How Reliable Are Systematic Reviews in Empirical Software Engineering?

Stephen MacDonell, Martin Shepperd, Barbara Kitchenham and Emilia Mendes

Abstract

BACKGROUND – the systematic review is becoming a more commonly employed research instrument in empirical software engineering. Before undue reliance is placed on the outcomes of such reviews it would seem useful to consider the robustness of the approach in this particular research context.
OBJECTIVE – the aim of this study is to assess the *reliability* of systematic reviews as a research instrument. In particular we wish to investigate the consistency of process and the stability of outcomes.
METHOD – we compare the results of two independent reviews undertaken with a common research question.
RESULTS – the two reviews find similar answers to the research question, although the means of arriving at those answers vary.
CONCLUSIONS – in addressing a well-bounded research question, groups of researchers with similar domain experience can arrive at the same review outcomes, even though they may do so in different ways. This provides evidence that, in this context at least, the systematic review is a robust research method.

Index Terms

Empirical software engineering, meta-analysis, systematic review, cost estimation.

Stephen MacDonell is with the School of Computing & Mathematical Sciences, Auckland University of Technology, Private Bag 92006, Auckland 1142, NZ.
Martin Shepperd is with the Department of I.S. & Computing, Brunel University, West London, UB8 3PH, UK.
Barbara Kitchenham is with the School of Computing & Mathematics, Keele University, Keele, Staffordshire, ST5 5BG, UK.
Emilia Mendes is with the Department of Computer Science, The University of Auckland, Private Bag 92019, Auckland 1142, NZ.

1 INTRODUCTION

For some years software engineering researchers have endeavored to collect empirical evidence to support or rebut emerging ideas concerning methods, technology and so forth. Groups such as the Software Engineering Lab at the University of Maryland [1] and the Fraunhöfer Institute for Experimental Software Engineering [2] have been at the forefront of such research.

Following on from this work, there has been a move to explicitly position software engineering as an *evidence-driven* discipline (see for example Kitchenham *et al.* [3] and Dybå *et al.* [4]). The reasons are not hard to determine. Software is ubiquitous and therefore its societal and economic impact is considerable. Consequently there is an urgency to ensure we advocate and deploy best practice. However, establishing what constitutes best practice and in what context is an empirical question. So there are increasing numbers of empirical studies to evaluate different software engineering practices and techniques through a mixture of experiment and quasi-experiment, case study, observation study, action research and ethnography.

Perhaps unsurprisingly, the empirical results have not always been consistent, so understanding what the body of evidence indicates can be a non-trivial task and not particularly amenable to *ad hoc* methods such as the narrative literature review. Thus there is now a move to adopt more rigorous approaches to identifying and synthesizing *all* the available evidence by means of the systematic review methodology, popularised by the medical community.

Whilst the authors of this paper are supportive of this position we pose the question: how reliable are systematic literature reviews (SLRs) as a research instrument? In other words, how much does the outcome of a systematic review depend upon who is involved in the process and the minutiae of micro decisions that any research process must entail? We believe this to be an important question since highly significant decisions may be made based upon the outcome of systematic reviews. Therefore we wish to have confidence that such review outcomes are stable and insensitive to minor quirks of the process. A positive result should reassure those considering the use of an SLR that the method is robust. Note that this is a separate problem from determining whether

'sufficient' primary studies exist in order yield a definitive answer to a specific SLR question. However, we would observe in passing, that one role of the SLR is to help direct our research energies to those areas where they are most needed.

In order to address the question of systematic review reliability, we compare in detail, two independent reviews that address the same research question. Since both studies are already published [5], [6] this paper considers the SLR outcomes only to the extent that is necessary to evaluate the reliability of the process. In other words, for the purposes of this paper the topic of the reviews is less important than the fact that we have two independently conducted SLRs that focus on the *same* research question.

The remainder of this paper is organised as follows. We provide a brief history of the systematic review and how it has been adopted by the software engineering community. Then we review related work looking at systematic reviews in all disciplines. This is followed by a description of our research method and our 'meta-protocol'. We then compare the outcomes of the two systematic reviews and consider where the approaches differed and the extent to which this had an impact upon the review outcomes. The paper concludes by discussing the significance of this study, to what extent it might be generalised, and areas for follow up work.

2 Related Work

2.1 Systematic Reviews and Meta-Analyses

Combining results from more than one primary study is an important part of research; in particular reconciling inconsistent results. There has been a perceived need to address this problem since the work of the statistician Karl Pearson at the beginning of the twentieth century.

Presently the lead in evidence-informed policy and practice lies with medicine [7]. In 1972, Archie Cochrane criticized medicine for not organizing its knowledge in any systematic, reliable, and cumulative way. The result was that health care was inconsistent, often ineffective and sometimes even harmful to patients. In October 2003, an international collaboration named after him was set up with the aim of encouraging and publishing systematic reviews of health care interventions. The two main

methodological principles of the Cochrane Collaboration are the need for unbiased comparisons of interventions (i.e. randomized controlled trials) and the importance of aggregating evidence from different studies to obtain reliable estimates of effects. These principles were based on substantial bodies of evidence showing that studies that do not attempt to minimize bias and reviews based on subsets of relevant studies can result in misleading policy and practice. The equivalent for "evidence on social and behavioural interventions and public policy, including education, criminal justice, and social welfare, among other areas" is the Campbell Collaboration.

Studies in other disciplines have confirmed that non-systematic reviews can be biased, may miss relevant papers and may lead to incorrect conclusions. For example, Oakley and colleagues found six non-systematic literature reviews of older people and health accident prevention which included a total of 137 reviews. However, only 33 studies were common to at least two reviews, and only one study was treated consistently in all six reviews [8], [9]. Similar results were found with two non-systematic reviews of anti-smoking education in young people. A total of 27 studies were included in both reviews but only three studies were common to both [10].

Experts can also be wrong. Nobel Laureate Linus Pauling reviewed literature about the common cold non-systematically and concluded that very large doses of vitamin C were beneficial [11]. Knipschild [12] tested Paulings claim with a separate exhaustive search finding 61 trials of which only 15 appeared methodologically sound. He concluded that even mega doses of vitamin C could not prevent a cold though it might shorten its duration. Paulings review missed five of the top 15 studies.

As individuals we can also be biased in our selection of reference material. Shadish [13] surveyed authors of over 280 articles in psychological journals and found that more often than not studies get cited simply because they supported the authors own argument, and not because the study was particularly reliable.

Approaches to the challenge of systematic evidence synthesis broadly fall into the following categories:

- *narrative review articles:* these provide an informal, qualitative summary of an *ad hoc* selection from a body of literature. They are characterized by limited and superficial approaches to combining primary study results.

- *bibliometric analysis:* a quantitative review of a body of literature. Statistical methods are used to reveal the historical trends or patterns of authorship, publication and usage in subject fields rather than the specific content of papers.
- *systematic literature reviews (SLR):* in contrast to the qualitative narrative review described above, these represent a repeatable method for identifying *all* relevant primary studies that satisfy the inclusion criteria of an explicit and publicly available protocol to answer a specific research question, e.g. the relative effectiveness of different requirements elicitation techniques [14].
- *meta-analysis:* a statistical analysis from the pooled results from two or more primary studies. Where a single study is being re-analysed this is sometimes referred to as a secondary analysis.
- *prospectively planned meta-analysis:* here groups of researchers across multiple centres take part in the joint planning and conduct of the data collection and analysis. Although this approach leads to a reduction in differences between studies [15], any problems in the design of single studies are of course multiplied.

2.2 Systematic Reviews in Empirical Software Engineering

Turning to software engineering, we see that there has been interest in the idea of building a body of evidence since the 1990s, for example, Basili *et al*. [16] suggested that individual studies should be seen as part of a 'family of studies' rather than isolated events. Thus, studies could be replicated and context variables varied so that a framework for organizing related studies could be built. However, a framework, making explicit the different models, and documenting key choices and rationales of experimental design used in each experiment, is required. Although this process might be seen as desirable in itself, it does not go as far as meta-analysis in that it concentrates on replicating studies and refining results, rather than combining results from a number of separate yet, hopefully, comparable studies.

Other researchers such as Hayes [17], Pickard *et al*. [18] and Miller [19] started to consider the extent to which empirical results might be pooled for meta-analysis. The difficulty that these researchers identify is that few primary studies provide access to raw data, or other experimental details; consequently, results from individual studies

frequently are neither generalizable [18] nor reliable [19]. Indeed Pickard *et al*. argue that without agreed sampling protocols for properly defined software engineering populations, and a set of standard measures recorded for all empirical studies, meta-analyses cannot be conducted [18]. Thus, for the present, we turn to the systematic literature review.

Kitchenham *et al*. [20] performed a systematic review of systematic literature reviews published between January 2004 and June 2007 in 10 major software engineering journals and in the proceedings from three major software engineering conferences, to address the questions:

- How much SLR activity has there been since 2004?
- What research topics are being addressed?
- Who is leading the SLR research?
- What are the limitations of current research?

They found 20 papers that included a literature review performed with some degree of rigour (whether or not the authors referred to their survey as"systematic") and 14 that they did not consider methodologically rigorous enough to be called systematic. An evaluation of the quality of the systematic literature reviews using the DARE criteria [21] found that all but two studies scored more than 2 on a 4-point scale and the quality appeared to be increasing. Eight of the studies considered research trends rather than the assessment of alternative technology. Of the twelve studies that addressed more detailed research questions, six addressed cost estimation topics and three addressed testing topics, others addressed individual topics (CMM effectiveness, Software Architecture Evaluation methods, COTS development). Authors of systematic literature reviews were primarily European researchers, in particular researchers at the Simula Research Laboratory. The main limitations of current systematic literature reviews were that many omitted a quality assessment of the included papers, and that, to date, few reviews are developing evidence-based guidelines appropriate for practitioners.

2.3 Reliability of Systematic Reviews in Other Disciplines

As mentioned previously, systematic literature reviews have been adopted as a tool for summarising evidence in a range of other disciplines (from psychiatry to social policy)

with the goal of being objective and repeatable. One of the concerns raised has been the infrequent occurrence of reviews that lead to definitive conclusions. Of course there are many explanations [22]. These include high levels of heterogeneity amongst the primary studies [23] but also methodological issues as embodied by, for example, an SLR of primary study appraisal tools which are designed to provide support for reviewers [24]. Such tools are designed to assist reviewers assess the quality and relevance of primary studies. This study found much diversity in tools, approaches and that there was no 'gold standard' approach. Clearly this is a potential source of unreliability and variance in the outcome of a systematic review.

3 RESEARCH QUESTION

From the foregoing discussion we see that systematic reviews are a potentially important research tool for software engineers and that a growing number of such reviews have been performed and published. Therefore, the main research question of interest here is: How reliable are systematic reviews? We narrow the question somewhat by focusing upon two independent reviews of a topic drawn from software engineering – the comparison of within-company and cross-company models in accurately predicting software development effort. Supplementary questions are:

- Q1: What are the similarities and differences between the search strategies applied?
- Q2: What are the similarities and differences between the data extracted from each study?
- Q3: What are the similarities and differences between the data aggregation strategies?
- Q4: How much effort was expended and how was this distributed[1]?

3.1 The 'Meta-Protocol'

In late 2005 two teams of researchers with similar backgrounds agreed to undertake *independent* systematic reviews of the same research question in order to address the issue of review reliability. The particular question addressed by each SLR was:

1. Although this question appeared in our meta-protocol it was not addressed by Team SLR1, hence no comparisons were possible and the question is only retained for completeness.

> "What evidence is there that cross-company estimation models are at least as good as within-company estimation models for predicting effort for software projects?"

The motivation for these reviews was to determine whether software project managers are better advised using their own local data or whether they can make do with data derived from other software development organizations. Note that the two reviews have been independently published as standalone pieces of research (SLR1) [5] (extended beyond our meta-protocol in [25]) and (SLR2) [6].

The two teams were as follows:

- Team SLR1: Barbara Kitchenham, Emilia Mendes, Guilherme Travassos
- Team SLR2: Stephen MacDonell, Martin Shepperd

The teams and review task can be characterised as follows. All five researchers had considerable experience both as researchers and also as published authors within the field under investigation. The problem domain was not perceived to be very extensive but it suffered from being somewhat ill-specified with no agreed terminology within the community. Moreover, in some cases candidate studies were principally concerned with other phenomena e.g. a comparison of different prediction methods, and thus results relevant to our review could be 'deeply buried'.

The two teams negotiated and agreed on the meta-level issues relevant to the reviews, themselves the subject of a study 'meta-protocol'. This specified the main and supplementary research questions stated above, the research method to be adopted including research topic and questions[2], team composition, and high-level data extraction and aggregation processes.

The meta-protocol also specified the basis for the comparison of the reviews. Among the criteria to be considered were the following (each related to one of the four questions listed above):

- Q1.1: Sources searched – where did the teams look for studies?
- Q1.2: Search strategy – how was the search executed (automatically, manually)?

2. The two teams formulated the research issue as two slightly different but logically equivalent research questions.

- Q1.3: Terms used in searching – what fields or similar were considered in the search, over what period of time?
- Q1.4: Papers found – retrieved or located as being potentially relevant
- Q1.5: Papers discarded – those studies not selected, and why
- Q1.6: Papers included – those selected as primary studies
- Q2.1: Analysis approach – issues considered and steps followed
- Q3.1: Analysis outcomes – interpretations and conclusions drawn
- Q4: Effort expended in various activities – determining the protocol, data extraction, data aggregation, write-up of outcomes.

These aspects of the two reviews are explored in the following section.

An agreed meta-protocol was needed to ensure that the teams addressed a common research question, crucial to the notion of performing a comparison. However, other aspects of each review (where to search, basis for inclusion) were left to the individual teams to determine. While members of the two teams had worked together previously they had not undertaken an SLR, and so there is no basis for assuming that prior common experiences influenced decisions such as where to search for primary studies. Note also that once the meta-protocol had been agreed the two teams did not communicate with regard to the review until both teams had completed the task.

4 RESULTS

In order to highlight the similarities and differences across the two studies we have constructed flowcharts that represent the activities undertaken in each, and show them side by side so that a mapping one to the other can be considered (see Figures 1, 2 and 3).

4.1 Comparing search strategies (Q1)

Q1.1: Sources searched: Access to electronic and physical resources differed across the two teams, meaning that they considered different research sources in their respective reviews.

Table 1 summarises the sources searched and the numbers of studies retrieved by the two SLRs. A blank indicates the source was not used. A number indicates the count

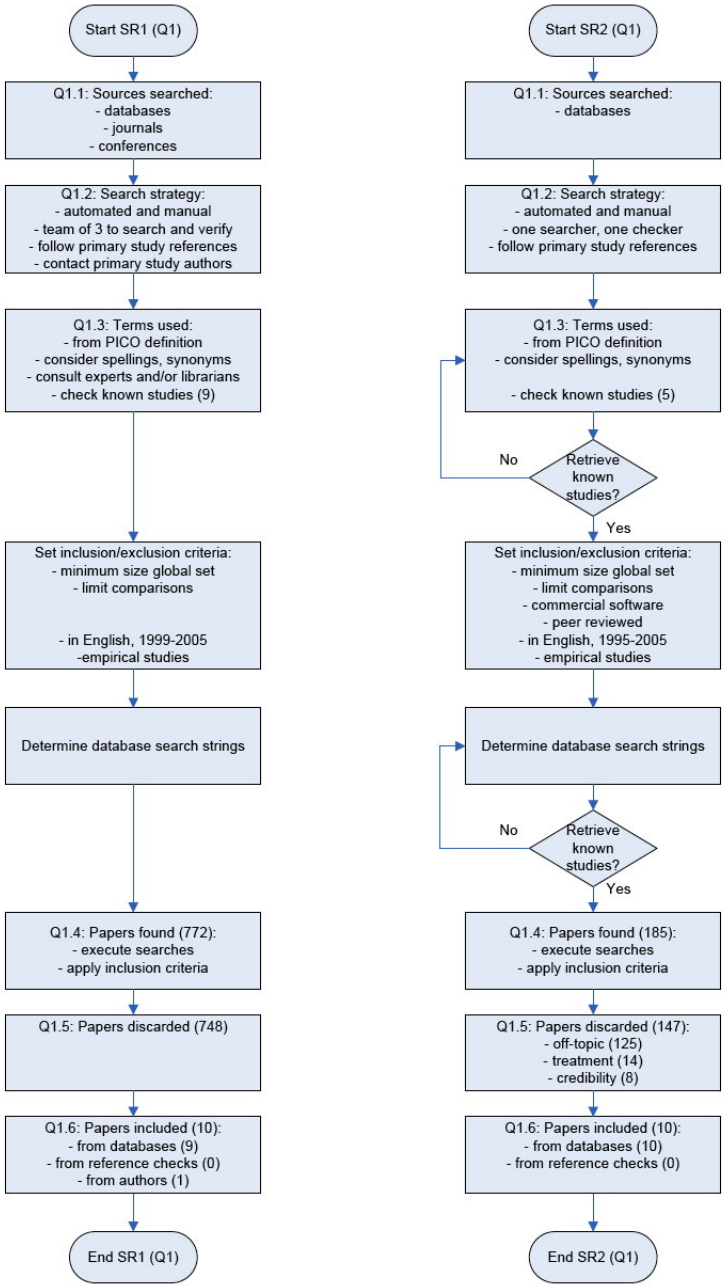

Fig. 1. Flow Chart to Compare Search Strategies of the Systematic Review Teams

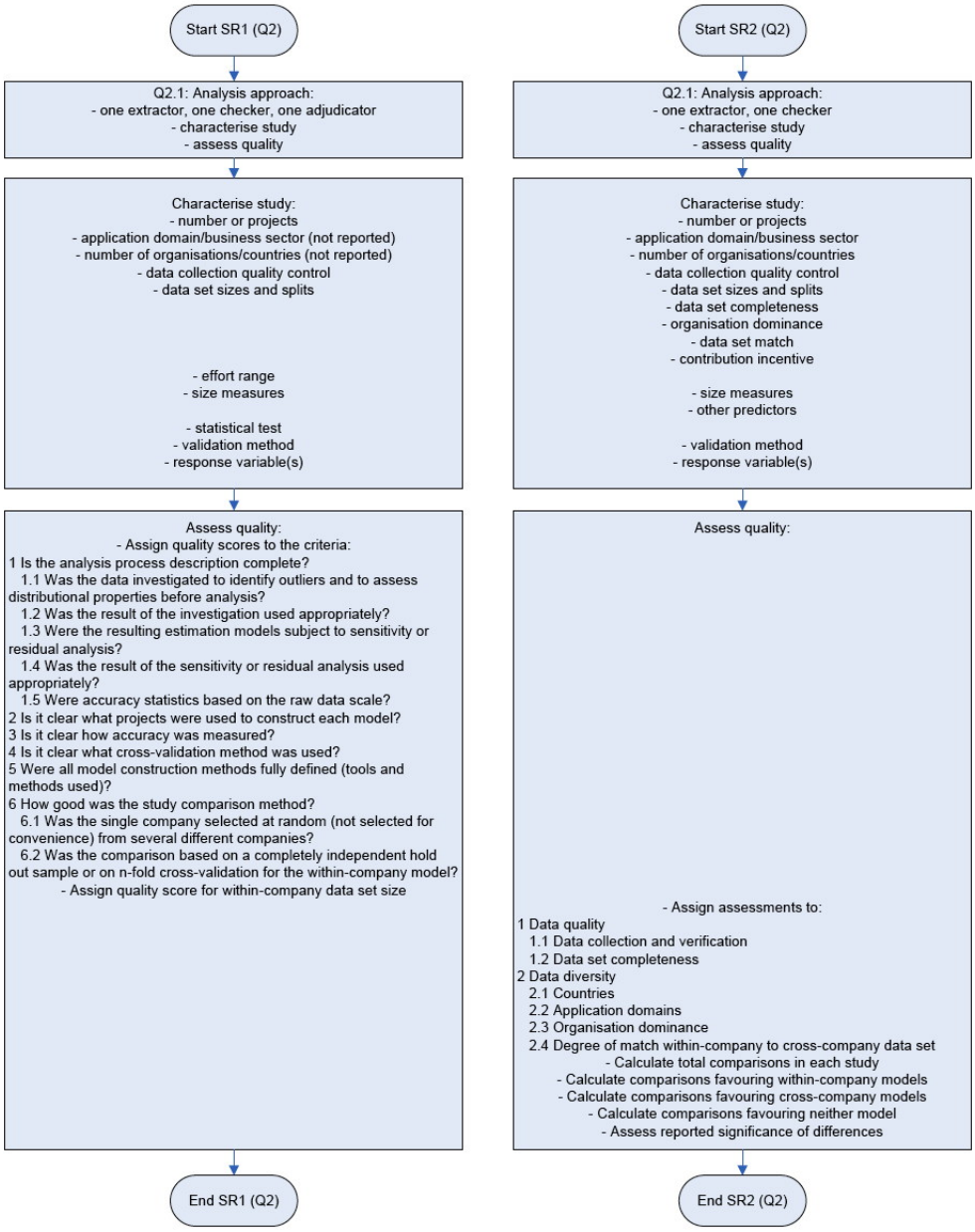

Fig. 2. Flow Chart to Compare Data Extraction and Quality Checking Processes of the Systematic Review Teams

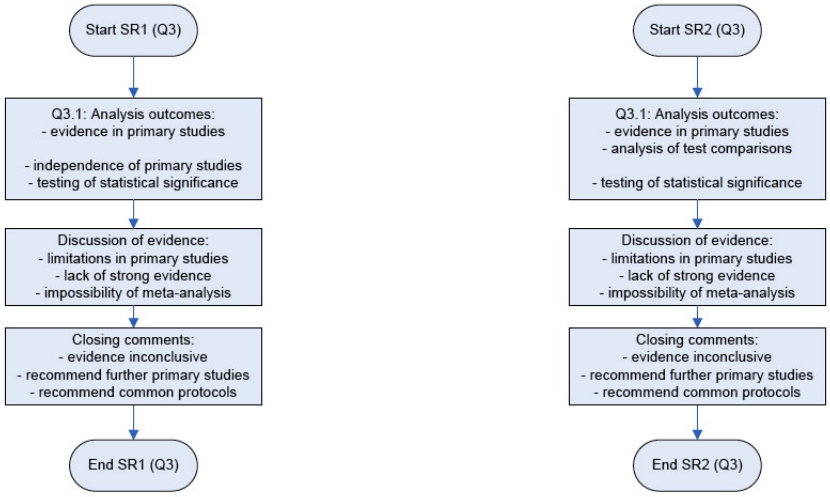

Fig. 3. Flow Chart to Compare Data Aggregation and Synthesis Processes of the Systematic Review Teams

of papers retrieved and the value in parentheses gives the number of studies actually retained for the SLR. Note that SLR2 performed a single search that combined the results of EI Compendex and INSPEC (via the Engineering Village portal). It can be seen that SLR1 utilised a more targeted set of journal and conference resources in comparison to SLR2, the latter relying more on broader database sources with a view to being more inclusive.

Q1.2: Search strategy: Both SLRs used a combination of automated database searching and manual citation analysis to search for potentially relevant studies. Where possible, queries were executed against all of the electronic sources (see the following subsection for further details). Each identified primary study was then considered as a source in its own right, in that each such primary study cited previous research potentially relevant to the SLR. The resources expended in searching differed across the two reviews - SLR1 utilised multiple searchers from the team (of three people) whereas SLR2 (comprising two people) used a searcher/checker approach. SLR1 also contacted researchers known to be working in this field so as to try to identify further relevant studies.

Q1.3: Terms used: The SLR1 search string was constructed primarily on the basis of the team's PICO definition, specifying the Population of interest, the Intervention, the

Source	SLR1	SLR2
D: ACM Digital Library	0(0)	15(2)
D: Blackwell/Synergy		5(0)
D: EI Compendex	60(1)	9(9)
D: INSPEC	224(4)	
D: EBSCOhost		3(2)
D: Expanded Academic ASAP		1(1)
D: IEEE Xplore	9(5)	30(6)
D: ProQuest		24(2)
D: Scholar Google		34(6)
D: ScienceDirect	453(4)	45(1)
D: Springerlink		11(1)
D: Wiley Interscience		0(0)
D: WoK Proceedings		5(5)
D: WoK Web of Science	19(4)	3(3)
J: Empirical Soft Eng	0(0)	
J: Info & Softw Technol	1(1)	
J: Management Science	0(0)	
J: Softw Process Improv & Pract	0(0)	
C: Evaluat & Assess in Softw Eng	1(1)	
C: Intl Conf on Softw Eng	2(2)	
C: Intl Symp on Softw Metrics	3(2)	

TABLE 1

Comparison of bibliographic sources utilisation (where D denotes database, J journal and C conference)

Comparison intervention, and the Outcomes sought. Synonyms used in known studies were then added to the string - the SLR1 team members were familiar with nine previous study and a further one in-press. In SLR2, the search string was determined more directly from the terms used in previously known studies, five in this case. By 'known studies', we mean that the sources were already stored electronically (either in personal bibliographic databases or in relevant directories) by one or more members of the teams.

Within-team discussions led to the addition of a small number of further search terms. Variations in spelling and structure (e.g. terms with and without hyphenation) were

considered and accounted for in the queries formed. At the same time, pilot testing against the search engines was also undertaken in order to consider the impact of search engine capabilities – the handling of Boolean combinations, sub-query nesting, the impact of hyphenation and so on. At the end of this process SLR1 had formulated a comprehensive single concatenated string (verified by domain experts) while SLR2 used a combination of three strings (using wildcards where possible), dealing with the function, the object and the context respectively of potentially relevant studies (e.g. cost model*, software project*, company specific). The resulting queries (in their generic forms due to differences in search capabilities between various databases) are shown in Appendix 1.

Throughout this iterative process SLR2 assessed and refined the search terms in order to ensure that the five known studies would be retrieved, in effect a capture-recapture process. In contrast, SLR1 relied on the *in principle* relevance of the terms to the known studies, that is, relying on a robust PICO definition to ensure that all relevant primary studies would be found. The two SLRs also adopted different time-spans in their search for material. SLR1 used the year of publication of the first-known primary study as their start year, resulting in a search period of 1999-2005, whereas SLR2 adopted effectively a ten-year time-window, producing a search period of 1995 to mid-2005. Whenever enabled by the search engine, full text as well as title + abstract + keyword searches were undertaken in both reviews. Returned results were limited to those published in English.

Other inclusion criteria applied in SLR1 were:

- data from more than 2 or 3 companies in the cross-company data set
- comparisons of single-organisation models to cross-company models (i.e. not to general cost-estimation models)
- validation data set based on single-organisation data only

Other inclusion criteria applied in SLR2 were:

- data from a minimum of 5 projects per company for at least 2 companies in the cross-company data set
- comparisons of single-organisation models to cross-company models (i.e. not to

Source	Missed Papers
D: ACM Digital Library	2
D: ScienceDirect	1
J: Management Science	1
C: International Symposium on Software Metrics	1

TABLE 2
Known papers missed in the search by SLR1

general cost-estimation models)

- substantially software projects (i.e. not hardware or co-design)
- commercial projects (i.e. not student projects)
- demonstrably peer reviewed (i.e. more than review of abstracts; not Technical Reports, postgraduate dissertations and undergraduate work).

Q1.4: Papers found: In sum (and including duplicates), SLR1 retrieved 772 potentially relevant papers and SLR2 identified 185 candidate studies (see Table 1). In comparing the two sets of results it appears that the more focused search approach employed in SLR2 resulted in a greater level of precision in search outcomes. Also of note is the fact that all nine studies known to one or both teams of reviewers were able to be retrieved from one or more databases. Table 2 indicates that specific SLR1 searches did not always have perfect recall in that known relevant articles were not retrieved. Table 3 provides an aggregate picture when combining individual database searches. Overall there was little difference in recall, however, the precision level for SLR2 was approximately 5 times greater than for SLR1.

Q1.5: Papers discarded: Of the 772 potentially relevant papers retrieved by SLR1, 748 were discarded as not being relevant to the study, using the following process. Each team member had specific responsibility for search, retrieval and recommendation of studies for a subset of sources so 24 document files were then exchanged among the members, with each file comprising a list of titles and abstracts for the papers retrieved and a recommendation as to whether or not the team member believed that the paper

	SLR1	SLR2
Retrieved	772	185
Detailed reviewed	24	38
Relevant retrieved	9	10
Total Relevant	10	11
Precision	0.012	0.054
Recall	0.900	0.909

TABLE 3
Overall precision and Recall of the SLR1 and SLR2 Search Strategies

met the inclusion criteria. The other two team members reviewed the recommendations and the team arrived at final consensus decisions. While disagreement among the team would have meant that the full paper would be reviewed, no such disagreements arose. After removing duplicates, a total of nine primary studies remained from SLR1.

In SLR2, 147 of the 185 retrieved studies were discarded as irrelevant leaving 38 for detailed review. The largest proportion by far (approximately 85 percent) were discarded as being off the topic of the review, while around 10 percent were not included due to the study treatment: for example, they did not use independent validation processes to assess model accuracy and so were model-fitting rather than prediction studies. The remaining studies were discarded due to questions over their credibility, leaving a final set of ten primary studies (less duplicates). Note that in SLR2 the decisions regarding inclusion/exclusion were made almost entirely by the searcher, with substantial discussion occurring only in relation to two of the eventually discarded papers. Decisions were made on the basis of title and abstract (and in SLR2, keyword) reviews in the first instance and then full paper reviews if needed.

Q1.6: Papers included: As a result of the first stage of the search process SLR1 had identified nine primary studies and SLR2 ten. Nine of these were in common across both SLRs, the difference being that SLR2 had included a 2003 paper published in the proceedings of the *International Symposium on Software Metrics*, authored by Mendes, Mosley and Counsell, that the SLR1 team had discarded (see Table 4). In this particular

study the comparison of within- and cross-company models had been undertaken using independent rather than common validation sets. While this was noted in the analysis by SLR2, in hindsight it should not have been included as a primary study.

Both teams then conducted a secondary search process as noted above. SLR1 undertook a detailed comparative analysis of the citations included in each of the primary studies whereas SLR2 used a simple scan and review of the references in each. Neither process led to the identification of further candidate primary studies. SLR1 also additionally asked for feedback from the authors of the identified primary studies regarding possible ongoing work, a step not taken by SLR2. This led to SLR2 missing a relevant primary study (authored by Mendes, Lokan, Harrison and Triggs) that had been accepted for publication but had not gone to print at the time of the review. As this work had been co-authored by one of the SLR1 team it was naturally added to their list of primary studies. This meant that both reviews had selected ten primary studies, with nine of these in common (see Table 4).

While this may be considered to be a small number of primary studies such an outcome was impossible to know in advance. Furthermore, some significant meta-analyses in medicine have utilised as few as eight primary studies, and large SLRs can themselves be problematic if they suffer extensive heterogeneity among studies.

4.2 Comparing data extraction and analysis approaches (Q2)

Q2.1: Analysis approach: The overall approach taken to data extraction and analysis was the same in both SLRs, in part by design (through the meta-protocol) and in part by coincidence. In particular, the six high level questions addressed by both SLRs were the same, described formally in SLR1 as:

1) Is the analysis process description complete?
    - Was the data investigated to identify outliers and to assess distributional properties before analysis?
    - Was the result of the investigation used appropriately?
    - Were the resulting estimation models subject to sensitivity or residual analysis?
    - Was the result of the sensitivity or residual analysis used appropriately?
    - Were accuracy statistics based on the raw data scale?

Primary Study	Included by SLR1	Included by SLR2
S1: Maxwell, Van Wassenhove and Dutta (1999)	Yes	Yes
S2: Brian, El Emam, Surmann, Wieczorek and Maxwell (1999)	Yes	Yes
S3: Briand, Langley and Wieczorek (2000)	Yes	Yes
S4: Jeffery, Ruhe and Wieczorek (2000)	Yes	Yes
S5: Jeffery, Ruhe and Wieczorek (2001)	Yes	Yes
S6: Wieczorek and Ruhe (2002)	Yes	Yes
S7: Lefley and Shepperd (2003)	Yes	Yes
S8: Kitchenham and Mendes (2004)	Yes	Yes
S9: Mendes and Kitchenham (2004)	Yes	Yes
S10: Mendes, Lokan, Harrison and Triggs (2005)	Yes	No
S11: Mendes, Mosley and Counsell (2003)	No	Yes
Totals	10	10

TABLE 4
Primary studies identified in each SLR

2) Is it clear what projects were used to construct each model?
3) Is it clear how accuracy was measured?
4) Is it clear what cross-validation method was used?
5) Were all model construction methods fully defined (tools and methods used)?
6) How good was the study comparison method?
   - Was the single company selected at random (not selected for convenience) from several different companies?
   - Was the comparison based on a completely independent hold out sample or on n-fold cross-validation for the within-company model?

That said, the more detailed analysis steps described for questions 1 and 6 were addressed explicitly in SLR1 whereas they were considered as part of a more holistic assessment of quality in SLR2. SLR1 assigned a quality score to all of the above factors so

that a total quality indicator could be determined. In contrast, SLR2 assigned subjective labels to the items above to provide an indicative signal of study quality.

In addressing the above questions the same quality issues were considered by both SLRs. Aspects of the cross-company data set were recorded, including the number of projects, the associated application domains and business sectors, the number of organisations and countries represented (although the latter two were not reported by SLR1), and any quality controls over the collection or verification of the data. The sizes of the specific data sets (both cross- and within-company) were also noted, with SLR1 assigning a quality score to each study depending on the size of the within-company data set. Both reviews recorded information on the actual metrics utilised in each study. SLR1 noted the size measures employed, whilst SLR2 also noted other predictor variables considered in modelling. SLR2 also reported the various modelling methods employed in each study along with details of the split of the data sets into model-building and model-validation subsets. Both reviews also noted several aspects of each study's validation approach, including the method (e.g. hold-out vs. leave-one-out) and the specific response variables considered (e.g. MRE, absolute residuals and goodness of fit).

In addition to the difference in assessment approach (SLR1's factor-based quality scoring vs. SLR2's holistic quality assessment), several specific differences in analysis were evident. SLR1 made note of the range of effort values in each cross-company data set and of the statistical test utilised to compare models. SLR2 considered the degree of completeness in each data set, the extent to which the cross-company data set was dominated by records from a small number of organisations, the degree to which the within-company data matched the cross-company data and whether any incentive existed to encourage submission to the cross-company repository. In assessing the quality of the studies SLR2 then considered the number of comparisons in each that favoured one modelling approach over another as well as those that were indifferent, and the extent to which such findings were statistically significant.

In terms of process, the two teams also used a different approach. Given that the SLR1 team comprised three people and that two of these had authored one or more primary study they randomly allocated the studies to extractor, checker and adjudicator roles,

with the caveat that a study author could not be an extractor. All three contributed to final decisions regarding the data extracted. In comparison, SLR2 used one data extractor and one checker for all ten studies.

4.3 Comparing aggregation strategies (Q3)

Q3.1: Analysis outcomes: Both reviews then considered the weight of evidence in relation to the research question that was the subject of the primary studies. In doing so, SLR1 took account of the evidence presented in each study and the testing of statistical significance as well as issues of independence between studies. SLR2 used the data on the number of comparisons in each primary study as the principal means of determining a study's outcome, along with consideration of significance testing undertaken.

Over the nine primary studies common to the two SLRs there was agreement on the interpretation of their results, although this was conveyed differently in each SLR. SLR1 provided the following summary:

- Studies S2, S3 and S6 were interpreted as showing that cross-company models were not significantly worse than within-company models
- Studies S4, S5, S8 and S9 were interpreted as showing that cross-company models were significantly worse than within-company models
- Studies S1 and S7 were interpreted as being inconclusive, primarily due to the absence of significance testing

SLR2 stated the outcomes as follows:

- Studies S2, S3 and S6 were interpreted as favouring cross-company models
- Studies S4, S5, S8 and S9 were interpreted as favouring within-company models
- Studies S1 and S7 were interpreted as being inconclusive, primarily due to the absence of significance testing

The interpretation provided by SLR1 maps more correctly to the PICO definition (particularly in relation to the nominated Intervention) and to the SLR1 expression of the research question. The SLR2 summary reflects the conclusions of the primary studies - considering studies S2, S3 and S6 in particular, their authors recommended that, in light of there being no statistically significant difference between the two approaches,

cross-company data sets should be used. This is because it enables access to more data, and more rapidly, since organisations do not have to await the accumulation of local data via project completions.

The more detailed consideration of each study's validation process afforded by SLR1 led to three studies then being excluded from the final aggregation of outcomes. Whilst these studies had met the inclusion and quality criteria they had each employed the same two data sets as a previous study, and therefore their results could not be regarded as providing independent evidence. While SLR2 had detected and noted the problem of a lack of independence in studies, this was not acted upon in terms of the associated studies being excluded from the final aggregation. SLR1 noted further that even identifying which specific data sets had been used in a study was not always straightforward.

Both SLRs identified other limitations in the primary studies, particularly in relation to data quality, model construction and experimental design, some of which had been acknowledged by the authors of the primary studies. Also noted by both SLRs was the consequent lack of strong evidence in the primary studies—individually and collectively—and the impossibility of meta-analysis due to application of different analysis and validation methods and the use of different response variables. Overall, both SLRs determined that the evidence favouring one approach or another was inconclusive. As a result both recommended that further primary studies needed to be undertaken and reported, and that this work would be most useful (from a meta-analysis point of view) if conducted based on common experimental and reporting protocols.

5 Discussion

While in many respects the two SLRs are similar, especially when considered at a fairly high level, differences between the two are evident right across the range of review activities. The differences and their impact are considered here, along with more general lessons learned from the research.

Looking specifically at the search phase (Q1), several insights can be drawn. First, while there was no single digital portal through which all the published primary studies were found, all (apart from the in-press study) were in fact 'findable' in this way. As

it turns out, for the research question being addressed here, searching a small number of sources across title / abstract / keyword entries would have enabled all relevant published studies to be found. Second, the tightly-defined and piloted search string used and verified in SLR2 was found to be far more cost-effective in terms of recall and precision than the more general version used in SLR1. SLR2 used a structured approach to constructing queries and in addition used the known papers in a rigorous capture-recapture approach. Third, the importance of the author follow-up activity was reinforced, that is, contacting the authors of primary studies to identify recently conducted relevant work not yet in the public domain. SLR2 did not do this, and missed a relevant study as a consequence. While it is correct to say that in this case this omission did not influence the overall outcome of the review (in that SLR1 reached the same overall view having included the additional study) this may not always be so.

Differences in the two review processes were more evident in the data extraction and analysis phase (Q2). While similar data quality and diversity issues were traversed, SLR1 adopted a more fine grained and quantitative approach to assessing these aspects and then took a rather holistic approach to the assessment of evidence. SLR2 took almost the opposite stance, using subjective judgements to characterise each primary study but then forming a view on the evidence in a more quantitative way. In spite of these differences, the aggregation approaches and outcomes largely coincided (Q3), with agreement on the interpretation of the nine common primary studies.

Both SLRs considered the combined weight of evidence along with issues of experimental design to arrive at an overall outcome that the evidence was not sufficient to enable a definitive conclusion with respect to the within-company/cross-company question. However, the more fine-grained approach used in SLR1 did lead that review to exclude three studies whilst SLR2's approach, which relied on the inclusion criteria (perhaps too rigidly), led to those studies being retained in the aggregation. In keeping with its intended broader treatment, SLR1 went on to identify *why* the studies' outcomes did not converge, a question that was not considered explicitly in SLR2. Similar issues for further consideration and recommendations for additional primary studies were provided by both reviews, with SLR1 making a specific recommendation regarding study independence.

In summary, the primary studies selected for analysis were almost identical in both SLRs and the conclusions reached were also the same, despite differences in review design and execution. This suggests that in this case, inconsistencies in the systematic review process did not adversely affect the stability of the outcomes. In this regard the systematic review approach proved to be reasonably robust and could be considered as a reliable research method.

There are at least three threats to the validity of our study that need to be acknowledged. First, there is a question over the extent to which the two reviews are representative of systematic reviews in general. These were rather small-scale reviews (leading to the identification of just eleven candidate primary studies), undertaken by researchers very familiar with the field, with several of the primary studies authored by three of the five reviewers. In addition, the reviewers were known to one another as members of a small research community. It is difficult to estimate the extent to which these factors may have confounded the outcomes of the work. However, as described previously, each team went to significant lengths to ensure that potential bias (arising, for example, from primary study authorship) was minimised through careful management of task allocations.

Second, it is possible that the outcome of both reviews, while the same, was wrong, and that other review teams may reach a different outcome to that reported here. Questions over consistency of process and stability of outcomes would therefore arise. What can be emphasised is that both teams generally followed standard approaches to systematic reviews. Furthermore, the two teams worked entirely independently post the definition of the meta-protocol until the two review reports had been completed. In principle then, there is some basis for applicability beyond this case.

Third, the searches of electronic databases were limited by the capabilities of (and inconsistencies in) the various search engines. Those databases that had limited capabilities in terms of performing complex multi-term searches caused both teams to use work-arounds — such actions could lead to important studies being missed. In this case we are confident that all relevant primary studies were identified, given that a wide range of sources was searched and the retrieval outcomes led to almost the same set of primary studies being selected in both SLRs. Inter-rater agreement regarding the

interpretation of individual studies was not assessed, so some degree of individual reviewer subjectivity could have influenced review outcomes.

The outcome of our investigation is two more general questions:

- Is the outcome more important than the means of achieving it?
- What has contributed to the differences between our two reviews?

So first we consider whether the outcome is more important than the process. Since systematic reviews emphasize the objective, transparent and repeatable nature of the process the answer must be "no" because without some knowledge of how the outcome was achieved it is impossible to trust the results, which is somewhat self-defeating. We believe there are a number of defining characteristics for a systematic review without which the process cannot be considered as such. These are (i) defining, reviewing and applying an explicit protocol, (ii) reviewing all papers that are retrieved by the defined search strategy or strategies, (iii) extracting and combining the results, (iv) testing the reliability of each of these processes and (v) documenting the process and process outcomes, thereby enabling other researchers to check the validity of any conclusions. Any SLR that fails to address the above five characteristics is clearly problematic. However, many of the differences between our two SLRs were at a more detailed level and therefore potentially less harmful.

Second, why did the differences that we have identified arise? Were they just a function of the relative novelty of SLRs in empirical software engineering and our consequent inexperience? Two specific issues may be relevant. In many other disciplines, most notably medicine, there is considerably more consensus in what constitutes a high quality primary study (randomized controlled trials are seen as the "gold standard") and what is an appropriate response variable. However, in software engineering we are dealing with primary studies that are diverse in quality and approach, so it was not possible to do a meta-analysis. Consequently each SLR team had to devise their own method of presenting and summarizing their results. Summarizing and aggregating results qualitatively is ill-defined, difficult and is likely to lead to some differences between the SLRs.

6 CONCLUSIONS

In this paper we have investigated the issue of the reliability of systematic reviews as a research method in the context of empirical software engineering. We conducted two reviews of the same research question in parallel teams, based on an agreed meta-protocol. Differences were observed in the activities undertaken by the two teams, but these did not appear to have any adverse effects on the reviews' conclusions, which coincided.

We therefore conclude that in this case the systematic literature review proved to be robust to differences in process and produced stable outcomes.

There are implications for both the conduct of systematic reviews and for the notion of their reliability. With respect to undertaking reviews, our experience suggests that a specific and verified search string increases the likelihood of finding studies and reduces search workload. Search strings that are more general in an attempt to be inclusive may miss studies and can lead to more effort being expended. Given the challenges encountered in searching a diverse set of sources, the provision of assistance from information retrieval/library science specialists may be beneficial. A further recommendation is that reviewers should contact the authors of the identified primary studies as a means of identifying further relevant work.

With respect to the usefulness of systematic reviews as a research method in empirical software engineering, the findings of this study suggest that they are robust to differences in people and process and that, if performed according to high-level guidelines for good practice, their outcomes can be relied upon. Those working in the domain, whether students undertaking an SLR under supervision or experienced researchers, should draw some level of reassurance from this conclusion.

Finally, it is also pertinent to ask whether a systematic review is necessary for a very specialized topic with a relatively small number of relevant papers. We observe that in the informal reviews provided in the original primary studies only 2 of the 9 studies we found referenced all papers known at the time they were written (see [5])[3]. These

3. Note, we exclude the Mendes, Lokan, Harrison and Trigg paper because that was written after we had done our searches

primary studies were also written by experts so this provides additional confirmation the value of systematic rather than informal reviews even when dealing with a small body of work.

Finally, it is also pertinent to ask whether a systematic review is necessary for a very specialized topic with a relatively small number of relevant papers. We do not have a definitive answer to this question but we note that Kitchenham et al. [5] analyzed the citations in each of the 10 primary studies they selected. Only three of the papers cited all the papers that were available when each was published, and one of those papers was co-authored by one of us (Mendes) after the search process for S1 was complete.

Further work includes a research project called EPIC, (EPSRC Project Number EP/E046983/1) on which Kitchenham is currently engaged. She is undertaking a series of case studies to assess the reliability of systematic literature reviews (SLR) and the latest version of the SLR guidelines can be found at [26].

APPENDICES

SLR1 Team Search Query

```
(software OR application OR product OR Web OR WWW OR Internet OR World-Wide
Web OR project OR development) AND (method OR process OR system OR technique
OR methodology OR procedure) AND (cross company OR cross organisation OR cross
organization OR cross organizational OR cross organisational OR crosscompany
OR cross-organisation OR cross-organization OR cross-organizational OR
cross-organisational OR multi company OR multi organisation OR multi
organization OR multi organizational OR multi organisational OR multicompany
OR multi-organisation OR multi-organization OR multi-organizational OR
multi-organisational OR multiple company OR multiple organisation OR multiple
organization OR multiple organizational OR multiple organisational OR
multiple-company OR multiple-organisation OR multiple-organization OR
multiple-organizational OR multipleorganisational OR within company OR within
organisation OR within organization OR within organizational OR within
organisational OR within-company OR within-organisation OR within-organization
```

```
OR within-organizational OR within-organisational OR single company OR single
organisation OR single organization OR single organizational OR single
organisational OR single-company OR single-organisation OR single-organization
OR single-organizational OR single-organisational OR company-specific) AND
(model OR modeling OR modelling) AND (effort OR cost OR resource) AND
(estimation OR prediction OR assessment)
```

SLR2 Team Search Query

```
(("cost model" OR "cost estimate" OR costimation OR `"cost prediction" OR
`"effort prediction" OR `"estimating cost" OR `"estimating effort") AND
("software project" OR "software product" OR "software development" OR "web
application" OR "web project" OR "web development")) AND ("company specific"
OR "company external" OR "cross company" OR "individual company" OR "multi
company" OR "multi organization" OR "multi organisation" OR "within company")
```

ACKNOWLEDGMENTS

Martin Shepperd and Steve MacDonell were partly supported by the Engineering and Physical Sciences Research Council (EPSRC) of the UK under grant EP/D003504. Barbara Kitchenham's research is funded by the EPSLRC project EP/E046983/1.

REFERENCES

[1] V. Basili and K. Freburger, "Programming measurement and estimation in the software engineering laboratory," *J. of Systems & Software*, vol. 2, pp. 47–57, 1981.

[2] H. Rombach, V. Basili, and R. Selby, Eds., *Experimental Software Engineering Issues: A Critical Assessment and Future Directions*, ser. LNCS. Berlin: Springer-Verlag, 1993.

[3] B. Kitchenham, T. Dybå, and M. Jørgensen, "Evidence-based software engineering," in *27th IEEE Intl. Softw. Eng. Conf. (ICSE 2004)*. Edinburgh: IEEE Computer Society, 2004.

[4] T. Dybå, B. Kitchenham, and M. Jørgensen, "Evidence-based software engineering for practitioners," *IEEE Software*, vol. 22, no. 1, pp. 58–65, 2005.

[5] B. Kitchenham, E. Mendes, and G. Travassos, "A systematic review of cross- vs. within-company cost estimation studies," in *10th Intl Conf Empirical Assessment in Soft. Eng. (EASE)*, 2006.

[6] S. MacDonell and M. Shepperd, "Comparing local and global software effort estimation models reflections on a systematic review," in *1st Intl. Symp. on Empirical Softw. Eng. & Measurement*, Madrid, 2007.

[7] A. Oakley, D. Gough, S. Oliver, and J. Thomas, "The policy of evidence and methodology: lessons from the eppi-centre," *Evidence and Policy*, vol. 1, no. 1, pp. 5–31, 2005.

[8] S. Oliver, G. Peersman, A. Harden, and A. Oakley, "Discrepancies in findings from effectiveness reviews: The case of health promotion for older people in accident and injury prevention," *Health and Education Journal*, pp. 66–77, 1999.

[9] A. Oakley, "Social science and evidence-based everything. the case of education," *Educational Review*, vol. 54, pp. 277–286, 2002.

[10] A. Oakley and D. Fullerton, *A systematic review of smoking prevention programmes for young people*. London: EPPI Centre, Institute for Education, 1995.

[11] L. Pauling, *How to Live Longer and Feel Better*. Atlantic Books., 1986.

[12] P. Knipschild, "Some examples of systematic reviews," *British Medical Journal*, vol. 309, p. 719 721, 1995.

[13] W. Shadish, "Author judgements about work they cite: Three studies from psychological journals," *Social Studies of Science*, vol. 25, pp. 477–498, 1995.

[14] A. Davis, O. Dieste, A. Hickey, N. Juristo, and A. Moreno, "Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review," in *14th IEEE International Requirements Engineering Conference (RE06)*. Minneapolis, MO: IEEE Computer Society, 2006, pp. 179–188.

[15] M. Blettner, W. Sauerbrei, B. Schlehofer, T. Scheuchenpflug, and C. Friedenreich, "Traditional reviews, meta-analyses and pooled analyses in epidemiology," *Intl. J. of Epidemiology*, vol. 28, no. 1, pp. 1–9, 1999.

[16] V. R. Basili, F. Shull, and F. Lanubile, "Building knowledge through families of experiments," *IEEE Transactions on Software Engineering*, vol. 25, no. 4, pp. 456–473, 1999.

[17] W. Hayes, "Research synthesis in software engineering: a case for meta-analysis," in *6th IEEE International Softw. Metrics Symp.* Boca Raton, Fl: IEEE Computer Society, 1999, pp. 143–151.

[18] L. Pickard, B. Kitchenham, and P. Jones, "Combining empirical results in software engineering," *Information & Software Technology*, vol. 40, no. 14, pp. 811–821, 1998.

[19] J. Miller, "Can results from software engineering experiments be safely combined?" in *IEEE 6th Intl. Metrics Symp.*, L. Briand, Ed. Boca Raton, Fl.: IEEE Computer Society, 1999.

[20] B. Kitchenham, P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering a systematic literature review," *Information & Software Technology*, vol. 51, no. 1, pp. 7–15, 2009.

[21] DARE, "What are the criteria for the inclusion of reviews on DARE?" Centre for Reviews and Dissemination, Tech. Rep., 2007.

[22] M. Petticrew, "Why certain systematic reviews reach uncertain conclusions," *British Medical Journal*, vol. 326, no. 7392, p. 756, 2003.

[23] J. Higgins, S. Thompson, J. Deeks, and D. Altman, "Measuring inconsistency in meta-analysis," *British Medical Journal*, vol. 327, pp. 557–560, 2003.

[24] P. Katrak, A. E. Bialocerkowski, N. Massy-Westropp, S. Kumar, and K. A. Grimmer, "A systematic review of the content of critical appraisal tools," *BMC medical research methodology [electronic resource]*, vol. 4, no. 1, p. 22, 2004.

[25] B. Kitchenham, E. Mendes, and G. Travassos, "Cross versus within-company cost estimation studies: A systematic review," *IEEE Transactions on Software Engineering*, vol. 33, no. 5, pp. 316–329, 2007.

[26] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering, version 2.3," Keele University, UK, Tech. Rep. EBSE Technical Report EBSE-2007-01., 2007.