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 strings 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.

© 2013 Elsevier B.V. All rights reserved.
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.
- 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.](image-url)
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:

- Information and Software Technology (IST): 6 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:

- 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 Dyba and Dingsøyr [6] since it is applicable to most types of study but unlike them we did not reject 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.

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).
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 spreadsheet 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 spreadsheet 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 reviews 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 $100 \times \frac{45}{410} = 11\%$.

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 $100 \times \frac{54}{3360} = 1.6\%$.

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

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 10

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

### Table 11

<table>
<thead>
<tr>
<th>Paper Id</th>
<th>Study Id</th>
<th>First author</th>
<th>Context of results</th>
<th>Study type</th>
<th>Percentage quality score</th>
<th>Main results</th>
</tr>
</thead>
<tbody>
<tr>
<td>P8</td>
<td>S7</td>
<td>Budgen</td>
<td>Discussed 6 examples of mapping studies</td>
<td>Informal literature review</td>
<td>100 × 4.75/9 = 52.8</td>
<td>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.</td>
</tr>
<tr>
<td>P32</td>
<td>S29</td>
<td>Jalali</td>
<td>Two search processes applied to same RQs</td>
<td>Case study</td>
<td>100 × 90/10 = 90</td>
<td>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.</td>
</tr>
<tr>
<td>P34</td>
<td>S31</td>
<td>Kitchenham</td>
<td>Comparison of two RAs doing same SR</td>
<td>Case study</td>
<td>100 × 7.5/10 = 75</td>
<td>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.</td>
</tr>
<tr>
<td>P35</td>
<td>S32</td>
<td>Kitchenham</td>
<td>Mapping study and comparison of papers found by SLRs addressing similar questions</td>
<td>Example</td>
<td>100 × 87.5/10 = 87.5</td>
<td>Overlapped with P38</td>
</tr>
<tr>
<td>P7</td>
<td>S6</td>
<td>Budgen</td>
<td>Reviewed 43 SRs (from a total of 145) that include information useful for teaching</td>
<td>Tertiary study</td>
<td>100 × 7/8 = 87.5</td>
<td>3 of 43 SRs included information that would be useful for teaching SE. Coverage of design is patchy and missing for other core areas.</td>
</tr>
<tr>
<td>P8</td>
<td>S7</td>
<td>Budgen</td>
<td>Discussed 6 studies – 3 undertaken by novices</td>
<td>Informal literature review</td>
<td>100 × 4.75/9 = 52.8</td>
<td>Overlapped with P38</td>
</tr>
<tr>
<td>P34</td>
<td>S31</td>
<td>Kitchenham</td>
<td>Two postgraduate research associates</td>
<td>Case study</td>
<td>100 × 7.5/10 = 75</td>
<td>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.</td>
</tr>
<tr>
<td>P38</td>
<td>S34</td>
<td>Kitchenham</td>
<td>Studies done by 6 students (3 undergrads, 3 PhD)</td>
<td>Opinion survey</td>
<td>100 × 7.5/9 = 86.4</td>
<td>All students were able to undertake the SR, but MSc students found restricted timescales challenging.</td>
</tr>
<tr>
<td>P45</td>
<td>S39</td>
<td>MacDonnell</td>
<td>Two separate research groups doing the same SR</td>
<td>Case study</td>
<td>100 × 8/8 = 100</td>
<td>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.</td>
</tr>
<tr>
<td>P65</td>
<td>S55</td>
<td>Cruzes</td>
<td>7 subjects compared with an expert</td>
<td>Quasi-experiment</td>
<td>100 × 5.75/10 = 57.5</td>
<td>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.</td>
</tr>
</tbody>
</table>
Table 13
Studies of the search and selection process.

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

Table 14
Text analysis supporting the SR process.

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

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 P33, snowballing appeared to work successfully in some cases and not others. P33 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.

<table>
<thead>
<tr>
<th>Paper ID</th>
<th>Study ID</th>
<th>Approach</th>
<th>Tools used</th>
<th>Study context and main results</th>
</tr>
</thead>
<tbody>
<tr>
<td>P11</td>
<td>S10</td>
<td>Information Extraction ([Text Mining])</td>
<td>Site Content Analyser (<a href="http://www.siticontentanalyzer.com">http://www.siticontentanalyzer.com</a>)</td>
<td>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.</td>
</tr>
<tr>
<td>P26</td>
<td>S23</td>
<td>Visual Text Mining ([VTM])</td>
<td>ReVis [27]</td>
<td>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.</td>
</tr>
<tr>
<td>P27</td>
<td>S24</td>
<td>VTM</td>
<td>PEX [19]</td>
<td>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.</td>
</tr>
<tr>
<td>P46</td>
<td>S40</td>
<td>VTM</td>
<td>PEx</td>
<td>Using an oracle of 40 papers selected by 2 researchers: A found 8.67 articles/h, B &amp; C found 24.49 and 23.53 articles/h, with precision of 82.8% (A), 81.28% (B) and 92% (C).</td>
</tr>
<tr>
<td>P50</td>
<td>S44</td>
<td>Meta-Searcher and Automated text retrieval</td>
<td>No specific tools mentioned</td>
<td>Discussed the use of such tools in other disciplines.</td>
</tr>
<tr>
<td>P56</td>
<td>S48</td>
<td>Ontology with textual analysis</td>
<td>Stanford Parser; SPARQL</td>
<td>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).</td>
</tr>
<tr>
<td>P57</td>
<td>S49</td>
<td>Linked data approach and text mining</td>
<td>DBpedia (<a href="http://dbpedia.org">http://dbpedia.org</a>); OpenCalais Web service (<a href="http://viewer.opencalais.com">http://viewer.opencalais.com</a>), Naïve Bayes tool.</td>
<td>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%.</td>
</tr>
<tr>
<td>P68</td>
<td>S58</td>
<td>Sentence classification</td>
<td>Bekwe-SanJuan algorithm; Agarwal’s algorithm; Teufel’s algorithm</td>
<td>Compared three different sentence-classification methods on set of 5W 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.</td>
</tr>
</tbody>
</table>

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.
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. 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 researchers can use to assess case studies. P31 presents a quality checklist for qualitative methods for an SR on agile methods and concept of strength of evidence. P52 presents a quality checklist for testing experiments and discusses some attempts to validate the checklist.

Table 16
Quality checklists and quality evaluation.

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

Table 17
Studies investigating data analysis and synthesis.

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

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; instead, they obtain a consensus about the quality of a paper given a quality checklist.

"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).

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 Dyba 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. It 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.

P16 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

<table>
<thead>
<tr>
<th>Paper</th>
<th>Study</th>
<th>First author</th>
<th>Topic</th>
<th>Study context</th>
<th>Study type</th>
<th>Quality score (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P16</td>
<td>S14</td>
<td>da Silva</td>
<td>Research questions</td>
<td>53 SRs</td>
<td>Tertiary study</td>
<td>100 × 7/7 = 100</td>
</tr>
<tr>
<td>P19</td>
<td>S16</td>
<td>Dieste</td>
<td>Updating an SR</td>
<td>Updating a complex SR</td>
<td>Lessons learnt</td>
<td>100 × 5/6 = 83.3</td>
</tr>
<tr>
<td>P25</td>
<td>S22</td>
<td>Felizardo</td>
<td>Graphical reporting</td>
<td>Re-analyzing an existing SR</td>
<td>Experiment</td>
<td>100 × 7/10 = 70</td>
</tr>
<tr>
<td>P49</td>
<td>S43</td>
<td>Petersen</td>
<td>Mapping study process</td>
<td>10 Mapping studies + example</td>
<td>Example and literature review</td>
<td>100 × 5.5/8 = 68.7</td>
</tr>
<tr>
<td>P64</td>
<td>S54</td>
<td>Bowes</td>
<td>SLR Tool (SLuRp)</td>
<td>Use on a complex SR</td>
<td>Discussion</td>
<td>NA</td>
</tr>
</tbody>
</table>
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 Compendex 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 study cases 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 jour-
nals 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 digi-
tal 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 dis-
tributed team to exhibit problems (P58), so we welcome such ini-
tiatives. 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, high-
lighting any limitations as well as reporting benefits. In particular,
we distrust the idea of automatic extraction of results from pri-
mary 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 estima-
tion method [9,23]. However, MMRE is still used in cost estimation
studies. If tools are used to extract data from cost estimation stud-
ies, 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 re-
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 dif-
ficult 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 fol-
lowing changes would appear to be appropriate:

1. To remove the proposal for constructing structured ques-
tions 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/exclu-
dion 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 appropri-
ate 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 con-
struction 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 stud-
ies 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 pos-
sible to apply to the wide range of studies included in this SR. How-
ever, 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 check-
list 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 con-
sideration of the empirical study type and its size/scope. How-
ever, 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 explic-
itly 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 valid-
dated 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.


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

<table>
<thead>
<tr>
<th>Issue Id</th>
<th>Issue text</th>
<th>Type</th>
<th>Suggestion for guidelines Yes/No</th>
<th>Novice issues Yes/No</th>
<th>Education issues Yes/No</th>
<th>Position in Paper</th>
<th>Stage in SR Process addressed</th>
<th>Importance (either text or number of “votes”)</th>
<th>Related Issue</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>For each issue/problem raised/problemsolution proposed specify the issue/problem using the same text as the papers authors</td>
<td>Advice (including Best practice)</td>
<td>Education (including training, gaining experience)</td>
<td>Page number or Table number or Id</td>
<td>Research question/Protocol/Search/Selection/Data extraction/Quality Assessment/Data Aggregation/Data Synthesis/Reporting</td>
<td>A ratio indicating number of votes out of possible votes. Or an textual indication of relative importance</td>
<td>Reference to any related issue</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Appendix B. Papers excluded from the SR during data extraction

<table>
<thead>
<tr>
<th>Authors</th>
<th>Title</th>
<th>Source</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boell S.K., Cecez-Kecmanovic D.</td>
<td>Literature reviews and the hermeneutic circle</td>
<td>Australian Academic and Research Libraries</td>
<td>General critique of SRs. Not SE oriented</td>
</tr>
<tr>
<td>Brereton P.</td>
<td>A study of computing undergraduates undertaking a systematic literature review</td>
<td>IEEE Transactions on Education</td>
<td>SLR was not a software engineering topic</td>
</tr>
<tr>
<td>Budgen, D., John Bailey, Mark Turner, Barbara Kitchenham, Pearl Brereton, Stuart Charters</td>
<td>Lessons from a cross domain investigation of empirical practices</td>
<td>EASE 2008</td>
<td>Preliminary version of Budgen et al., 2009, so also rejected</td>
</tr>
<tr>
<td>de Almeida Biolchini, Jorge Calmon, Paula Gomes Mian, Ana Candida Cruz Natali, Tayana Uchoa Conte, Guilherme Horta Travassos</td>
<td>Scientific research ontology to support systematic review in software engineering</td>
<td>Advanced Engineering Informatics</td>
<td>No clear implications for SR processes</td>
</tr>
<tr>
<td>Nakagawa E.Y., Feitosa D., Felizardo K.R.</td>
<td>Using systematic mapping to explore software architecture knowledge</td>
<td>ICSE</td>
<td>Just a straightforward mapping study</td>
</tr>
<tr>
<td>Major L., Kyriacou T., Brereton O.P.</td>
<td>Systematic literature review: Teaching novices programming using robots</td>
<td>IET Seminar Digest, 2011</td>
<td>Failed inclusion criteria. Primarily an SR not aimed at investigating SR process issues</td>
</tr>
<tr>
<td>Ramey J., Rao P.G.</td>
<td>The systematic literature review as a research genre</td>
<td>IEEE International Professional Communication Conference</td>
<td>General discussion. Not SE oriented</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Paper Number</th>
<th>Study Number</th>
<th>Authors</th>
<th>Year</th>
<th>Title</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>S1</td>
<td>Babar, Muhammad Ali, He Zang</td>
<td>2009</td>
<td>Systematic literature reviews in software engineering: Preliminary results from interviews with researchers</td>
<td>International Symposium on Empirical Software Engineering and Measurement (ESEM)</td>
</tr>
<tr>
<td>Paper Number</td>
<td>Study Number</td>
<td>Authors</td>
<td>Year</td>
<td>Title</td>
<td>Source</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>------------------------------------------------------------------------</td>
<td>------</td>
<td>----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>P6</td>
<td>S5</td>
<td>Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M., Khalil, M.</td>
<td>2007</td>
<td>Lessons from applying the systematic literature review process within the software engineering domain</td>
<td>Journal of Systems and Software (JSS), 80 (4), 571–583</td>
</tr>
<tr>
<td>P7</td>
<td>S6</td>
<td>Budgen, D., Drummond, S., Brereton, P. and Holland, N.</td>
<td>2012</td>
<td>What Scope is there for Adopting Evidence-Informed teaching in SE</td>
<td>International Conference on Software Engineering (ICSE)</td>
</tr>
<tr>
<td>P9</td>
<td>S8</td>
<td>Chen, Lianping, Muhammad Ali Babar and He Zhang</td>
<td>2010</td>
<td>Towards an Evidence-Based Understanding of Electronic Data Sources</td>
<td>EASE</td>
</tr>
<tr>
<td>P10</td>
<td>S9</td>
<td>Cruzes D., Mendonca M., Basili V., Shull F., Jino M.</td>
<td>2007</td>
<td>Using context distance measurement to analyze results across studies</td>
<td>ESEM</td>
</tr>
<tr>
<td>P13</td>
<td>S12</td>
<td>Cruzes, D.S., Tore Dybå</td>
<td>2011</td>
<td>Recommended Steps for Thematic Synthesis in Software Engineering</td>
<td>ESEM</td>
</tr>
<tr>
<td>P14</td>
<td>S13</td>
<td>Cruzes, Daniela, Tore Dybå</td>
<td>2010</td>
<td>Synthesizing evidence in software engineering research</td>
<td>ESEM</td>
</tr>
<tr>
<td>P18</td>
<td>S15</td>
<td>Dieste, O. and Padua, A.G.</td>
<td>2007</td>
<td>Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews</td>
<td>ESEM</td>
</tr>
<tr>
<td>P21</td>
<td>S18</td>
<td>Dieste, O., Enrique Fernández, Ramón García-Martínez, Natalia Juristo</td>
<td>2011</td>
<td>The risk of using the Q heterogeneity estimator for software engineering experiments</td>
<td>EASE</td>
</tr>
<tr>
<td>P22</td>
<td>S19</td>
<td>Dieste, O., Grimán, A., Juristo, N. and Saxena, H.</td>
<td>2011</td>
<td>Quantitative determination of the relationship between internal validity and bias in software engineering: consequences for systematic literature reviews</td>
<td>ESEM</td>
</tr>
<tr>
<td>P23</td>
<td>S20</td>
<td>Dybå, T., Dingsøyr, T., G.K. Hanssen</td>
<td>2007</td>
<td>Applying systematic reviews to diverse study types: an experience report</td>
<td>ESEM</td>
</tr>
</tbody>
</table>

(continued on next page)
<table>
<thead>
<tr>
<th>Paper Number</th>
<th>Study Number</th>
<th>Authors</th>
<th>Year</th>
<th>Title</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>P24</td>
<td>S21</td>
<td>Dybå, T., Torgeir Dingsøyr</td>
<td>2008</td>
<td>Strength of evidence in systematic reviews in software engineering</td>
<td>ESEM</td>
</tr>
<tr>
<td>P25</td>
<td>S22</td>
<td>Felizardo K.R., Riaz M., Sulayman M., Mendes E., MacDonell S.G., Maldonado J.C.</td>
<td>2011</td>
<td>Analyzing the use of graphs to represent the results of systematic reviews in software engineering</td>
<td>Proceedings – 25th Brazilian Symposium on Software Engineering, SBES 2011</td>
</tr>
<tr>
<td>P26</td>
<td>S23</td>
<td>Felizardo, K.R., Andery, G.F., Paulovich, F.V., Minghim, R., Maldonado, J.C.</td>
<td>2012</td>
<td>A visual analysis approach to validate the selection review of primary studies in systematic reviews</td>
<td>IST, 54 (10), 1079–1091</td>
</tr>
<tr>
<td>P27</td>
<td>S24</td>
<td>Felizardo, Katia Romera, Elisa Yumi Nakagawa, Daniel Feitosa, Rosane Minghim and José Carlos Maldonado</td>
<td>2010</td>
<td>An Approach Based on Visual Text Mining to Support Categorization and Classification in the Systematic Mapping Activity in Systematic Literature Reviews</td>
<td>EASE</td>
</tr>
<tr>
<td>P28</td>
<td>S25</td>
<td>Felizardo, Katia Romero; Norsaremah Salleh, Rafael Messias Martins, Emilia Mendes, Stephen G. Macdonell and José Carlos Maldonado</td>
<td>2011</td>
<td>Using Visual Text Mining to Support the Study Selection Activity in Systematic Literature Reviews</td>
<td>ESEM</td>
</tr>
<tr>
<td>P29</td>
<td>S26</td>
<td>Fernandez-Saez A.M., Bocco M.G., Romero F.P.</td>
<td>2010</td>
<td>SLR-Tool a tool for performing systematic literature reviews</td>
<td>ICSOFT 2010 – Proceedings of the 5th International Conference on Software and Data Technologies ESEM</td>
</tr>
<tr>
<td>P31</td>
<td>S28</td>
<td>Ivarsson M., Gorschek T.</td>
<td>2011</td>
<td>A method for evaluating rigor and industrial relevance of technology evaluations</td>
<td>ESE</td>
</tr>
<tr>
<td>P33</td>
<td>S30</td>
<td>Kitchenham, B.A. Li, Z., Burn, A.J.</td>
<td>2011</td>
<td>Validating Search Process in Systematic Literature Reviews: Repeatability of Systematic Literature Reviews</td>
<td>EAST</td>
</tr>
<tr>
<td>P34</td>
<td>S31</td>
<td>Kitchenham B., Pearl Brereton, Zhi Li, David Budgen &amp; Andrew Burn</td>
<td>2011</td>
<td>Mapping study completeness and reliability – a case study</td>
<td>EASE</td>
</tr>
<tr>
<td>P35</td>
<td>S32</td>
<td>Kitchenham, B., Pearl Brereton and David Budgen</td>
<td>2012</td>
<td>Mapping study completeness and reliability – a case study</td>
<td>IST, 53 (6), 638–651</td>
</tr>
<tr>
<td>P36</td>
<td>S33</td>
<td>Kitchenham, B.A., Budgen, D., Pearl Brereton, O.</td>
<td>2011</td>
<td>Using mapping studies as the basis for further research – A participant-observer case study</td>
<td>ESEM</td>
</tr>
<tr>
<td>P37</td>
<td>S33</td>
<td>Kitchenham, Barbara A., David Budgen and O. Pearl Brereton</td>
<td>2010</td>
<td>The value of mapping studies – A participant-observer case study</td>
<td>EASE</td>
</tr>
<tr>
<td>P38</td>
<td>S34</td>
<td>Kitchenham, B., Pearl Brereton, David Budgen</td>
<td>2010</td>
<td>The educational value of mapping studies of software engineering literature</td>
<td>ICSE</td>
</tr>
<tr>
<td>P39</td>
<td>S35</td>
<td>Kitchenham, B., Pearl Brereton, David Budgen, Zhi Li</td>
<td>2009</td>
<td>An Evaluation of Quality Checklist Proposals – A participant-observer cases study</td>
<td>EASE</td>
</tr>
<tr>
<td>P40</td>
<td>S36</td>
<td>Kitchenham, B., Pearl Brereton, Mark Turner, Mahmood Niazi, Stephen G. Linkman, Riallet Pretorius, David Budgen</td>
<td>2009</td>
<td>The impact of limited search procedures for systematic literature reviews A participant-observer case study.</td>
<td>ESEM</td>
</tr>
<tr>
<td>P42</td>
<td>S37</td>
<td>Kitchenham, B.A., Andrew J. Burn, Zhi Li</td>
<td>2009</td>
<td>A Quality Checklist for Technology-Centred Testing Studies</td>
<td>EASE</td>
</tr>
<tr>
<td>Paper Number</td>
<td>Study Number</td>
<td>Authors</td>
<td>Year</td>
<td>Title</td>
<td>Source</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------</td>
<td>---------</td>
<td>------</td>
<td>-------</td>
<td>--------</td>
</tr>
<tr>
<td>P44</td>
<td>S38</td>
<td>Kitchenham, B.A., Sjoberg, D.I.K., Dyba, T., Pfahl, D., Brereton, P., Budgen, D., Host, M., Runeson, P.</td>
<td>2012</td>
<td>Three empirical studies on the agreement of reviewers about the quality of software engineering experiments</td>
<td>IST, 54 (8), 804–819</td>
</tr>
<tr>
<td>P45</td>
<td>S39</td>
<td>MacDonell, S., Shepperd, M., Kitchenham, B., Mendes, E.</td>
<td>2010</td>
<td>How reliable are systematic reviews in empirical software engineering?</td>
<td>IEEE Transactions on Software Engineering (TSE), 36 (5), 676–687</td>
</tr>
<tr>
<td>P46</td>
<td>S40</td>
<td>Malheiros, Viviane, Erika Hohn, Roberto Pinho, Manoel Mendonca, Jose Carlos Maldonado</td>
<td>2007</td>
<td>A Visual Text Mining approach for Systematic Reviews</td>
<td>ESEM</td>
</tr>
<tr>
<td>P47</td>
<td>S41</td>
<td>Oates, Briony J., Graham Capper</td>
<td>2009</td>
<td>Using systematic reviews and evidence-based software engineering with masters students</td>
<td>EASE</td>
</tr>
<tr>
<td>P48</td>
<td>S42</td>
<td>Petersen, K., Ali, N.B.</td>
<td>2011</td>
<td>Identifying strategies for study selection in systematic reviews and maps</td>
<td>ESEM</td>
</tr>
<tr>
<td>P51</td>
<td>S45</td>
<td>Riaz, Mehwish,; Muhammad Sulayman, Norsaremah Salleh and Emilia Mendes Runeson, P and Höst, M.</td>
<td>2010</td>
<td>Experiences Conducting Systematic Reviews from Novices' Perspective</td>
<td>Empirical Software Engineering 14, 131–164</td>
</tr>
<tr>
<td>P52</td>
<td>S27</td>
<td>Skoglund, Mats and Per Runeson</td>
<td>2009</td>
<td>Guidelines for conducting and reporting case study research in software engineering Reference-based search strategies in systematic reviews</td>
<td>EASE</td>
</tr>
<tr>
<td>P53</td>
<td>S46</td>
<td>Staples, M., Niazi, M.</td>
<td>2007</td>
<td>Experiences using systematic review guidelines Towards Evidence-Based Ontology for Supporting Systematic Literature Review</td>
<td>JSS, 80 (9), 1426–1437</td>
</tr>
<tr>
<td>P55</td>
<td>S47</td>
<td>Sun, Yueming, Ye Yang, He Zhang, Wen Zhang, Qing Wang</td>
<td>2012</td>
<td>Towards Evidence-Based Ontology for Supporting Systematic Literature Review</td>
<td>EASE</td>
</tr>
<tr>
<td>P57</td>
<td>S50</td>
<td>Turner, M., Barbara Kitchenham, Pearl Brereton, David Budgen</td>
<td>2008</td>
<td>Lessons learnt Undertaking a Large-scale Systematic Literature Review</td>
<td>EASE</td>
</tr>
<tr>
<td>P58</td>
<td>S51</td>
<td>Zhang, He and Muhammad Ali Babar</td>
<td>2010</td>
<td>On Searching Relevant Studies in Software Engineering</td>
<td>EASE</td>
</tr>
<tr>
<td>P59</td>
<td>S52</td>
<td>Zhang, He, Muhammad Ali Babar</td>
<td>2011</td>
<td>An Empirical Investigation of Systematic Reviews in Software Engineering</td>
<td>ESEM</td>
</tr>
</tbody>
</table>

(continued on next page)
## References


### Appendix C (continued)

<table>
<thead>
<tr>
<th>Paper Number</th>
<th>Study Number</th>
<th>Authors</th>
<th>Year</th>
<th>Title</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>P62</td>
<td>S51</td>
<td>Zhang, He, Muhammad Ali Babar, Paolo Tell</td>
<td>2011</td>
<td>Identifying relevant studies in software engineering</td>
<td>IST, 53 (6), 626–637</td>
</tr>
<tr>
<td>P63</td>
<td>S53</td>
<td>Zhang, He, Muhammad Ali Babar, Xu Bai, Juan Li, Huang, Liguo</td>
<td>2011</td>
<td>An Empirical Assessment of A Systematic Search Process for Systematic Reviews</td>
<td>EASE</td>
</tr>
<tr>
<td>P64</td>
<td>S54</td>
<td>Bowes, David, Hall, Tracy and Beecham, Sarah</td>
<td>2012</td>
<td>SLuKp – A tool to help large complex systematic literature reviews deliver valid and rigorous results</td>
<td>Evidential Assessment of Software Technologies (EAST)</td>
</tr>
<tr>
<td>P65</td>
<td>S55</td>
<td>Cruzes, D. Mendonca, M., Basili, V., Shull, F. and Jino, N.</td>
<td>2007</td>
<td>Extracting information from Experimental Software Engineering papers</td>
<td>International Conference of the Chilean Society of Computer Science, SCC’07</td>
</tr>
<tr>
<td>P66</td>
<td>S56</td>
<td>Mian, P., T. Conte, A. Natali, J. Biolchini, E. Mendes, G. Travassos</td>
<td>2005</td>
<td>Lessons learned on applying systematic reviews to software engineering</td>
<td>Proceedings of the 2nd Experimental Software Engineering Latin American Workshop (ESELAW’05), Brazil</td>
</tr>
<tr>
<td>P67</td>
<td>S57</td>
<td>Mohagheshi, P., Conradi, R.</td>
<td>2006</td>
<td>Vote counting for combining quantitative evidence from empirical studies – an example.</td>
<td>Proc ISESE ’06, pp. 22–42</td>
</tr>
<tr>
<td>P68</td>
<td>S58</td>
<td>Torres, José Alberto S., Cruzes, Daniela and Salvador, Laís do Nascimento</td>
<td>2012</td>
<td>Automatic Results Identification in Software Engineering Papers. Is it possible?</td>
<td>12th International Conference of Computational science and Its Applications</td>
</tr>
</tbody>
</table>


