relevant papers at the workshop.

We both undertook an independent manual search of the main sources from 2005 to June 2012 (with ESEM 2012 being searched using the published program) and classified each paper as included or excluded. The emphasis of the manual search was on including papers unless they were clearly irrelevant. The results of the two searches were collated and any papers we disagreed about were read and then discussed. If we could not come to an agreement about a paper we classified it as "include".

3.2.2. Stage 1 citation-based search and selection

To support the manual search, an automated search based on citation analysis (also known as forward snowballing) was performed. Kitchenham searched SCOPUS for all papers referencing the following papers:

- Kitchenham, B.A., S. Charters (2007). Guidelines for performing systematic literature reviews in software engineering (Search date 25th June 2012).
- Kitchenham, B. (2004). Procedures for undertaking systematic reviews.
- Kitchenham, B., Tore Dybå, and Magne Jørgensen (2004). Evidence-based Software Engineering. ICSE (Search date 25th June 2012).
- Dybå, Tore, Barbara Kitchenham, and Magne Jørgensen (2005). Evidence-based Software Engineering for Practitioners. IEEE Software (Search date 25th June 2012).
- Brereton, Pearl, Barbara A. Kitchenham, David Budgen, Mark Turner, Mohamed Khalil (2007). Lessons from applying the systematic literature review process within the software engineering domain (Search date 29th June 2012).

After removing duplicates, we both evaluated each paper for inclusion in the set of candidate papers based on title and abstract. The main emphasis was to include papers unless they were clearly irrelevant. The decisions of each author were collated. Papers which both authors agreed to include were included and any papers which both authors agreed to exclude were excluded. Any papers for which the inclusion/exclusion assessment differed between authors were discussed until either agreement was reached or the paper was provisionally included.

3.2.3. Stage 2 selection processes

Papers in the known set, the manual search inclusion set, and the automated search inclusion set were collated into a set of candidate papers. Any papers excluded in one search and selection process but included in the other or the known set, were identified and further discussed. If we could not come to a decision about a paper it was included. The final set of selected papers entered the data extraction process.

The final inclusion/exclusion decision took place when full papers were read in parallel with data extraction and quality assessment. On finding a paper whose relevance was questionable, the researcher notified the other researcher and explained why the paper was suspect. The other researcher either agreed to exclude the paper or entered into discussion about its relevance. Discussions continued until we both agreed about the final classification of the paper.

3.2.4. Stage 3 search and selection

Stage 3 took place in parallel with data and quality extraction from studies identified in Stage 1 and Stage 2. It comprised three main tasks:

1. Search process validation. See Section 3.3.
2. Backward snowballing. Once the search process and initial data and quality extraction was completed, the references of the selected papers were reviewed and any missing candidate papers were assessed against the inclusion/exclusion criteria. The set of selected papers was updated to include any additional relevant papers found by snowballing.
3. Approaching individual researchers. After snowballing, we approached any researchers or research group that produced more than two papers included in the set of selected papers and asked them if they had any other papers or research reports related to SR methodology. Any such papers were added to the set of selected papers.

3.2.5. Primary study identification

The relationship between selected papers and primary studies can be complex since several different papers can report the same study and a paper can report several different studies. In our study the relationship was mapped as follows:

- Papers were each given a unique identifier of the form PX where X is a unique integer.
- Each paper was given a study number of the form SY where Y is an integer but not necessarily unique. If a paper reported the same study as another paper each was given the same study identifier.
- If papers by the same authors refer to the same topic but use different materials/subjects for validation, they were given different study numbers.
- If papers reported multiple studies, we distinguished between validation replications i.e. studies using the same experimental method but different materials and independent validations i.e. validation that use different experimental methods and/or, in

the case of formal experiments, use different human subjects. Replication validations were treated as multi-case case studies (and were only given one quality assessment since the methodology was the same) Replication validations increase the scope/size of a study not its quality. Independent validations were treated as separate studies and were given individual quality assessments. Separate studies reported in the same paper were given an additional identifier i.e. SY.a, SY.b to identify the individual study.

3.3. Search and selection validation

There were two aspects to search validation:

1. The papers found by the manual search and the citation based search were compared with the set of known papers to assess the completeness of the manual and citation search. If the manual search and selection process was performed effectively only the papers published in sources that were not searched would have been missed. If the automated search was performed effectively only the papers either not yet published or not indexed (including all EASE papers except EASE 2011) would have been missed.
2. SCOPUS was used to find papers that cited the Biolchini et al. [2] guidelines for systematic reviews. Papers relating to assessing SR methodology were identified and compared with the set of all papers that were found by the search process or were already known.

Selection validation was based on the Kappa agreement achieved between the authors for the manual and automated searches.

3.4. Inclusion and exclusion criteria

The aim of this systematic review was to identify and classify papers related to SR methodological issues in the context of software engineering, including papers related to quality assessment of primary studies. The inclusion criteria were therefore:

1. That the main objective of the paper which may be a primary, secondary or tertiary study was either to discuss or investigate a methodological issue related to systematic reviews. This inclusion criterion defines the basic scope of our study.
2. That the paper discusses or investigates the construction of and/or evaluation of quality instruments used to assess the quality of primary studies or the general strength of evidence. Quality evaluation of primary studies is an important and difficult element of a software engineering SR, so we decided to include papers that investigated quality evaluation, even if they were not primarily aimed at improving SR methods.
3. That the paper must have a software engineering context. To keep our study to manageable levels, the scope of our study was restricted to SE related papers. We feel this is justified because many of the problems being addressed are related to limitations of SE digital sources and the empirical methods used in SE.
4. That the paper must be written in English. We did not believe that many important studies would be published in languages other than English. For example, although many SR related papers have been published by South American authors, the majority of their studies were written in English. The same is true of studies reported by Northern European researchers.
5. That short papers which fulfill the above criteria be included. We had no reason to believe that short papers would fail to provide sufficient levels of detail to report focused methodological studies.

Note that different papers related to the same study were kept in the set of selected papers but identified as linked papers.

The exclusion criteria were:

1. Secondary or tertiary studies whose main objective was to report the results of a systematic review or mapping study. Thus we excluded papers that commented on problems with searches or other processes as part of reporting an SR or mapping study. This decision was to ensure that papers included in our study would have undertaken a systematic investigation of the methodology issue, as well as to avoid the need to find and read every systematic review published in the software engineering domain.
2. Papers discussing EBSE principles. EBSE is a wider topic that systematic reviews thus papers on general EBSE topics were outside the scope of our study.
3. Methodological studies with general (i.e. non-software engineering) focus. To restrict our search to manageable levels, we did not try to find methodologically-based studies performed outside the SE domain.
4. Papers for which only PowerPoint presentations or extended abstracts were available. Studies reported only by abstract or slides would not provide sufficient information to be included in the set of selected papers.
5. Papers producing guidelines for performing or reporting primary studies (i.e. empirical studies performing evaluations of a methodology) as opposed to guidelines for quality evaluation of primary studies. Procedures for performing or reporting primary studies are outside the scope of our study.

In particular, our selected papers excluded:

1. The three tertiary studies which were aimed at classifying software engineering SRs, i.e. Kitchenham et al. [15,12–14] and da Silva et al. [4]. These studies discuss the quality of SRs but are not primarily about the SR methodology.
2. Papers that describe guidelines for SRs in software engineering [17,2,16]. The most recent set of guidelines will be assessed in the light of recommendations obtained from the primary studies in this study in terms of how it should be amended or extended.
3. Papers reporting studies that developed or evaluated guidelines for performing empirical studies or reporting empirical studies rather than evaluating the quality of empirical studies. For example, the paper by Verner et al. [40] produced guidelines for performing cases studies, and would be excluded. Similarly, the guidelines for reporting experiments produces by Jedlitschka et al. [10] are also excluded. In contrast, although their main purpose was to produce guidelines for conducting and reporting of case studies, the paper by Runeson and Höst [31] includes a checklist for readers which can be considered to be a quality checklist, so we include their paper in our set of included papers.

3.5. Quality assessment

Our primary studies were of many different types: case studies, surveys, secondary studies, etc. Rather than using multiple instruments to assess the quality of the study, we classified the type of study and used a generic set of questions to evaluate its quality. We used the quality instrument developed by Dybå and Dingsøyr [6] since it is applicable to most types of study but unlike them we did not rejected discussion papers or lessons learnt papers. We hoped that using a common generic set of criteria would make it possible to compare the quality of papers using different

research methods. However, this proved problematic as discussed below.

We intended to use discussion studies and lessons learnt studies to identify issues and/or problems associated with using systematic literature reviews in the context of software engineering. Also we intended to include good quality lessons learnt studies together with empirical studies when we assessed the strength of evidence associated with any suggestions for SR process change.

The checklist we used was:

1. Is the paper based on research (or is it a discussion paper based on expert opinion)? Yes/No.
2. What research method was used: Experiment, Quasi-Experiment, Lessons learnt, Case study, Opinion Survey, Tertiary Study, Other (specify)? Note This is to be based on our reading of the paper not the method claimed by the author of the paper.
3. Is there a clear statement of the aims of the study? Yes/Partly/No. Score as 1, 0.5, 0. Interpolation is permitted.
4. Is there an adequate description of the context in which the research or observation was carried out? Yes/Partly/No. Score as 1, 0.5, 0. Interpolation is permitted.
5. Was the research method appropriate to address the aims of the research? Yes/Partly/No/Not applicable (i.e. Expert Opinion). Score as 1, 0.5, 0 or mark NA. Interpolation is permitted for numerical values.
6. Was the recruitment strategy (for human-based experiments and quasi-experiments) or experimental material or context (for lessons learnt) appropriate to the aims of the research? Yes/Partly/No/Not applicable (i.e. Expert Opinion). Score as 1, 0.5, 0 or mark NA. Interpolation is permitted for numerical values.
7. For empirical studies (apart from Lessons Learnt), was there a control group or baseline with which to evaluate SR procedures/techniques? Yes/Partly/No/Not applicable (i.e. Lessons Learnt or Expert opinion) Score as 1, 0.5, 0 or mark NA. Interpolation is permitted for numerical values.
8. For empirical studies (apart from Lessons Learnt), was the data collected in a way that addressed the research issue? Yes/Partly/No/Not applicable (i.e. Lessons learnt or Expert opinion). Score as 1, 0.5, 0 or mark NA. Interpolation is permitted for numerical values.
9. For empirical studies (apart from Lessons Learnt), was the data analysis sufficiently rigorous? Yes/Partly/No/Not applicable (i.e. Lessons Learnt or Expert opinion). Score as 1, 0.5, 0 or mark NA. Interpolation is permitted for numerical values.
10. Has the relationship between researcher and participants been considered to an adequate degree? Yes/Partly/No. Score as 1, 0.5, 0. Interpolation is permitted.
11. Is there a clear statement of findings? Yes/Partly/No. Score as 1, 0.5, 0. Interpolation is permitted.
12. Is the study of value for research or practice? Yes/Partly/No. Score as 1, 0.5, 0. Interpolation is permitted.

We both extracted the quality data from each primary study independently. The results were collated and any disagreements were discussed until agreement was reached. The quality extraction was done in parallel with data extraction. We did not plan to exclude any studies based on the quality score because the quality score related to the validation exercise and a methodology proposal might be worth considering even if its evaluation was poorly performed. We intended to use the quality score to assess the overall weight of evidence but this proved problematic as discussed below.

We immediately noted some problems with the approach:

- Although we identified broadly which questions would be inappropriate for certain types of study, we found some questions were inappropriate due to the context of the study. For example, if the study was based on Monte Carlo simulation or another researcher's SR results, question 10 concerning the relationship between subjects and experimenters would be inappropriate. This is discussed in more detail in Section 4.5. In practice, we not only assessed independently the value of each question for a specific study, but also assessed independently whether the question was appropriate in the context of the study.
- Our assessment of validation method type differed frequently from that of the authors of the study. The most common differences were that, if a case study was based on an opinion survey we called it "Opinion survey" rather than "Case Study", and if a study was based on a post-hoc re-analysis of a SR we called it an "Example" not a "Case Study" keeping the term "Case study" for an evaluation that was performed as part of undertaking an SR. We also identified very small experiments (e.g. 4 or fewer subjects) and small examples (e.g. one that only considered a part of the relevant data set or a small part of a specific task).
- We found that using the checklist, small studies could obtain good scores although, by nature of their limited size, they could provide only limited evidence of the value of the methodology. For example, if the aim of the study was to undertake a preliminary feasibility study of a methodology, it could score well on all checklist questions although overall it could only be said to provide very limited evidence of the real value of the methodology. Furthermore, some lessons learnt and experience papers scored quite well because only a relatively few checklist questions were relevant. It seemed that the quality score should only be used to differentiate between studies of the same type and size. For this reason, we did not exclude papers based on their quality score but we considered the type of study, its size and its quality score when discussing the overall weight of evidence in favour of the methodology.

3.6. Data extraction

Kitchenham extracted standard information from each paper, i.e.:

- Primary study ID.
- Author(s).
- Title.
- Publication venue.
- Date of publication.
- Publication details for journal (Volume and Issue).
- Page numbers (if available).

We both extracted the primary study specific data for each paper that was based on a preliminary categorization of the known studies. It included:

1. Type of Paper: Problem identification and/or problem solution (PI) or Experience Paper, Opinion Survey or Discussion paper (E).
2. Scope of the study: Mapping studies/Conventional Systematic review/Both/Updating studies/Other (which must be specified).
3. Summary of aims of Study.
4. Main topics covered (NOT mutually exclusive):
   a. Educational issues: Yes/No.
   b. SR Participant Viewpoint: Experience Researcher (E)/Novice (N)/Not specified (NS).
   c. Research questions: Yes/No.

   d. SR claims: Repeatability, Auditability, Objectivity, Value, Other (Specify).
   e. Protocol Development: Yes/No.
   f. Search processes: Yes/No.
   g. Search validation/evaluation: Yes/No.
   h. Selection processes: Yes/No.
   i. Quality evaluation of primary studies: Yes/No.
   j. Data Extraction: Yes/No.
   k. Data Synthesis: Yes/No.
   l. Reporting: Yes/No.
5. Method proposed: Name or description (e.g. Quasi-Gold Standard, Visual Text Mining).
6. Validation/Evaluation performed: Yes/No.
7. Actual Validation method (as judged by each researcher): Experiment, Quasi Experiment, Tertiary Study, Case study, Data Mining (i.e. papers analyzing historical data sets), Opinion survey (Interview), Opinion Survey (Questionnaire), Lesson Learnt, Example, Other (to be specified).
8. Claimed Validation method (as specified by authors of paper).
9. Summary of main results. Note details of lessons learnt and opinion survey results will be collected in a separate word file.
10. Any process recommendations (suggested by data extractors).

A data collection form was set up in an Excel spread sheet and finalized after both authors trialed the data extraction on several papers.

Discussion papers, lessons learnt papers and opinion surveys were treated differently from other studies. Relevant fields for lessons learnt, surveys and discussion papers were included in the Excel spread sheet depending on the scope of the paper. If the scope of the paper was very broad (i.e. all aspects of an SR and/or the results included comments from a large variety of subjects), no attempt was made to complete the Excel form. If the paper covered a very specific topic and had a limited number of results and process recommendations, the Excel sheet was completed for the paper.

Lessons learnt and opinion papers that had a broad scope had a text based data extraction form for each study that permitted individual textual elements to be extracted. The format of this form is shown in Appendix A.

For the standard data extraction process, we both extracted the data from each primary study independently. The results were collated and any disagreements were discussed until agreement was reached. For the textual data extraction, Kitchenham performed the extraction and Brereton checked the extraction.

3.7. Data and quality extraction reliability

Kitchenham checked the level of agreement achieved for data and quality extraction. In the case of quality extraction the Pearson correlation coefficient was found between the values for each assessor for each paper both for the number of appropriate questions and for the average quality score for each paper. In the case of data extraction, the agreement with respect to the study categories was assessed using the Kappa statistic.

3.8. Data aggregation and synthesis

Information from lessons learnt, surveys and discussion studies was reviewed and any process issues raised by these studies was collated and recorded in the data collection form shown in Appendix A. The problems and advice mentioned in more than one paper were collated by comparing the results extracted from each study and looking for similarities, using an approach similar to the meta-ethnography approach proposed by Noblit and Hare [24]. This was done in three stages. Firstly Kitchenham extracted individual issues from the text and tables in the terminology used in the paper, linking the issue to its position in the paper. This was then checked by Brereton and any disagreements noted. Next, Kitchenham extracted from each paper the issues that seemed most important (i.e. were mentioned by many subjects in a specific paper, were mentioned in several other papers, or corresponded to our own experience). In addition, repeated issues (e.g. issues that were mentioned both in the discussion and the conclusions) were identified as single issues. The extracted issues were summarized using a more consistent terminology. The summarization involved abstracting specific themes in cases where many different specific issues were raised (for example problems with constructing search strings resulted in a number of differently specified problems). Then the issues from each paper were integrated into two lists, one for problems and one for advice, by comparing the important issues from each paper and including any issue that was mentioned at least twice. The lists for each paper and the integrated lists were checked by Brereton and all disagreements were discussed and resolved.

Studies covered by the classification scheme were grouped into sets of studies addressing similar issues – note some primary studies were relevant to several different categories. Within each category, papers were grouped with respect to the specific technique being proposed or the particular task in the SR process. Some categories were not analyzed explicitly because only one or two papers investigated that issue (i.e. protocol production and research questions). Other categories were concatenated into joint categories (i.e. novice participant type and education, searching, search validation and selection, quality evaluation and checklists, data extraction and data synthesis). In one case we noted a specific technology (i.e. textual analysis) was recommended for a variety of different tasks. We treated papers discussing the use of textual analysis as a separate set of related papers. After grouping related papers, we used narrative synthesis to discuss the results reported by papers addressing similar topics.

Results from each set of related primary studies were collated and assessed for:

- Consistency (i.e. the extent to which results reported on a specific issue from different studies were consistent).
- Strength of evidence based on the number, type and quality of studies that reported the results.

After our initial aggregation, we reviewed the recommendations found in the individual papers. We classified each recommendation based on whether it was relevant to the guidelines, was already covered in the guidelines, had already been mentioned during our discussion of the paper, or needed to be included in the discussion. We also looked for any general trends that had not been previously discussed but indicated an issue that needed to be addressed in the guidelines. We integrated the results from our synthesis with the recommendations we found in the individual papers. These recommendations were then used to specify changes required to the current guidelines.

3.9. Limitations of the research method

One significant limitation is that we would be collecting data from some papers that we ourselves authored. This can lead to two problems:

- We may base our assessment of the answers to data extraction questions on our understanding of our papers not just the information that was reported, potentially losing traceability.
- We may be systematically too lax (or stringent) in our evaluation of the quality of our own papers.

There is no way to completely avoid personal bias. We performed our extractions independently and tried to be rigorous in assessing the reason for any disagreements, if necessary tracking the issue to parts of the paper's text. The final extraction was agreed by both researchers to correspond to data reported in the paper.

Another limitation is that we restricted our automated search to citation analysis of five specific EBSE and SR related papers, so may have missed papers that would have been found by a broader search. The reason for our restriction was twofold: firstly, we wanted to avoid large numbers of papers from outside the SE domain, secondly, we expected that SE researchers commenting on process issues would base any criticism on SE related guidelines. We also used citations of the Biolchini et al. [2] guidelines as a check on the set of papers found by our automated search since these guidelines were written by an independent group of researchers.

We based our search on only one digital source i.e. SCOPUS. Since we based our automated search on citation analysis (i.e. forward snowballing), we were restricted to a general indexing system that supported such analysis. To reduce any bias introduced by using a single digital indexing source, we also performed a manual search of important sources, undertook backward snowballing (i.e. searching the reference lists of the primary studies we found by our main search process) and approached individual authors to determine whether they had published any relevant material we had missed.

A final limitation is our decision to exclude papers that mentioned process issues as an additional issue as part of an SR or mapping study. This was again necessary both to restrict our primary studies to those that would have collected information about methodological issues systematically and to reduce the number of papers we needed to read to manageable levels, but it means we may have missed some relevant papers.

4. Included and excluded studies and validity

This section reports the outcome of our search and selection process and presents our various validity checks, including the reliability of our data extraction and quality assessment process.

4.1. Stage 1 and Stage 2 search and selection

Our initial informal search identified 55 known papers. Subsequently three papers were removed, leaving 52 known papers.

Our citation search, performed during June 2012 using SCOPUS, found 410 unique papers (see Table 1). After we assessed each paper individually, our initial inclusion/exclusion assignments agreed for 398 papers and disagreed for 12 papers giving a Kappa agreement of 0.844 (see Table 2). The precision of the automated search and selection process was $100 \times 45/410 = 11\%$.

Our manual search took place between July–August 2012 (including review of accepted papers for ESEM 2012). The results are shown in Table 3. The Kappa values for each source and overall showed good levels of agreement.

After reading and discussing the 15 papers we originally disagreed about, 11 of the papers were included in the data extraction phase. Thus, a total of 54 candidate papers were found by the manual search. The precision of the manual search and selection process was $100 \times 54/3360 = 1.6\%$.

Table 1
Automated search results.

Source paper	Papers found
Kitchenham [17]	178
Kitchenham and Charters [16]	150
Brereton et al. [1]	80
Kitchenham et al. [18]	96
Dybå et al. [7]	75
Unique papers	410

Table 2
Automated search selection process.

Results of assessing title and abstract	Results
Initial Agreed Include	37 papers
Initial Agreed Exclude	361 papers
Disagreed	12 papers
Kappa	0.844
Agreed include after discussion	8 papers
Agreed exclude after discussion	4 papers
Final number included	45 papers

After collating the known papers, the papers found by the automated search and the papers found by the manual search we identified a total of 76 papers to include in the quality and data extraction process. However, there were anomalies in the results i.e. some papers included in one set of papers were found but excluded in another set. After discussing anomalies between the different search and selection processes three papers were removed from the set of known papers because they were rejected during the manual selection. Two were more relevant to EBSE rather than SRs:

- Rainer et al. [29].
- Rainer and Beecham [28].

One was a poster not a full paper:

- Woodall and Brereton [39].

Thus 73 papers entered the quality extraction and data extraction process. During data and quality extraction 10 papers were found to violate the detailed inclusion/exclusion criteria. These papers and the reasons for their exclusion are reported in Appendix B. Thus, 63 unique papers were included in the initial set of selected papers.

However, another five papers were found after the initial data and quality extraction:

- 10 candidate papers were found by snowballing the references of the initially selected papers. After assessing each paper, we agreed that three of the papers should be included.
- After contacting the most prolific authors (i.e. Dybå, Cruzes, Dieste, Maldonado, Zhang, Babar) we located one extra paper.
- After attending the EAST 2012 workshop, we found one more paper.

Thus, our final set of selected papers comprised 68 unique papers (see Appendix C). However, not all the papers reported unique primary studies (see Section 4.3).

4.2. Manual and automated search validation

Table 4 shows the effectiveness of the manual and automated searches relative to the known set of papers. The overall assessment of the process was based on the number of unique papers

Table 3
Results of manual search.

Source	Papers Agreed Include Phase 1	Papers Agreed Exclude Phase 1	Papers Disagreed Phase 1	Papers Total	Kappa
Evaluation and Assessment in Software Engineering (EASE)	18	111	8	139	0.783
Empirical Software Engineering and Measurement (ESEM)	16	317	5	338	0.857
Empirical Software Engineering Journal (ESE)	3	177	1	181	0.854
Information and Software Technology (IST)	2	710	0	712	1
Journal of Systems and Software (JSS)	2	1333	0	1335	1
International Software Engineering Conference (ICSE)	2	710	1	713	0.799
Total	43	3360	15	3418	0.849

Table 4
Effectiveness of manual and automated search.

Search process	Number of known papers found	Number of known papers that should have been found	Percentage of Known papers that should have been found
Manual Search	45	46	$100 \times 45/46 = 97.8$
Citation Search	29	36	$100 \times 29/36 = 80.6$
Overall	47	49	$100 \times 47/49 = 95.9$

found by the overall search process (i.e. papers found by both manual and automated searches were counted only once). Note this analysis was completed prior to data extraction and includes the 10 papers that were subsequently rejected.

Each search reached a reasonable level of effectiveness although the manual search was more effective. However, the manual search had worse precision than the automated search (1.6% compared with 11%). The automated search missed most of the papers published in EASE proceedings because until 2010 the EASE proceedings (although available online) were not indexed by SCOPUS (or any other indexing system). The automated search also missed some papers because SCOPUS did not immediately recognize the Kitchenham [17] guidelines (which appeared in a technical report not a published paper) as a document that should have its citations collated.

A citation search using SCOPUS based on the guidelines produced by Biolochini et al. [2] undertaken on 25th October 2012 found 48 papers of which six were methodology papers and all six had already been found by our search process.

4.3. Relationship between papers and primary studies

The papers included in this review are shown in Appendix C. The first 63 papers were found by phases 1 and 2 of the search and selection process, the last five papers were found by phase 3 of the search and selection process. 10 papers were duplicate reports of previously reported studies (i.e. 9 journal papers were based on previous conference papers and in one case two separate conference papers reported the same study). Different papers reporting the same primary study have different papers numbers but share a study number. In these cases, data was extracted from the most recent paper and if necessary additional information was sought from the earlier papers. The duplicate reports that were excluded from the data extraction are shown in italics in Appendix C. We have cited the duplicate reports to increase the repeatability of our study. If we included only the most recent paper, other researchers would not know whether other related papers they found had been found by our search process and rejected (as duplicates) or not found at all.

Six papers reported multiple studies but two of these papers were duplicate reports of studies. Five of the multiple study papers reported two primary studies and one reported three primary studies. The four unique multi-study papers in this SR reported a total of 7 primary studies. Multiple studies are identified by a letter (*a*, *b*, or *c*) added to the study number. Note however, as discussed in Section 3.2.5 we have not counted as separate studies, papers that reported several primary studies where the primary studies used the same methodology and addressed the same research questions. In this case the multiple studies are treated as close replications. The impact of the replication is to increase the size/scope of the primary study not to change the quality score.

One paper (P61) referred to three different studies that were reported in two previous conference papers (P1 and P60). However, the study reported in P1 was only reported very briefly in P61, so we have treated the three papers as reporting one study in P1 and two studies in both P60 and P61, thus contributing three independent studies to this SR. So, overall in answer to RQ1 which asked what papers relating to SR methodology were published during the period 2005 to October 2012, we found 68 papers discussing issues related to SR methodologies which related to 63 unique studies.

4.4. Data extraction and quality assessment reliability

Although we defined the data collection process in our protocol and discussed our first few extractions to try and achieve consistency, the initial inter-rater reliability of our extractions was problematic.

The reliability assessment of our quality evaluation was based on 54 papers. "Pure" discussion paper i.e. papers that did not include a validation element were not evaluated for quality i.e. papers P13, P24, P50, P52, P64. We also initially disagreed about whether four papers which had only limited validation should be treated as discussion papers or validation papers. These four papers were excluded from the assessment of quality assessment reliability.

We had expected some of the criteria to be inappropriate for specific types of paper as noted in Section 3.5, however, we found that in some cases we disagreed about whether a specific quality question was relevant or not based on the particular study not just the type of study. The Pearson correlation between the number of questions each of us believed to be relevant for 54 studies was 0.67 which although statistically significant ($p < 0.001$) shows a disturbing level of disagreement. Reliability was even worse for the average scores for each study, where the correlation between our scores was 0.54 which is statistically significant ($p < 0.001$) but still disappointingly low.

Table 5
Initial agreement with respect to study categories during data extraction.

Data extracted	Categories	Agreement	Total assessment	Kappa
Type of study	Problem or solution investigation paper/Discussion, opinion survey, lessons leant	58	63	0.795
Focus of study	SRs/Mapping study/Both/Not applicable	37	63	0.413
Education/training related	Yes (identifying claim)/No (not applicable/blanks counted as No)	47	49	0.810
Takes a specific viewpoint	Novice/Expert/Both/Not applicable	28	49	0.277
Protocol related	Yes/No (not applicable/blanks counted as No)	46	49	0.347
Discussed SLR claims	Yes (specified claims not considered for kappa analysis)/No (not applicable/blanks counted as No)	43	49	0.624
Research question related	Yes/No (not applicable/blanks counted as No)	46	49	0.846
Related to search process	Yes/No (not applicable/blanks counted as No)	46	49	0.840
Related to search validation	Yes/No (not applicable/blanks counted as No)	47	49	0.778
Related to paper selection	Yes/No (not applicable/blanks counted as No)	46	49	0.847
Related to quality assessment	Yes/No (not applicable/blanks counted as No)	43	49	0.689
Related to data extraction	Yes/No (not applicable/blanks counted as No)	38	49	0.543
Related to data synthesis	Yes/No (not applicable/blanks counted as No)	37	49	0.372
Related to reporting	Yes/No (not applicable/blanks counted as No)	43	49	0.344
Validation method	Example, Experiment, Quasi Experiment, Lessons learnt, Opinion survey, Case study, Tertiary study (excluding other specified types)	22	33	0.507

Our initial agreement with respect to categorical data is shown in Table 5. The number of studies in each category is not identical. Some of the papers had data collected in a different manner because they were broad lessons learnt or opinion survey papers (see the studies reported in Table 6, except for study 52a which was a tertiary study and subjected to the normal data extraction process). These papers were excluded from the Kappa analysis except for the initial assessment of paper type and focus of study which was collected for all studies. In addition, the validation methods reliability was only applied to studies that included a validation element (i.e. not simple discussion papers) and restricted to studies that were classified according to the categories indicated in the table. Many studies were classified into types we had not anticipated such as "Monte-Carlo simulation", "Observational Studies", "Correlation studies". Also some studies used multiple methods. If there was clear distinction between individual empirical methods in studies applying multiple methods we separated them into different studies, but when a single study used a variety of different approaches (e.g. some qualitative data and some quantitative data) to address the same research question, we felt it was inappropriate to treat them as separate studies. Overall, one of the main reasons for disagreement was that studies often mentioned several steps in the SR process but reported in detail only one or two steps. We only recognized somewhat late in the data extraction process that we were only interested in categorizing a study against SR steps that were discussed or investigated in detail, not against all the steps that were mentioned as part of the evaluation exercise.

Our reliability was particularly poor with respect to deciding whether the study focused on a particular type of SR (conventional SR or mapping study), whether the study took a specific viewpoint (i.e. novice, expert, or both), whether it was protocol related, whether it related to data extraction and whether it related to data synthesis. Of these categories, we have only considered papers related to novices and papers related to data aggregation and synthesis explicitly in our aggregation. In the case of studies related to novices, this category is fortunately confounded with the educational category for which we achieved better agreement. In the case of data aggregation and synthesis many of our disagreements were caused by making different assumptions about what was meant by "analysis" and what was meant by "synthesis". For aggregation purposes we have considered these categories together.

4.5. Quality extraction trends

We observed some differences in the quality scores for different types of study (see Fig. 4). Tertiary studies exhibited the largest quality scores while examples and small experiments exhibited usually relatively low quality scores. Most case studies were high quality but two case studies had relatively low quality scores.

As mentioned in Section 3.6, questions were often inappropriate. The number of inappropriate questions is shown for each question in Fig. 5. Question 7 (Was there a control group or baseline with which to evaluate SR procedures/techniques?) and Question 10 (Has the relationship between researcher and participants been considered to an adequate degree) were the questions that we deemed inappropriate most frequently. Q7 was deemed inappropriate if there were no participants (e.g. the study used results from other studies, or was based on Monte Carlo simulation), or the study was a lessons learnt study where participants and researchers were known to be the same individuals. Q10 was deemed inappropriate on the same basis as Q7.

5. Results

This section discusses each of the papers we included in our study in the context of papers with similar characteristics.

5.1. General lessons learnt and opinion survey papers

We identified eight broad scope lessons learnt and opinion papers reporting nine unique studies (see Table 6). Generally, the papers seemed to be of reasonable quality for the type of papers with the quality score varying from 70% to 100%. However, in many cases (particularly the lessons learnt papers) the assessment was

Table 6
General lesson learnt and opinion survey papers.

Paper	Study	First author	Type of study	Basis of recommendations	Overall quality (% of relevant questions)
P1	S1	Babar	Opinion survey (semi-structured interviews)	Survey of three "leaders", eight "followers" and six "novices". Later extended to include eight more followers and one more novice (reported in P61)	100 × 7/9 = 77.7
P6	S5	Brereton	Lessons learnt	Three SRs (one completed, one in progress, one abandoned)	100 × 6/6 = 100
P23	S20	Dybå	Lessons learnt	One SR	100 × 4/5 = 80
P51	S45	Riaz	Opinion survey	Three novices (each undertaking an SR) plus one expert	100 × 7.5/9 = 83.3
P54	S47	Staples	Lessons learnt	One SR	100 × 6/6 = 100
P58	S50	Turner	Lessons learnt	One large SR	100 × 4.25/6 = 70.8
P61	S52a	Zhang	Tertiary study	Found and assessed 148 SRs	100 × 7.5/8 = 93.7
P61	S52b	Zhang	Opinion survey	52 SR authors and 27 traditional reviewers	100 × 8.25/9 = 91.7
P66	S56	Mian	Lessons learnt	Several SRs	100 × 3.75/5 = 75

made based on a limited number of questions rather than all 10 numerical questions because we judged many of the quality questions were inappropriate in specific circumstances.

Tables 7 and 8 report respectively the problems and advice mentioned in at least two studies. It appears that the three most significant problems are:

1. Digital libraries are not well-suited to complex automated searches (mentioned five times). In addition the lack of standardized keywords was mentioned twice.
2. The time and effort needed for SRs (mentioned four times). In addition the time taken for protocol construction was mentioned twice.
3. The problem of quality assessment of papers based on different research methods (mentioned four times).

We have assessed the importance of a problem or piece of advice in terms of the number of papers that mention it. However, the individual papers may not be completely independent because in the case of P1, the reported opinions came from the authors of some of the lessons learnt studies, for example, as two of the "leaders", Kitchenham and Dybå both contributed to the opinion survey reported in P1 but Kitchenham also contributed to two of the lessons learnt papers (i.e. P6 and P58) and Dybå also contributed to the study reported in P23. Furthermore there may be other overlaps of which we are unaware among the "novices" and "followers".

We have separated the papers into papers published between 2005 and 2007 and papers published in or after 2008, since from 2008 the new version of the guidelines was available. Many of the issues are mentioned in both time periods, but there are several differences:

- Three early papers comment on the criticality of research questions while two later papers comment on the difficulty of defining research questions.
- Two later papers comment on the need for domain knowledge.
- Two early papers mention the need for tools to support SRs. Later papers (and one early paper) emphasize rather that the process is time-consuming which tends to support the need for tools.

The first two differences may be because most of the early papers were written by experienced researchers who addressed issues related to their own topics of interest. In contrast, P1 and P51 include issues raised by novices (i.e. research students), who would not necessarily have had enough domain knowledge to identify specific topics of interest or detailed research questions when they started their studies.

There also appear to be some issues that are particularly problematic for mapping studies as opposed to conventional SRs:

- Using structured questions to construct search strings would not be very helpful for mapping studies that are searching for papers on a specific topic as opposed to a comparison of specific technologies.
- Paper selection is more difficult for mapping studies because it is harder to define inclusion/exclusion criteria for broad topic areas – as we noted in this study it is hard to be certain how best to react to papers that mention a topic issue in passing rather than have the topic of interest as the main focus of the paper.

We only found one example of conflicting advice. Two papers suggested using an extractor and a checker, whereas one paper which used that approach felt it had allowed invalid data collection procedures to go unnoticed.

Table 7 provides a preliminary answer to RQ3 which asked what problems had been observed by SE researchers undertaking SRs while Table 8 addresses RQ4 which asks what advice or techniques have been proposed to address SE problems and the extent of evidence supporting them. In the following sections, we consider the problems and advice presented in other empirically based studies and discussion papers.

5.2. Benefits delivered by SRs

We were interested to identify the extent to which SE research had confirmed that SRs deliver their claimed benefits and whether or not other benefits/advantages had been observed. The general claims for SRs are based on the scientific rigour of the methodology which leads to:

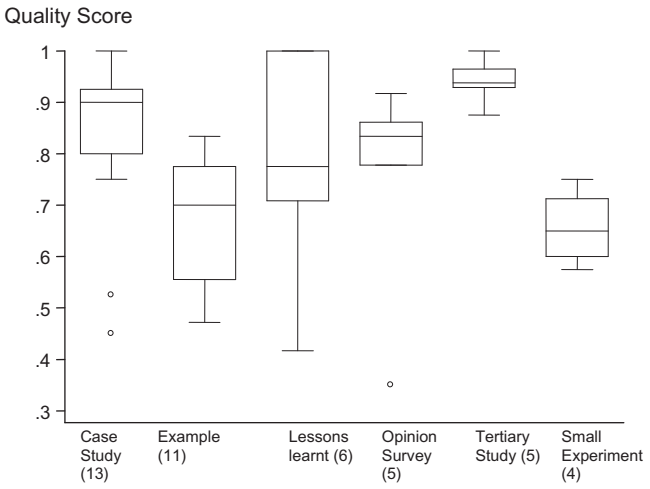

Fig. 4. Quality score for different types of study (number of studies in parenthesis).

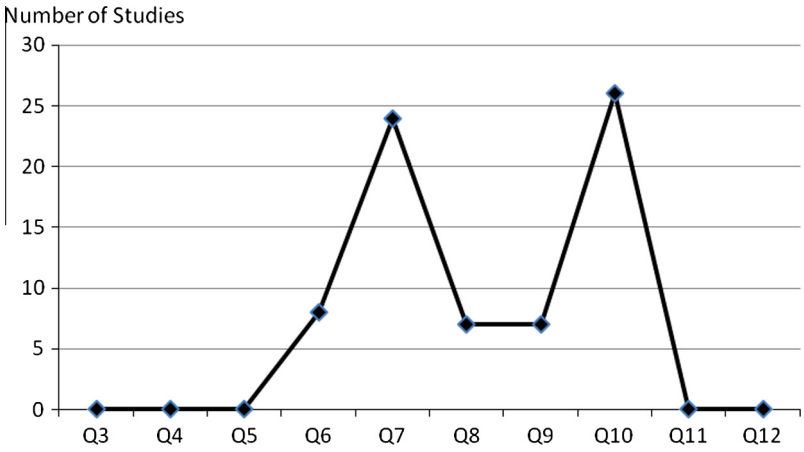

Fig. 5. The number of inappropriate questions per question.

Table 7
Problems identified by lessons learnt and opinion survey papers.

Problem/issue	Mentioned in papers published before 2008	Mentioned in papers published after 2007
Digital library interfaces & functionality inappropriate for SRs	Brereton (P6); Dybå (P23); Staples (P54); Mian (P66)	Babar (P1); Riaz (P51)
Time/effort consuming	Mian (P66)	Babar (P1); Riaz (P51); Zhang (P61)
Protocol will take a long time and/or will be revised	Brereton (P6)	Babar (P1)
IT and software engineering abstracts are poor	Brereton (P6); Dybå (P23)	Riaz (P51)
Qualitative studies complicate SR procedures	Dybå (P23); Brereton (P6);	Babar (P1)
Paper selection/Inclusion exclusion	Staples (P54)	Babar (P1); Riaz (P51)
Defining research questions is difficult		Babar (P1); Riaz (P51)
Quality assessment depends on study type	Brereton (P6); Dybå (P23)	Zhang (P61)
Managing quality evaluation of mixed study types	Dybå (P23)	Riaz (P51)
Data model and data extraction forms may change during extraction	Staples (P54);	Riaz (P51); Turner (P58)
Structured questions not appropriate	Staples (P54)	Riaz (P51)
Space constraints for papers	Brereton (P6)	Riaz (P51)
Choosing appropriate digital libraries	Dybå (P23)	Riaz (P51)
Need domain knowledge		Babar (P1); Riaz (P51)
Papers omit information	Dybå (P23); Staples (P54)	Riaz (P51)
Need tool/methods to support SRs	Staples (P54); Mian (P66)	
SE keywords are not standardized	Dybå (P23)	Mian (P66)

1. Reduction of experimenter bias. That is researchers are encouraged to establish a process by which *all relevant* publications are identified and included in the SR and avoid personal preferences for certain papers or against other papers We do not mean that every study is included in the aggregated data, since researchers may decide to reject low quality papers. However, all papers with high and low quality should be identified, and the rejection of low quality papers should be justified.
2. Increased repeatability/consistency of results. That is researchers from different organisations should get essentially the same results if they address the same research questions. Again aggregations might vary due to issues such as the inclusion or not of low quality studies, but also if researchers make different choices about the digital libraries they search or the time period they include, however, differences should be explicable in terms of the detailed processes used.
3. Auditability. That is SR processes should be fully reported in a clear and understandable manner. Other researchers and readers of an SR report should be in a position to assess the rigour of the SR process used and thus be able to assess the scientific credibility of its results.

Babar and Zhang have considered the value of SRs in papers P1 and P61 which are introduced in Section 5.1. Their results were based on a series of innovative different studies: structured interviews, a tertiary study of existing SE SRs and a survey of authors of SRs and conventional literature reviews. Their tertiary study identified that SRs get more citations than conventional literature reviews. In addition their semi-structured interviews and surveys identified numerous benefits (see Table 9). It was interesting that many of the benefits were mentioned not only by researchers undertaking systematic reviews but also by researchers undertaking conventional reviews. We note that many benefits mentioned by those surveyed were personal benefits. It appears that many researchers believe that SRs are direct benefits to them as researchers which may explain some of the popularity of SRs.

In addition to the personal opinions from broad surveys reported in P1 and P61, other researchers have undertaken empirical studies to investigate SR claims (see Table 10). This table includes papers that have discussed the value of mapping studies as well as studies that have considered conventional SRs. The major difference according to P8 is that mapping studies are intended to "scope" the literature in a topic area and identifies "clusters" of studies suitable for SRs and "gaps" in the current research that suggest the need for more primary studies (i.e. empirical studies investigating the specific methodology). However, mapping studies also have a requirement for rigour, so share many characteristics of conventional SRs.

With respect to conventional SRs, P45 presents an example of a planned study where two research groups undertook independent

Table 8
Advice given by lessons learnt and opinion survey papers.

Advice	Mentioned in papers published before 2008	Mentioned in papers published after 2008
Guidelines work well – so read them	Dybå (P23); Staples (P54)	Babar (P1)
Defining research questions is critical	Brereton (P6); Dybå (P23); Staples (P54)	
Get your protocol validated externally	Brereton (P6)	Babar (P1)
Consult domain expert to help with search strings	Mian (P66)	Riaz (P51)
Do pilot review or mapping study before SR	Brereton (P6); Dybå (23); Mian (P66)	Babar (P1)
Do bookkeeping, record as much as you can during the review	Brereton (P6)	Babar (P1)
You should have good reasons for everything you do, justify your process (particularly the search process)	Brereton (P6)	Babar (P1)
Have one extractor & one checker	Brereton (P6); Staples (P54)	Contrary view – Turner (P58)

Table 9
Benefits/value of SRs.

Benefits/value	Benefit type	Mentioned by
New research findings	Scientific advances	Babar (P1); Zhang (P61)
Learning from studies	Personal	Babar (P1)
Recognition from community	Personal	Babar (P1)
Paper publication	Personal	Babar (P1)
Working experience	Personal	Babar (P1)
Learning research skills	Personal	Babar (P1)
Clear statement and structure of state of the art	Scientific advances	Zhang (P61)
SRs provide a systematic way of building evidence	Methodology	Zhang (P61)
More reliable findings based on synthesis of literature	Methodology	Zhang (P61)
Repeatability	Methodology	Zhang (P61)
Identification of problem areas for new research	Scientific advances	Zhang (P61)
A source for supporting practitioners' decisions about technology selection	Industry	Zhang (P61)

SRs addressing the same research questions. Both groups of researchers were domain experts and experienced researchers. They each identified 10 studies relating to the same topic of which 9 studies were identical. The conclusions they drew from the papers they aggregated were essentially the same. This case study was of high quality and, therefore, provides strong evidence that the SR methodology encourages repeatability although individual studies may exhibit differences. In contrast, P34 reports two SRs addressing the same research issue undertaken independently by two research associates (RAs) which showed no evidence of repeatability. The RAs found different studies between themselves and different studies than those reported in a previous expert literature review. These two results suggest that the extent of repeatability achieved is very dependent on both the domain experience and the research experience of the researchers.

Two papers (P32, and P41) looked at results obtained by different search processes in the context of mapping study trends. In both cases high level trends were quite stable and differences due to the different search process (and thus a different selection of primary studies), were only visible when detailed results were compared. These results suggest that mapping studies can be useful even if incomplete. This interpretation was reinforced by paper P35 which reported a mapping study finding more clusters than an expert review, and missing no clusters. However, the sets of identified clusters were all incomplete when compared with other SRs addressing the specific cluster topics. These results are also supported by paper P36 which found clear benefits from mapping studies in terms of confirming the availability of studies for SRs and providing sets of primary studies suitable for subsequent SRs but warned that mapping studies cannot be guaranteed to be complete and may quickly become out of date.

Finally, P38 investigated the education value of mapping studies. Six students, three undergraduate and three postgraduates, were asked to report their experiences of doing mapping studies. The problems they reported are noted in Section 5.3.1. However, in terms of benefits, they identified that:

- Mapping studies teach students how to search literature and organize the papers found.
- PhD students find a mapping study a valuable means of initiating their research activities.
- Undertaking a mapping study provides students with transferable research skills.

Additionally, some students found the activity challenging, some found it enjoyable and some indicated that it gave them a good overview of the topic area.

These results provide an initial answer to RQ2 which asks to what extent research has confirmed the claims of the SR methodology.

5.3. Main topic areas addressed by studies

In this section we discuss the topics addressed by the remaining 54 studies (i.e. those other than broad lessons leant and opinion surveys). The studies are collated with respect to the main topic(s) addressed as identified by our classification system (see Section 3.6) after aggregating related categories and including a separate discussion of studies recommending the use of textual analysis tools.

5.3.1. Education and novice related papers

In addition to papers P1 and P51 discussed in Section 5.1, another eight studies, concentrated on educational and/or novice related issues (see Table 11). These papers include two papers (i.e. P34 and P38) which have already been discussed from the viewpoint of mapping study benefits in Section 5.2. We have explicitly included these papers in this section because we are considering a different aspect of their results. We observed that two of the papers exhibited an overlap in context – paper P8 discussed three studies also discussed in paper P38. Thus for the purpose of aggregation we excluded paper P8.

P7 was rather different from the other papers. It investigated the extent to which information obtained from SRs could inform Software Engineering teaching. The study found 43 SRs containing information that could influence SE teaching.

Two papers discussed methods of teaching students. P3 reported using outcomes from an existing SR to provide hands-on

Table 10
Studies investigating the value of SRs and mapping studies.

Paper Id	Study Id	First author	Context of results	Study type	Percentage quality score
P8	S7	Budgen	Discussed 6 examples of mapping studies	Informal literature review	100 × 4.75/9 = 52.8
P32	S29	Jalali	Two search processes applied to same RQs	Case study	100 × 90/10 = 90
P34	S31	Kitchenham	Comparison of two RAs doing same SR	Case study	100 × 7.5/10 = 75
P35	S32	Kitchenham	Mapping study and comparison of papers found by SLRs addressing similar questions	Example	100 × 87.5/10 = 87.5
P36	S33	Kitchenham	Four mapping studies used as basis for subsequent work plus 2 external examples	Opinion survey	100 × 8.25/9 = 91.7
P38	S34	Kitchenham	Studies done by 6 students (3 undergrads, 3 PhD)	Opinion survey	100 × 7.75/9 = 86.4
P41	S36a	Kitchenham	Comparison of a broad (automated) and restricted (manual) search.	Case study	100 × 9.25/10 = 92.5
P45	P39	MacDonnell	Two separate research groups doing the same SR	Case study	100 × 8/8 = 100

Table 11
Papers discussing educational and novice-related issues.

Paper Id	Study Id	First author	Context of results	Study type	Percentage quality score	Main results
P3	S3	Baldassarre	Student class (size unspecified)	Opinion survey	100 × 3.5/10 = 35	Students were able to select papers and extract data as part of a hands-on training exercise. Students found the exercises useful and felt the SR method was useful
P5	S4	Brereton	1 MSc student	Case study	100 × 4.5/10 = 45	MSc Student was able to perform SR in restricted timescale but found it difficult. The main problem was study inclusion/exclusion – experts selected far fewer papers and recognized multiple reporting of the same study
P7	S6	Budgen	Reviewed 43 SRs (from a total of 145) that include information useful for teaching	Tertiary study	100 × 7/8 = 87.5	43 of 145 candidate papers included information that would be useful for teaching SE. Coverage of design is patchy and missing for other core areas
P8	S7	Budgen	Discussed 6 studies – 3 undertaken by novices	Informal literature review	100 × 4.75/9 = 52.8	Overlapped with P38
P34	S31	Kitchenham	Two postgraduate research associates	Case study	100 × 7.5/10 = 75	Two research associates were both given the same task (find empirical studies of unit testing). RAs both found different sets of studies which also differed from studies found by an expert review. RAs included papers that experts rejected and vice versa
P38	S34	Kitchenham	Studies done by 6 students (3 undergrads, 3 PhD)	Opinion survey	100 × 7.75/9 = 86.4	All students were able to undertake the SR, but MSc students found restricted timescales challenging
P47	S41	Oates	43 SRs produced by Masters students	Observational study and opinion survey	100 × 7.5/10 = 75	Masters students can do SRs. Students performed less well on aspects of the process relating to article evaluation (both the criteria and the actual evaluation)
P65	S55	Cruzes	7 subjects compared with an expert	Quasi-experiment	100 × 5.75/10 = 57.5	Data extractions by 7 subjects each looking at 3 papers were compared with that of an expert. There was less agreement between subjects and expert that had been hoped. Extracting outcomes was less reliable than context information. Results were better for experiments than case studies

examples for students. P65 described a method for effectively reading and extracting information from papers which aimed to assist novices to identify and extract data required to address SR questions.

Table 12 identifies issues mentioned in the education and novice related papers. These issues were also noted in P1 and P51. Thus, the same issues have been found by different researchers using different empirical methods, so can be regarded as reasonably robust.

Table 12
Common issues.

Issue	Demonstrated by
Novices can do SRs/Mapping Studies	P3, P5, P38, P47 Contrary view: P34
Time and effort required is major problem for undergrads/MSc students.	P5, P38, P47
Paper selection (i.e. inclusion/exclusion) is difficult for novices	P5, P34, P47

5.3.2. Searching and search validation

Table 13 reports papers that suggest means of searching the digital libraries and performing the study selection process. Note more papers related to study selection are discussed later in Section 5.3.3.

Two papers investigated the overlap between search results from different digital libraries. P9 used objective metrics to assess overlap which demonstrated overlaps more clearly than P2. The overlaps found by P9 were as might be expected: General indexing systems overlap while publishers' sites did not overlap. Although the agreement between ACM and IEEE to allow their search engine access to both digital libraries may have changed that observation. P9 pointed out that using indexing systems reduces the need for searching some publishers' sites (e.g. Springer Link and Wiley

Table 13
Studies of the search and selection process.

Paper id	Study id	First author	Approach	Study context	Study type	Percentage quality score
P2	S2	Bailey	Digital library effectiveness	Search process of two different SRs	Example	100 × 4.5/8 = 59.4
P9	S8	Chen	Digital library effectiveness	Search process of two different SRs	Example	100 × 6.5/9 = 72.2
P17	S15	Dieste	Strings to find empirical studies	Re-analysis of a previous SR	Example	100 × 9/9 = 100
P32	S29	Jalali	Citation –based Search (i.e. Snowballing)	Two search processes applied to same RQs	Case study	100 × 9/10 = 90
P33	S30	Kitchenham	Search validation using a reference set of papers	Known set of papers	Example	100 × 7.5/10 = 75
P48	S42	Petersen	Review of current practice	Assessed 139 SRs	Tertiary study	100 × 6.75/7 = 96.4
P53	S46	Skoglund	Citation –based Search (i.e. Snowballing)	Used three previous SRs to illustrate process	Example	100 × 7.5/9 = 83.3
P62	S51	Zhang	Search and validation – using quasi-gold standard	Two SRs	Case study	100 × 0.25/10 = 92.5
P63	S53	Zhang	Search and validation – using quasi-gold standard	Two SRs	Case study	100 × 10/10 = 100

Table 14
Text analysis supporting the SR process.

Paper ID	Study ID	First author	Study type	Percentage quality score
P11	S10	Cruzes	Small example	100 × 0.5/8 = 6.25
P26	S23 (a)	Felizardo	Example	100 × 5/9 = 55.6
P26	S23 (b)	Felizardo	Small experiment	100 × 6.25/10 = 62.5
P27	S24	Felizardo	Example	100 × 7.75/10 = 77.5
P28	S25	Felizardo	Small experiment	100 × 5.75/10 = 57.5
P29	S26	Fernandez-Saez	Discussion paper	N/A
P46	S40	Malheiros	Small experiment	100 × 7.5/10 = 75
P50	S44	Ramampiaro	Discussion paper	N/A
P56	S48	Sun	Small experiment	100 × 6.75/10 = 67.5
P57	S49	Tomassetti	Example	100 × 7.25/10 = 72.5
P68	S58	Torres	Example	100 × 5.5/9 = 61.1

Interscience). This is similar to the point made by Dybå et al. in P23 that they could have saved time and effort for general searches by using ACM, IEEE, plus two indexing systems rather than searching multiple publishers' digital libraries. P9 recommended researchers to report the overlap metrics in their SRs. However, although the metrics are useful when studying the search engines, we are not convinced that the information needs to be reported in every SR.

P17 looked at the problem of finding empirical studies. The authors used Sjøberg et al.'s [33] set of 103 papers as a gold standard to develop strings that would identify experiments and quasi-experiments. They point out the tension between *precision* and *sensitivity* and suggest that using only the term "experiment" achieved good precision and sensitivity. However, they note that terms describing empirical methods are used inconsistently.

Two papers looked at the use of citations analysis (also referred to as snowballing) as a means of identifying primary studies. P32 compared forward snowballing (i.e. finding papers that cited papers found by a search process) and backward snowballing (i.e. looking at the references of the papers found by a search process) and reported that the two techniques gave quite similar results in terms of high level trends and may be more efficient when keywords include general terms such as "agile" that apply to many papers. They recommend using a combination of forward and backward snowballing. In the examples reported in P53, snowballing appeared to work successfully in some cases and not others. P53 also considers the use of critical papers as a starting point for forward citation analysis but the authors did not find this technique very successful for the cases they investigated. In this study, we found automated citation analysis (i.e. forward snowballing) using critical papers to be an effective means of identifying relevant papers.

Two papers proposed an integrated manual and automated strategy (P62, P63). Each of these papers reported a high quality two-case case study. The papers proposed an initial manual search be used to identify a set of known papers. The known papers then act as a *quasi-gold* standard to assist the construction of search strings and assess the quality of the resulting automated search by calculating the quasi-sensitivity of the automated search relative to the known papers. The papers also comment that automated textual analysis of the title, keywords, and abstract of known papers could be used to help construct appropriate search strings. The value of a known set of papers to help determine the search strings for an automated search has been reported in studies by other authors. Kitchenham et al. [12] used the papers found by a manual search to act as a known set of papers to help construct search strings and in P34, Kitchenham et al. used the results of a previous expert review as a set of known studies to assess the completeness of an automated search. In addition, we used a similar approach to assess the effectiveness of the searches performed in this study.

P33 suggested a possible refinement of the quasi-sensitivity concept. The authors suggest that the set of known papers should be split into two sets, and one set be used to construct search strings while the other independent set should be used to evaluate the effectiveness of the search process.

P48 took a rather different approach and reviewed how existing SRs had organized the process of agreeing inclusion/exclusion. This study identified 139 existing SRs in software engineering and identified the actual processes for reaching an inclusion/exclusion (I/E) decision reported in the studies. The strategies included identifying objective criteria for decisions (with the most common being calculating a measure of agreement), strategies for resolving disagreements/uncertainties (with the most common being discussion or adding another reviewer) and decision rules used to arrive at an I/E decision (with the most common being "at least one *uncertain* then include").

5.3.3. Textual mining approaches

Looking at papers selection, classification and data extraction we found a set of 10 papers reporting 11 studies that all proposed textual analysis tools to support the SR process. This is the largest cluster of studies that we found. They are listed in Table 14 which

Table 15
Text analysis supporting the SR process.

Paper ID	Study ID	Approach	Tools used	Study context and main results
P11	S10	Information Extraction (Text Mining)	Site Content Analyser (http://www.sitecontetanalyzer.com)	Discussion containing a small example. The example showed correlated articles had similar word frequency ratings and there were strong relationships between word frequencies and title
P26	S23 (a)	Visual Text Mining (VTM)	ReVis [27]	Re-analyzed a large SR (261 primary studies) and presented various visual displays of study information including citation and content maps to show clusters of similar studies. The VTM analysis found similar clusters to the original study
P26	S23 (b)	VTM	ReVis	4 PhD students: Two used VTM while two read papers to decide whether previous inclusion/exclusion decisions were valid. Results were similar but VTM was faster than reading
P27	S24	VTM	PEx [19]	Re-analyzed a published mapping study to show how visual text mining can use classification data to identify related clusters of studies. All but 2 of the 35 studies were clustered similarly to original paper
P28	S25	VTM	ReVis	4 PhD students: two using VTM diagrams had better performance and more reliable outcomes selecting primary studies compared to the two who read the abstracts
P29	S26	SLR-Tool (incorporating Text Mining)	Apache Lucerne (https://lucene.apache.org)	The tool incorporated textual analysis facilities
P46	S40	VTM	PEx	3 researchers studied 100 articles, two used VTM (B & C), one did not (A). Using an oracle of 40 papers selected by 2 researchers: A found 8.67 articles/h, B & C found 24.49 and 23.53 articles/h, with precision of 82.8% (A), 81.28% (B) and 92% (C)
P50	S44	Meta-Searcher and Automated text retrieval	No specific tools mentioned	Discussed the use of such tools in other disciplines
P56	S48	Ontology with textual analysis	Stamford Parser; SPARQL	An ontology of SRs for cost estimation was constructed. It was used to convert standard abstracts to structured abstracts. The use of the ontology tool to select appropriate papers and extract/aggregate data from those papers was compared with the effectiveness of 4 PhD students. The ontology tool found 11 papers – the students found fewer papers although among all students all papers were found. The students took between 7.5 and 10 h each. The tool also aggregated the data correctly using much less time (12 min compared with 31–39 min)
P57	S49	Linked data approach and text mining	DBpedia (http://dbpedia.org); OpenCalais Web service (http://viewer.opencalais.com). Naïve Bayes tool.	The authors report a process to support the second phase of data selection based on key words and a naïve Bayes classification process. The process was trialed on a part of a large cost estimation SR and reduced the number of papers needing manual review by 20%
P68	S58	Sentence classification	Ibekwe-SanJuan algorithm; Agarwal's algorithm; Teufel's algorithm	Compared three different sentence-classification methods on set of SW testing papers. Results were disappointing although the authors claimed that results of a study using a combined approach not reported in this paper were better

shows the study type and study quality score. In addition, P62 and P63 mention that the approach could be used to assist the construction of search strings. They report that they attempted to use the approach but they do not go into any details. Also, in P23 Dybå et al. reported that they rejected the use of a text analysis tool (NVivo) because of problems converting *pdf* to text.

We summarize the approaches, tools used and the study context and main results in Table 15. Several authors (in particular Cruzes, Maladona and Felizardo) contributed to a number of the studies, so the support for this concept cannot be judged simply by the number of papers mentioning the topic. However, by avoiding duplicate reports, we ensured that the study context of all papers including an evaluation of the proposed technique was different (i.e. different studies used different published SRs as background material or gave subjects different tasks). Thus favourable results from different studies can be assumed to provide independent support for the concept of textual analysis even if the authors overlap.

The general approach of studies proposing the use of text analysis tools is to use a text analysis tool to identify words or phrases that describe individual articles and count the frequency of important words or phrases in each article. Other analysis tools (such as visual display tools) can then be used to identify whether articles that are similar with respect to the frequency of those words or phrases are treated similarly in the SR. This approach can be used:

1. To refine automated search strings, P29.
2. To identify similar papers as part of the paper selection process, P11, P26, P28, P46, P50, P57. These studies offer a different approach to those discussed in Section 5.3.2. They use the tools to investigate whether included and excluded studies are similar with respect to studies' most important keywords and rely on SR researchers to interpret the information provided. The studies discussed in Section 5.3.2 provide a quantitative assessment of the search process effectiveness (although the SR researchers still need to decide whether the achieved effectiveness is sufficient).
3. To categorize and classify articles for a mapping study, P27.
4. To select articles that address a specific research question, P50, P56.
5. To extract the data needed to answer specific research questions P56, P68.

All but one of the 9 empirical studies reported favorable results. The exception (P68) commented that the results obtained for the three sentence classification methods were generally much worse using SE papers than the results reported by the algorithm developers. However, most of the studies were rather limited. Many of the empirical studies were small experiments, restricted examples, or retrospective re-analysis of existing SRs which aimed to demonstrate the feasibility of the approach rather than test the approach.

Table 16
Quality checklists and quality evaluation.

Paper	Study	First author	Study type	Study context	Quality score (percent)
P22	S19	Dieste	Correlation study	Re-analysis of previous two meta-analyses correlating checklist values to a measure of study bias	100 × 6.5/8 = 81.2
P24	S21	Dybå	Discussion	Discussion of using a checklist for qualitative methods for an SR on agile methods and concept of strength of evidence	N/A
P31	S28	Ivarsson	Case study	Application of an industrial relevance checklist in a large SR	100 × 5.25/ 10 = 52.5
P39	S35	Kitchenham	Case study	Tailoring a checklist for an SR from a large set of possible criteria	100 × 8/10 = 80
P41	S36b	Kitchenham	Case study	Comparing different quality assessment processes used in different but related SRs	100 × 8/9 = 88.9
P42	S37	Kitchenham	Limited validation	Developing a quality checklist for testing experiments	100 × 5.5/10 = 55
P44	S38a	Kitchenham.	Observation	Evaluating a quality assessment process in terms of number of assessors	100 × 7.5/9 = 83.3
P44	S38b	Kitchenham	Observation	Evaluating a quality assessment process in terms of impact of assessor discussion	100 × 8.5/9 = 91.7
P44	S38c	Kitchenham	Experiment	Evaluating a quality assessment process in terms of team size	100 × 8.5/10 = 85
P52	S27	Runeson	Discussion	Presentation of a checklist for readers of case studies	N/A

Table 17
Studies investigating data analysis and synthesis.

Paper	Study	First author	Study context	Study type	Quality score (percent)
P10	S9	Cruzes	Re-analysis of an existing literature review to illustrate the use of context variables to cluster studies	Example	100 × 7/10 = 70
P12	S11	Cruzes	Two teams tried two methods of case study aggregation	Example	100 × 4.25/9 = 47.2
P13	S12	Cruzes	Provided guidelines for thematic synthesis	Discussion	NA
P15	S13	Cruzes	Reviewed 49 SRs in terms of aggregation methods used	Tertiary study	100 × 6.5/7 = 92.9
P20	S17	Dieste	Compared four meta-analysis methods with respect to reliability and power	Monte Carlo simulation	100 × 7/8 = 87.5
P21	S18	Dieste	Confirmed that the Q test for heterogeneity is not very powerful	Monte Carlo simulation	100 × 7/8 = 87.5
P67	S57	Mohagheshi	SR based on 8 studies was used to illustrate the use of statistical vote counting	Example	100 × 3.5/7 = 50

To use Wieringa's terminology [38], the current studies are concerned with solution validation not implementation evaluation. Nonetheless some of the retrospective studies were of relatively good quality given their type (i.e. obtained a quality score of more than 70%) but none scored 80% or more.

5.3.4. Quality assessment and checklists

Studies that reported quality checklists and/or attempted to evaluate the quality evaluation process are shown in Table 16. Note we did not attempt to assess the quality of studies that presented a checklist without attempting to validate it.

P22 was an innovative study that attempted to assess the validity of a quality instrument by comparing the score obtained for each study with an objective measure of bias. The measure of bias was obtained by comparing the effect size reported in the paper with the overall effect size reported in a meta-analysis of the papers. They identified only three checklist items correlated with bias (note a negative correlation with bias is equivalent to a positive correlation with quality):

- "Are hypotheses being laid [*sic*] and are they synonymous with the goals discussed before in the introduction?" (Correlation of −0.744 with bias).
- "Does the researcher define the process by which he applies the treatment to objects and subjects (e.g. randomization)?" (Correlation of −0.694 with bias).
- "Are the statistical significances mentioned with the results?" (Correlation of −0.406 with bias).

P24, P31, P42, P52 all propose checklists that can be used to assess the quality of empirical studies. P31 suggests a checklist to determine the industrial relevance of empirical studies which might be of particular significance in the context of EBSE where it is intended that results of SRs should assist practitioners, P52 presents a quality checklist researchers can use to assess case studies. P42 describes the construction of a quality checklist for technology intensive testing experiments and discusses some attempts to validate the checklist. We note that the need for special quality checklists for SE studies applies also to cost estimation studies, usability studies and performance studies and other technology-intensive empirical studies and that none of the checklists from the medical domain are appropriate for these types of SE studies. P24 presents a quality checklist developed for an SR of agile methods. It also makes an important distinction between the quality evaluation of a study and an assessment of the overall strength of evidence associated with a topic of interest when the topic may have been investigated using a variety of different empirical methods.

Kitchenham and colleagues report a series of studies (S36b, S38a, S38b, S38c) that investigated the process by which researchers obtain a consensus about the quality of a paper given a quality checklist. They reported that using two researchers with a period of discussion did not necessarily deliver high reliability (where reliability in this context means consistency in the application of the checklist). They suggest using three or more researchers and taking an average of the "total score", obtained by converting the checklist questions to numerical values. Simple aggregation of scores appeared more efficient (i.e. involved less effort) than incorporating periods of discussion without seriously degrading reliability. In contrast, P22 recommends against using aggregate scores from numerical values of checklist items and recommends only using validated checklist items.

P39 investigated the proposal in the guidelines [16], that checklists could be tailored from a set of checklist items compiled from

existing medical and sociological text books. However, although use of a common set of checklist items lead to a common vocabulary, it was not helpful for novices who intended to develop a checklist for a specific SR. P39 notes that a generic checklist might be a useful starting point for quality checklists for human-based experiments and that researchers should work together to construct appropriate tailored checklists.

5.3.5. Data analysis and synthesis

Studies addressing the problem of data analysis and synthesis are shown in Table 17. P10 suggests the use of contextual information to cluster studies into groups of comparable studies. This is quite a common strategy for aggregating studies in an SR and, for example, in this paper we have grouped some studies according to the SR process they address and others we grouped according to the methodology they used. However, P10 gives a more complex example of using multiple criteria to characterize studies and cluster analysis to identify studies with similar characteristics. They produced a similar grouping to the original researcher.

In P15, Cruzes and Dybå undertook a tertiary study of software engineering SRs (excluding mapping studies) that investigated what types of syntheses methods were being used by SE researchers. They report that half the 49 SRs they reviewed did not contain any formal study synthesis, and of those that did two thirds performed a narrative or thematic synthesis. However, it is worth noting that many of the SRs they analyzed were published before SE researchers became aware of the difference between mapping studies and SRs, so "SR"s lacking synthesis may have been mapping studies that do not synthesize their results in the same way as SRs. Following up the issue of study synthesis, Cruzes and colleagues have provided guidelines for thematic synthesis (P13) and investigated the synthesis of case studies (P12).

The remaining three studies addressed issues related to meta-analysis. P20 used Monte Carlo simulation analysis to compare four meta-analysis methods (Weighted Mean Difference, WMD, Statistical Vote Counting, SVC, Parametric Response Ratio, RR, and Non-Parametric Response Ratio, NPRR) with respect to reliability and power. They suggest software engineers select the method that optimized reliability and power. However, it must be noted that there are other meta-analysis methods not covered by P20, for example using the correlation coefficient [30] or using various measures based on the proportion of variation accounted for by the treatment [25]. Also using Monte Carlo simulation, P21 confirmed that the Q test for heterogeneity is not very powerful. We note that many researchers prefer the I^2 test, although there are also concerns about its power [37]. P67 presents an example based on the SVC approach and points out that it is a useful method of combining empirical results when meta-analysis is not applicable due to small number of studies, diversity of measures and/or limited data on the scale of the effect or its significance.

5.3.6. Miscellaneous

The remaining five studies are reported in Table 18. P16 reports a study that classified the research questions reported in 53 SRs reported by Kitchenham et al. [12]. They found that 63% of research questions were exploratory and only 15% investigated causality. As might be expected 17 of the 18 studies classified as mapping studies reported exploratory studies. However, only 13 of the 32 studies classified as SRs asked causal questions which might mean that some of the SRs were really mapping studies and many mapping studies were published as SRs before software engineering researchers realized the difference between the two types of review.

P19 discusses practical problems experienced updating an SR. This should be compared with P36 which includes a report of our experiences updating our first tertiary study to include a wider search process and a longer time period. The method of aggregation used in the SR being updated by P19 was both novel and complex. In contrast, in P36 we found that updating a simple SR such as a mapping study was not such a major problem. However, we expect the issue of updating SRs to increase in importance as the existing body of SRs in SE increases.

P25 recommends the use of PEx to provide graphical representations of the results of SRs. In an experiment involving 24 participants, 8 participants were given information in graphs, 8 were given information tables and 8 were given information in both tables and graphs. There was no significant difference in comprehensibility; however, in terms of performance/time taken, graphs were the least time-consuming. In our opinion, researchers should use the most appropriate mechanism to answer the research question which in some cases may be graphs and in others tables. However, SRs should always provide full traceability to the source papers.

P49 presents a process model for mapping studies that is much more detailed than the discussion in P8 and demonstrates the value of bubble plots to report mapping study results.

P64 reports the SLuRp tool which can be compared with the SLR-TOOL reported in P29. Both tools aim to address all the SR processes and manage the problems associated with multiple researchers interacting with many primary studies. SLuRp emphasizes the importance of managing large-scale SRs involving a large distributed research team and providing a means of reliably monitoring the progress of the SR.

5.4. Recommendation for changes to the guidelines

In addition to the results discussed in Section 5.3, we looked at several other methods of identifying issues that might require a change to our Guidelines. P1 explicitly reported recommendations for changes to the Guidelines. The researchers taking part in structured interviews made several suggestions for improving the guidelines which, in order of popularity, were:

- More/better quality assessment guidelines (mentioned five times).
- More experiences and examples of good protocols (mentioned four times).
- Simplified "pocket" guide for people reviewing SRs and novices (mentioned four times).
- More references to statistical texts and details about meta-analysis (mentioned twice).

Table 18
Miscellaneous papers.

Paper	Study	First author	Topic	Study context	Study type	Quality score (percent)
P16	S14	da Silva	Research questions	53 SRs	Tertiary study	$100 \times 7/7 = 100$
P19	S16	Dieste	Updating an SR	Updating a complex SR	Lessons learnt	$100 \times 2.5/6 = 42.7$
P25	S22	Felizardo	Graphical reporting	Re-analyzing an existing SR	Experiment	$100 \times 7/10 = 70$
P49	S43	Petersen	Mapping study process	10 Mapping studies + example	Example and literature review	$100 \times 5.5/8 = 68.7$
P64	S54	Bowes	SLR Tool (SLuRp)	Use on a complex SR	Discussion	NA

- More explanation of how to deal with qualitative studies such as case studies (mentioned once).
- Templates for protocols and instructions on how to complete them (allowing for different types of SR) (mentioned once).

Most of these issues can be addressed. Unfortunately, the most requested change is the one for which there is very little practical help.

We also identified issues raised by other studies when we extracted process recommendations (if available) from each study. Some recommendations were already included in the guidelines (e.g. P16 recommended using a reporting standard for SRs but there is already a proposal in the guidelines) and others were merely a statement of the potential value of the proposed method (e.g. P26 S23a. concluded that visual text mining can improve the objectivity of the inclusion/exclusion process). However, we identified some further themes and issues that should be considered in addition to those identified in P1 and in the above discussion of the primary studies in particular:

- Many papers presented recommendations for mapping studies (i.e. P35, P36, P41 S36a, P49).
- Many papers presented recommendations for data synthesis of qualitative study types (i.e. P12, P13, P15, P24).
- Two papers recommended reporting how duplicate studies were managed (i.e. P5 and P35).
- Three papers reported checklists specifically designed to address empirical SE studies (P31, P42, and P53) which could usefully be referenced in the Guidelines.

6. Discussion

6.1. Specific research questions

Our four detailed research questions have been addressed by the results reported in Sections 4 and 5. In summary, RQ1 asked what papers report experiences of using the SR methodology and/or investigate the SR process in software engineering between the years 2005 and 2012 (to June). We found 68 papers discussing issues related to SR methodologies of relevance to our study which discussed 63 unique studies.

This might be regarded as a large number of studies when compared with the number of SRs published in software engineering, for example P7 found 145 SRs up to mid-2011. However, the final step of EBSE (i.e. "Evaluate performance and seek ways to improve it") positively encourages researchers to attempt to improve their process [7]. In addition, when we perform SRs we need to define our research plans in detail in our protocols and document the process in our final report. This emphasis on documenting process plans and outcomes fits well with case study research. Furthermore, the documented outcomes mean that other researchers can easily utilize the outcomes of previous SRs to test out new techniques or procedures. This is indeed what has happened. Many researchers performed case studies of the SR methodology and/or support tools as they undertook their SRs, or used the outcomes of previous SRs as input to their investigations of new approaches.

RQ2 asked to what extent research confirmed the claims of the SR methodology. As might be expected, it is clear that SR claims rely on researchers appropriately using the SR methodology. We are only likely to find reliable, auditable and consistent results when SRs are undertaken by experienced researchers with domain knowledge. However, this leads to a question mark over the results of SRs performed principally by research students. The studies that cover the issue of education confirm that the SR methodology can be used by students but we need to distinguish between undertaking an SR as a training exercise in order to understand the SR process and undertaking an SR as a research goal in its own right. P51 reports that three PhD students took between 8 and 9 months to perform an SR which is similar to a report by one of our students [20]. In spite of complaints that SRs take a long time, 9 months is not unreasonable in the timescale of a PhD. It also provides sufficient time to undertake a high quality SR. However, SRs undertaken by MSc students are usually constrained into a 2–3 month period which is likely to be insufficient both to learn the process and to perform a high-quality study.

Perhaps the most important benefits claimed for SRs were reported in P1 and P61. These are the discovery of new results and a clear structuring of the state of the art. These issues were the most frequently cited motivators for doing SRs by individuals in the structured interviews (7 of 26 and 5 of 26 respectively) and, in addition, 80% of the 52 SR authors responding to a questionnaire reported SRs can unexpectedly bring new research innovation.

Claims for mapping studies relate to their ability to scope the research available in a broad topic area and to identify gaps and clusters in the literature. Overall the evidence supports these claims and suggests that mapping study results in terms of identifying clusters and high level trends are quite resilient to different search processes. However, there is also evidence that mapping studies may miss significant numbers of relevant papers and should not be the basis for SRs without additional more focused searches.

Research question RQ3 asked what problems had been observed by SE researchers when undertaking SRs. A summary of problems and issues can be found in Tables 7 and 12. The evidence suggests that almost every aspect of the SR process has caused problems to some researchers. However, the top three issues appear to be:

1. Digital libraries in SE are not well-suited to complex automated searches.
2. The time and effort needed for SRs.
3. The problem of quality assessment of papers based on different research methods.

Research question RQ4 asked what advice and/or techniques related to performing SR tasks have been proposed and what is the strength of evidence supporting them. A summary of advice can be found in Table 8. A variety of methods and techniques were introduced in Section 5.3 and we discuss them below in the context of the three top SR problems.

The problem with digital libraries is not one that individual researchers can address since the digital libraries are owned and administered by the professional societies and publishers. Possible approaches include:

1. Identifying an appropriate set of libraries to search. Based on current advice, if researchers plan an automated search using search strings (as opposed to a citation analysis methods such as forward snowballing), we recommend searching IEEE, ACM which ensures good coverage of important journals and conferences and at least two general indexing systems such as SCOPUS, EI Compendix or Web of Science (P9, P23).
2. Using the "quasi-gold standard" the search process strategy proposed by Zhang and colleagues (P62 and P63) which is supported by results from two high-quality multi-case case studies and several other studies and provides a useful means of integrating manual and automated searches. Manual searches

should be based mainly on topic specific conferences and journals over a specified time period. However, to act as a quasi-gold standard, it is also useful to include some more general SE journal and conference sources (e.g. IEEE Transactions and the International Conference on Software Engineering). If the sources searched manually are not indexed by the current digital libraries (as was the case of the EASE conference before 2010), they cannot act as gold standard for automated searches.
3. Considering the use of citation analysis (i.e. snowballing) which can be useful in certain circumstances (P53 and this study) although the evidence also confirms that it is sometimes ineffective.

With respect to the time and effort required for SRs there were two proposals for tools to support the SR process as a whole (P29 and P64). In our own experience, it is easy for large SRs with a distributed team to exhibit problems (P58), so we welcome such initiatives. However, the proposed tools need to be evaluated by groups other than those who developed them before they can be unreservedly recommended.

Other researchers have proposed the use of tools (particularly textual analysis tools) to assist specific elements of the SR process (see Table 14). The appeal of textual analysis tools is that scientific articles are textual in nature, so tools that analyse text should be able to assist the SR process. There is substantial evidence of the *feasibility* of using such tools but we need more high quality large-scale studies that consider their impact in practice, highlighting any limitations as well as reporting benefits. In particular, we distrust the idea of automatic extraction of results from primary studies unless our ability to evaluate the quality of different studies improves. Many software engineering studies still use poor or invalid methods, for example, cost estimation researchers have known for many years that the Mean Magnitude Relative Error (MMRE) metric is biased and gives a better value for an estimation method that persistently underestimates than an unbiased estimation method [9,23]. However, MMRE is still used in cost estimation studies. If tools are used to extract data from cost estimation studies, without considering whether the study has used an invalid metric (i.e. without appropriate evaluation of study quality), the extracted results may be obtained very quickly but will be wrong.

Although we would not recommend automatic extraction of results, textual analysis tools can be used in parallel with human intensive methods to evaluate the consistency of the decisions made by the SR team. For example inclusion/exclusion decisions and study classification decisions can be assessed by investigating whether the SR research team have treated similar primary studies in the same way as proposed by P26. We would advise researchers undertaking SRs to trial such tools and report their findings.

With respect to the problem of assessing the quality of primary studies of different types, there has been little progress. Most of the research into quality evaluation has been directed at developing and/or evaluating quality instruments. Only one paper addressed the problem directly. P24 presented the GRADE approach to assess strength of evidence. However, in our opinion, the approach is difficult for experienced researchers, and likely to be infeasible for novice researchers.

6.2. Changes to guidelines

As well as addressing individual research questions, our overall motivation was to assess whether current research supported any changes to current guidelines for SRs in software engineering.

In terms of the primary studies included in this study the following changes would appear to be appropriate:

1. To remove the proposal for constructing structured questions and using them to construct search strings. It does not work for mapping studies and appears to be of limited value to SRs in general since it leads to very complex search strings that need to be adapted for each digital library.
2. To recommend the use of the Quasi-Gold standard approach to integrate manual and automated searches and evaluate the effectiveness of the search process.
3. To recommend that researchers *consider* the use of textual analysis tools to evaluate the consistency of inclusion/exclusion decisions and categorizations.
4. To remove the reference to using a data extractor and a data checker.
5. To include more information about data synthesis issues, particularly the problem of dealing with qualitative methods and studies utilizing mixed methods and provide appropriate references in the guidelines.
6. Either to include more advice on mapping studies or produce a separate set of guidelines for mapping studies.
7. To mention the need to report how duplicate studies are handled.
8. To emphasize the need to keep records of the conduct of the study.
9. To mention the use of citation-based search strategies (i.e. snowballing).
10. To include more examples and advice concerning the construction of protocols.
11. To included references to SE study-specific checklists.

It is also apparent that the discussion of quality checklists in the current guidelines is not useful. It is clear that there is no simple solution to the problem of assessing the quality of empirical studies in SE. We believe that the current unhelpful guidelines should be removed but it is not clear what should replace them. The checklist used in this study is fairly general and we found it possible to apply to the wide range of studies included in this SR. However, we found ourselves forced to assess appropriateness of the checklist items for each study, adding to the complexity of the quality assessment. We also note that applying the quality checklist will not identify invalid empirical practices such as the use of MMRE to compare cost estimation models. The best compromise we can suggest is to:

1. Use a checklist similar to the one proposed in P23 and apply it to all types of empirical study (even if some checklist elements are not applicable to some types of study) but to include consideration of the empirical study type and its size/scope. However, if you are concentrating on only a few different study types, it might be preferable to have tailored checklists for each type. For example, the checklist reported in P23 is not ideally suited for formal experiments, since it does not explicitly consider whether random allocation to treatment took place and whether the allocation to treatment was concealed [36].
2. Ensure that all researchers understand how to apply the quality checklist. Checklists need to be trialed by all researchers and the reasons for disagreements investigated.
3. With two researchers assess quality of primary studies, apply the checklists independently and use discussion to arrive at agreement. With more researchers use three independent assessors and take the mean score. It should also be noted that P22 disputed the value of checklists unless composed of validated items and, in particular, recommended against summing numerical values of checklist elements to form overall scores.

4. Consider the issue of the validity of the empirical methods separately for different types of study.
5. Consider the GRADE method for assessing overall strength of evidence (P24).

However, (apart from step 3) this advice is not supported by empirical evidence nor is it obvious how more empirical evidence could be gathered.

6.3. Limitations

We have already discussed the limitation of our research approach in Section 3.9. The main limitation arising from the conduct of the study is the relatively poor initial agreement we achieved on study quality. We discussed each disagreement until we arrived at a joint evaluation but we must accept that our assessment of a paper's quality score is likely to be rather error prone which in turn impacts the reliability of any assessment of strength of evidence. To address this we have reported not just the quality score but our assessment of the type of validation performed and the context of the validation which provide some additional indication of the stringency of the validation exercise.

Another important limitation of the conduct of our study was that we used the extractor-checker for extracting data from the broad lessons learnt and opinion survey papers. However, we ensured that all the information extracted from these papers was reported in the words of the authors of the papers and was linked back to the specific point in the paper where the issue was mentioned. We also used an analyst-checker process to integrate the results from these different papers. This was done because we were unsure initially how to manage the aggregation and synthesis process which meant that the approach could not be specified prior to undertaking it. Thus, we have increased the risk of missing some important issues, or misinterpreting issues that we found, compared with a study where all data extraction and aggregation was undertaken independently and then integrated.

7. Conclusions

This systematic mapping study has discussed 68 software engineering research papers reporting 63 unique primary studies addressing problems associated with SRs, advice on how to perform SRs, and proposals to improve the SR process. These studies have identified a number of common problems experienced by SE researchers undertaking SRs and various proposals to address these problems. We have identified numerous improvements that should be made to the SR guidelines [16], in particular, we believe that the current guidelines should be amended to remove unhelpful suggestions with respect to structured questions and search string construction and construction of quality checklists. They should also be changed to include recommendations related to using a quasi-gold standard and optional use of textual analysis tools. In addition, some changes must be made to advice related to quality checklists but it is not possible to avoid the inherent difficulty associated with quality assessment.

We believe that further research is required in several areas:

- The development and evaluation of tools to manage the SR process.
- The evaluation of textual analysis tools in prospective case studies (rather than post-hoc examples) and large scale experiments.
- Procedures for quality evaluation of SE papers when the primary studies have used a variety of different empirical methods.

Appendix A. Format of form for extracting lessons learnt and opinion survey textual data

Paper title:
Paper ID:
Study ID:
Extractor

Issue Id	Issue text	Type	Suggestion for guidelines Yes/No	Novice issues Yes/No	Education issues Yes/No	Position in Paper	Stage in SR Process addressed	Importance (either text or number of "votes")	Related Issue	Comment
	For each issue/ problem raised/ problem solution proposed specify the issue/ problem using the same text as the papers authors	Advice (including Best practice) Problem (including Challenge) Value (Benefit)			Education (including training, gaining experience)	Page number or Table number or Id	Research question/ Protocol/ Search/ Selection/ Data extraction/ Quality Assessment/ Data Aggregation/ Data Synthesis/ Reporting	A ratio indicating number of votes out of possible votes. Or an textual indication of relative importance	Reference to any related issue	

Appendix B. Papers excluded from the SR during data extraction

Authors	Title	Source	Reason
Boell S.K., Cecez-Kecmanovic D.	Literature reviews and the hermeneutic circle	Australian Academic and Research Libraries	General critique of SRs. Not SE oriented
Brereton P.	A study of computing undergraduates undertaking a systematic literature review	IEEE Transactions on Education	SLR was not a software engineering topic
Budgen D., Bailey J., Turner M., Kitchenham B., Brereton P., Charters S.	Cross-domain investigation of empirical practices	IET Software, 2009	More related to primary studies than SRs
Budgen, D., John Bailey, Mark Turner, Barbara Kitchenham, Pearl Brereton, Stuart Charters	Lessons from a cross domain investigation of empirical practices	EASE 2008	Preliminary version of Budgen et al., 2009, so also rejected
de Almeida Biolchini, Jorge Calmon, Paula Gomes Mian, Ana Candida Cruz Natali, Tayana Uchoa Conte, Guilherme Horta Travassos	Scientific research ontology to support systematic review in software engineering	Advanced Engineering Informatics	No clear implications for SR processes
Jorgensen M., Dyba T., Kitchenham B.	Teaching evidence-based software engineering to university students	2005 Proceedings – International Software Metrics Symposium	More related to EBSE than SRs
Nakagawa E.Y., Feitosa D., Felizardo K.R.	Using systematic mapping to explore software architecture knowledge	ICSE	Just a straightforward mapping study
MacDonnell, S.G. and M.J. Shepperd	Comparing Local and Global Software Effort Estimation Models – Reflections on a Systematic Review	ESEM 2007	Failed inclusion criteria. Primarily an SR not aimed at investigating SR process issues
Major L., Kyriacou T., Brereton O.P.	Systematic literature review: Teaching novices programming using robots	IET Seminar Digest, 2011	Failed inclusion criteria. Primarily an SR not aimed at investigating SR process issues
Ramey J., Rao P.G.	The systematic literature review as a research genre	IEEE International Professional Communication Conference	General discussion. Not SE oriented

Appendix C. Selected papers (rows in italics identify duplicate reports)

Paper Number	Study Number	Authors	Year	Title	Source
P1	S1	Babar, Muhammad Ali, He Zang	2009	Systematic literature reviews in software engineering: Preliminary results from interviews with researchers	International Symposium on Empirical Software Engineering and Measurement (ESEM)
P2	S2	Bailey J., Zhang C., Budgen D., Turner M., Charters S.	2007	Search engine overlaps: Do they agree or disagree?	Proceedings – ICSE 2007 Workshops: Second International Workshop on Realizing Evidence-Based Software Engineering, REBSE'07
P3	S3	Baldassarre, M.T., Nicola Boffoli, Danilo Caivano and Giuseppe Visaggio	2008	A Hands-On Approach for Teaching Systematic Review	PROFES Lecture Notes in Computer Science, 2008, Volume 5089/2008, 415–426, DOI: 10.1007/978-3-540-69566-0_33
P4	*S3*	*Baldassarre, M.T., Danilo Caivano, Barbara Kitchenham & Giuseppe Visaggio*	*2007*	*Systematic Review of Statistical Process Control: An Experience Report*	*Evaluation and Assessment in Software Engineering (EASE)*
P5	S4	Brereton P., Turner M., Kaur R.	2009	Pair programming as a teaching tool: a student review of empirical studies	Proceedings – 22nd Conference on Software Engineering Education and Training, CSEET 2009

Appendix C (*continued*)

Paper Number	Study Number	Authors	Year	Title	Source
P6	S5	Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M., Khalil, M.	2007	Lessons from applying the systematic literature review process within the software engineering domain	Journal of Systems and Software (JSS), 80 (4), 571–583
P7	S6	Budgen, D., Drummond, S., Brereton, P. and Holland, N.	2012	What Scope is there for Adopting Evidence-Informed teaching in SE	International Conference on Software Engineering (ICSE)
P8	S7	Budgen, D.; Turner, M.; Brereton, P. and Kitchenham, B.	2008	Using Mapping Studies in Software Engineering.	Proc. Of PPIG'08, Lancaster University, UK, pp. 195–204
P9	S8	Chen, Lianipng, Muhammad Ali Babar and He Zhang	2010	Towards an Evidence-Based Understanding of Electronic Data Sources	EASE
P10	S9	Cruzes D., Mendonca M., Basili V., Shull F., Jino M.	2007	Using context distance measurement to analyze results across studies	ESEM
P11	S10	Cruzes, D., Mendonça, M., Basili, V., Shull, F., Jino, M.	2007	Automated Information Extraction from Empirical Software Engineering: Is that possible?	ESEM
P12	S11	Cruzes, D.S., Dybå, T., Runeson, P., Höst, M.	2011	Case studies synthesis: Brief experience and challenges for the future	ESEM
P13	S12	Cruzes, D.S., Tore Dybå	2011	Recommended Steps for Thematic Synthesis in Software Engineering	ESEM
P14	*S13*	*Cruzes, Daniela, Tore Dybå*	*2010*	*Synthesizing evidence in software engineering research*	*ESEM*
P15	S13	Cruzes, Daniela, Tore Dybå	2011	Research synthesis in software engineering: A tertiary study	IST, 53 (5), 440–455
P16	S14	da Silva, Fabio Q.B., André L.M. Santos, Sérgio C.B. Soares, A. César C. França and Cleviton V.F. Monteiro.	2010	A Critical Appraisal of Systematic Reviews in Software Engineering from the Perspective of the Research Questions Asked in the Reviews	ESEM
P17	S15	Dieste O., Griman A., Juristo N.	2009	Developing search strategies for detecting relevant experiments	Empirical Software Engineering 14, 513–539
P18	*S15*	*Dieste, O. and Padua, A.G.*	*2007*	*Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews*	*ESEM*
P19	S16	Dieste O., Lopez M., Ramos F.	2008	Formalizing a systematic review updating process	Proceedings – 6th ACIS International Conference on Software Engineering Research, Management and Applications, SERA 2008
P20	S17	Dieste, O., Enrique Fernández, Ramón Garcia Martinez and Natalia Juristo	2011	Comparative Analysis of Meta-Analysis Methods: When to use Which?	ESEM
P21	S18	Dieste, O., Enrique Fernández, Ramón García-Martínez, Natalia Juristo	2011	The risk of using the Q heterogeneity estimator for software engineering experiments	EASE
P22	S19	Dieste, O., Grimán, A., Juristo, N. and Saxena, H.	2011	Quantitative determination of the relationship between internal validity and bias in software engineering: consequences for systematic literature reviews	ESEM
P23	S20	Dybå, T., Dingsøyr, T., G.K. Hanssen	2007	Applying systematic reviews to diverse study types: an experience report	ESEM

(*continued on next page*)

Appendix C (*continued*)

Paper Number	Study Number	Authors	Year	Title	Source
P24	S21	Dybå, T., Torgeir Dingsøyr	2008	Strength of evidence in systematic reviews in software engineering	ESEM
P25	S22	Felizardo K.R., Riaz M., Sulayman M., Mendes E., MacDonell S.G., Maldonado J.C.	2011	Analyzing the use of graphs to represent the results of systematic reviews in software engineering	Proceedings – 25th Brazilian Symposium on Software Engineering, SBES 2011
P26	S23 (a,b)	Felizardo, K.R., Andery, G.F., Paulovich, F.V., Minghim, R., Maldonado, J.C.	2012	A visual analysis approach to validate the selection review of primary studies in systematic reviews	IST, 54 (10), 1079–1091
P27	S24	Felizardo, Katia Romera, Elisa Yumi Nakagawa, Daniel Feitosa, Rosane Minghim and José Carlos Maldonado	2010	An Approach Based on Visual Text Mining to Support Categorization and Classification in the Systematic Mapping	EASE
P28	S25	Felizardo, Katia Romero; Norsaremah Salleh, Rafael Messias Martins, Emilia Mendes, Stephen G. Macdonell and José Carlos Maldonado	2011	Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews	ESEM
P29	S26	Fernandez-Saez A.M., Bocco M.G., Romero F.P.	2010	SLR-Tool a tool for performing systematic literature reviews	ICSOFT 2010 – Proceedings of the 5th International Conference on Software and Data Technologies
P30	*S27*	*Höst, M. and P. Runeson*	*2007*	*Checklists for Software Engineering Case Study Research.*	*ESEM*
P31	S28	Ivarsson M., Gorschek T.	2011	A method for evaluating rigor and industrial relevance of technology evaluations	ESE
P32	S29	Jalali, E. and Wohlin, Claes	2012	Systematic Literature Studies: Database Searches vs. Backward Snowballing	ESEM
P33	S30	Kitchenham, B.A. Li, Z., Burn, A.J.	2011	Validating Search Process in Systematic Literature Reviews	EAST
P34	S31	Kitchenham B., Pearl Brereton, Zhi Li, David Budgen & Andrew Burn	2011	Repeatability of Systematic Literature Reviews	EASE
P35	S32	Kitchenham, B., Pearl Brereton and David Budgen	2012	Mapping study completeness and reliability – a case study	EASE
P36	S33	Kitchenham, B.A., Budgen, D., Pearl Brereton, O.	2011	Using mapping studies as the basis for further research – A participant-observer case study	IST, 53 (6), 638–651
P37	*S33*	*Kitchenham, Barbara A., David Budgen and O. Pearl Brereton*	*2010*	*The value of mapping studies – A participant-observer case study*	*EASE*
P38	S34	Kitchenham, B., Pearl Brereton, David Budgen	2010	The educational value of mapping studies of software engineering literature	ICSE
P39	S35	Kitchenham, B., Pearl Brereton, David Budgen, Zhi Li	2009	An Evaluation of Quality Checklist Proposals – A participant-observer cases study	EASE
P40	*S36*	*Kitchenham, B., Pearl Brereton, Mark Turner, Mahmood Niazi, Stephen G. Linkman, Rialette Pretorius, David Budgen*	*2009*	*The impact of limited search procedures for systematic literature reviews A participant-observer case study.*	*ESEM*
P41	S36 (a,b)	Kitchenham, B.A., Brereton, P., Turner, M., Niazi, M.K., Linkman, S., Pretorius, R., Budgen, D.	2010	Refining the systematic literature review process-two participant-observer case studies	Empirical Software Engineering 15, 618–653
P42	S37	Kitchenham, B.A., Andrew J. Burn, Zhi Li	2009	A Quality Checklist for Technology-Centred Testing Studies	EASE

Appendix C (*continued*)

Paper Number	Study Number	Authors	Year	Title	Source
P43	*S38 (a,b)*	*Kitchenham, B.A., Sjoberg, D.I.K., Brereton, P., Budgen, D., Dyba, T., Host, M., Pfahl, D., Runeson, P.*	*2010*	*Can we evaluate the quality of software engineering experiments?*	*ESEM*
P44	S38 (a,b,c)	Kitchenham, B.A., Sjoberg, D.I.K., Dyba, T., Pfahl, D., Brereton, P., Budgen, D., Host, M., Runeson, P.	2012	Three empirical studies on the agreement of reviewers about the quality of software engineering experiments	IST, 54 (8), 804–819
P45	S39	MacDonell, S., Shepperd, M., Kitchenham, B., Mendes, E.	2010	How reliable are systematic reviews in empirical software engineering?	IEEE Transactions on Software Engineering (TSE), 36 (5), 676–687
P46	S40	Malheiros, Viviane, Erika Hohn, Roberto Pinho, Manoel Mendonca, Jose Carlos Maldonado	2007	A Visual Text Mining approach for Systematic Reviews	ESEM
P47	S41	Oates, Briony J., Graham Capper	2009	Using systematic reviews and evidence-based software engineering with masters students	EASE
P48	S42	Petersen, K., Ali, N.B.	2011	Identifying strategies for study selection in systematic reviews and maps	ESEM
P49	S43	Petersen, K.; Feldt, R.; Shahid, M. and Mattsson, M.	2008	Systematic Mapping Studies in Software Engineering.	EASE
P50	S44	Ramampiaro H., Cruzes D., Conradi R., Mendona M.	2010	Supporting evidence-based Software Engineering with collaborative information retrieval	Proceedings of the 6th International Conference on Collaborative Computing: Networking, Applications and Worksharing, CollaborateCom 2010
P51	S45	Riaz, Mehwish,; Muhammad Sulayman, Norsaremah Salleh and Emilia Mendes	2010	Experiences Conducting Systematic Reviews from Novices' Perspective	EASE
P52	S27	Runeson, P and Höst, M.	2009	Guidelines for conducting and reporting case study research in software engineering	Empirical Software Engineering 14, 131–164
P53	S46	Skoglund, Mats and Per Runeson	2009	Reference-based search strategies in systematic reviews	EASE
P54	S47	Staples, M., Niazi, M.	2007	Experiences using systematic review guidelines	JSS, 80 (9), 1426–1437
P55	*S47*	*Staples, Mark & Mahmood Niazi*	*2006*	*Experiences Using Systematic Review Guidelines*	*EASE*
P56	S48	Sun, Yueming, Ye Yang, He Zhang, Wen Zhang, Qing Wang	2012	Towards Evidence-Based Ontology for Supporting Systematic Literature Review	EASE
P57	S49	Tomassetti, Federico,; Giuseppe Rizzo, Antonio Vetro', Luca Ardito, Marco Torchiano & Maurizio Morisio	2011	Linked Data Approach for Selection Process Automation in Systematic Reviews	EASE
P58	S50	Turner, M., Barbara Kitchenham, Pearl Brereton, David Budgen	2008	Lessons learnt Undertaking a Large-scale Systematic Literature Review	EASE
P59	*S51*	*Zhang, He and Muhammad Ali Babar*	*2010*	*On Searching Relevant Studies in Software Engineering*	*EASE*
P60	*S52 (a,b)*	*Zhang, He, Muhammad Ali Babar*	*2011*	*An Empirical Investigation of Systematic Reviews in Software Engineering*	*ESEM*
P61	S52 (a,b)	Zhang, He, Muhammad Ali Babar	2013	Systematic Reviews in Software Engineering: An Empirical Investigation	IST, 55 (7), 1341–1354

(*continued on next page*)

Appendix C (*continued*)

Paper Number	Study Number	Authors	Year	Title	Source
P62	S51	Zhang, He, Muhammad Ali Babar, Paolo Tell	2011	Identifying relevant studies in software engineering	IST, 53 (6), 626–637
P63	S53	Zhang, He, Muhammad Ali Babar, Xu Bai, Juan Li, Huang, Liguo	2011	An Empirical Assessment of A Systematic Search Process for Systematic Reviews	EASE
P64	S54	Bowes, David, Hall, Tracy and Beecham, Sarah	2012	SLuRp – A tool to help large complex systematic literature reviews deliver valid and rigorous results	Evidential Assessment of Software Technologies (EAST)
P65	S55	Cruzes, D. Mendonca, M., Basili, V., Shull, F. and Jino, N.	2007	Extracting information from Experimental Software Engineering papers	International Conference of the Chilean Society of Computer Science, SCCC '07
P66	S56	Mian, P., T. Conte, A. Natali, J. Biolchini, E. Mendes, G. Travassos	2005	Lessons learned on applying systematic reviews to software engineering	Proceedings of the 2nd Experimental Software Engineering Latin American Workshop (ESELAW'05), Brazil
P67	S57	Mohagheshi, P., Conradi, R.	2006	Vote counting for combining quantitative evidence from empirical studies – an example.	Proc ISESE '06, pp. 24-2
P68	S58	Torres, José Alberto S., Cruzes, Daniela and Salvador, Laís do Nascimento	2012	Automatic Results Identification in Software Engineering Papers. Is it possible?	12th International Conference of Computational science and Its Applications

References

[1] Pearl Brereton, Barbara A. Kitchenham, David Budgen, Mark Turner, Mohamed Khalil, Lessons from applying the systematic literature review process within the software engineering domain, Journal of Systems and Software 80 (4) (2007) 571–583.

[2] J. Biolchini, P.G. Mian, A.C.C. Natali, G.T. Travassos, Systematic Review in Software Engineering. Technical Report RT-ES 679-05. PESC, COPPE/UFRJ, 2005.

[3] D. Budgen, M. Turner, P. Brereton, B. Kitchenham, Using mapping studies in software engineering, in: Proc. of PPIG'08. Lancaster University, UK, 2008, pp. 195–204.

[4] Fabio Q.B. da Silva, André L.M. Santos, Sérgio Soares, A. CésarC. França, Cleviton V.F. Monteiro, Felipe Farias Maciel, Six years of systematic literature reviews in software engineering: an updated tertiary study, Information and Software Technology 53 (9) (2011) 899–913.

[5] O. Dieste, A.G. Padua, Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews. ESEM 2007, 2007.

[6] T. Dybå, T. Dingsøyr, Empirical studies of agile software development: a systematic review, Information and Software Technology 50 (2007) 833–859.

[7] Tore Dybå, Barbara Kitchenham, Magne Jørgensen, Evidence-based software engineering for practitioners, IEEE Software 22 (1) (2005) 58–65.

[8] K. Felizardo, S. MacDonell, E. Mendes, J.C. Maldonado, A systematic mapping on the use of visual data mining to support the conduct of systematic literature reviews, Journal of Systems and Software 7 (2) (2012) 450–461.

[9] T. Foss, E. Stensrud, B. Kitchenham, I. Myrtveit, A Simulation study of the model evaluation criteria MMRE, IEEE Transactions on Software Engineering 29 (11) (2003) 985–995.

[10] A. Jedlitschka, M. Ciolkowski, D. Pfahl, Reporting experiments in software engineering, in: F. Shull, J. Singer, D. Sjøberg (Eds.), Guide to Advances Empirical Software Engineering, Springer, 2008.

[11] B. Kitchenham, Pearl Brereton, Zhi Li, David Budgen, Andrew Burn, Repeatability of Systematic Literature Reviews. EASE 2011, 2011.

[12] Barbara Kitchenham, Rialette Pretorius, David Budgen, Pearl Brereton, Mark Turner, Mahmood Niazi, Stephen G. Linkman, Systematic literature reviews in software engineering – a tertiary study, Information & Software Technology 52 (8) (2010) 792–805.

[13] B.A. Kitchenham, D.I.K. Sjoberg, P. Brereton, D. Budgen, T. Dyba, M. Host, D. Pfahl, P. Runeson, Can We Evaluate the Quality of Software Engineering Experiments? ESEM, 2010.

[14] B. Kitchenham, Pearl Brereton, David Budgen, The Educational Value of Mapping Studies of Software Engineering Literature. ICSE 2010, 2010.

[15] Barbara Kitchenham, Pearl Brereton, David Budgen, Mark Turner, John Bailey, Stephen G. Linkman, Systematic literature reviews in software engineering – a systematic literature review, Information & Software Technology 51 (1) (2009) 7–15.

[16] B.A. Kitchenham, S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE-2007-01, School of Computer Science and Mathematics, Keele University, 2007.

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

[18] B. Kitchenham, Tore Dybå, Magne 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. ISBN 0-7695-2163-0.

[19] A.A. Lopes, P. Pinho, F.V. Paulovich, R. Minhim, Visual text mining using association rules, Computers & Graphics 31 (3) (2007) 316–326.

[20] L. Major, T. Kyriacou, O.P. Brereton, Systematic literature review: teaching novices programming using robots, IET Seminar Digest, 2011.

[21] S. MacDonell, M. Shepperd, B. Kitchenham, E. Mendes, How reliable are systematic reviews in empirical software engineering?, IEEE TSE 36 (5) (2010) 676–687

[22] C. Marshall, P. Brereton, Tools to Support Systematic Literature Review in Software Engineering: A Mapping Study. Unpublished Manuscript, 2013.

[23] I. Myrtveit, E. Stensrud, Validity and reliability of evaluation procedures in comparative studies of effort prediction, Empirical Software Engineering 17 (1–2) (2012) 18–22.

[24] G.W. Noblit, R.D. Hare, Meta-Ethnography: Synthesizing Qualitative Studies, Sage, Thousand Oaks, 1988.

[25] S. Olejnik, J. Algina, Generalized eta and omega squared statistic: measures of effect size for some common research designs, Psychological Methods 8 (4) (2003) 434–447.

[26] K. Petersen, R. Feldt, M. Shahid, M. Mattsson, Systematic Mapping Studies in Software Engineering. EASE 2008, 2008.

[27] A.J. Quigley, M. Postema, H. Schmidt, ReVis: Reverse Engineering by Clustering and Visual Object Classification, in: Proceedings Australian Software Engineering Conference, 2000. IEEE, 2000, pp. 119–125.

[28] A. Rainer, S. Beecham, A Follow-Up Empirical Evaluation of Evidence Based Software Engineering by Undergraduate Students. EASE, 2008.

[29] A. Rainer, T. Hall, N. Baddoo, A Preliminary Empirical Investigation of the Use of Evidence Based Software Engineering by Under-Graduate Student. EASE, 2006.

[30] R. Rosenthal, M.R. DiMatteo, Meta-analysis: recent developments in quantitative methods for literature reviews, Annual Review of Psychology 52 (2001) 59–82.

[31] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, Empirical Software Engineering 14 (2009) 131–164.

[32] Mats Skoglund, Per Runeson, Reference-Based Search Strategies in Systematic Reviews. EASE 2009, 2009.

[33] 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 Software Engineering 31 (9) (2005) 733–753.
[34] M. Staples, M. Niazi, Experiences using systematic review guidelines, Journal of Systems and Software 80 (9) (2007) 1425–1437.
[35] Mark Staples, Mahmood Niazi, Experiences Using Systematic Review Guidelines. EASE, 2006.
[36] K. Schulz, I. Chalmer, R.J. Hayes, D.G. Altman, Empirical evidence of bias dimensions of methodological quality associated with estimates of treatment effects in controlled trials, JAMA 273 (1995) 408–412.
[37] K. Thorlund, G. Imberger, B.C. Johnston, M. Walsh, T. Awad, et al., Evolution of heterogeneity (I^2) estimates and their 95% confidence intervals in large meta-analyses, PLoS ONE 7 (7) (2012) e39471, http://dx.doi.org/10.1371/journal.pone.0039471.
[38] R. Wieringa, N. Maiden, N. Mead, C. Rolland, Requirements engineering paper classification and evaluation criteria: a proposal and a discussion, Requirements Engineering 11 (2006) 102–107.
[39] P. Woodall, P. Brereton, Conducting a Systematic Literature Review from the Perspective of a Ph.D. Researcher. EASE, 2006.
[40] June M. Verner, Jennifer Sampson, Vladimir Tosic, Nur Azzah Abu Bakar, Barbara Kitchenham, Guidelines for industrially-based multiple case studies in software engineering, RCIS 2009 (2009) 313–324.