Information and Software Technology 53 (2011) 440–455

Contents lists available at ScienceDirect

Information and Software Technology

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

Research synthesis in software engineering: A tertiary study

Daniela S. Cruzes [a,*], Tore Dybå [b]

[a] NTNU, NO-7491 Trondheim, Norway
[b] SINTEF, NO-7465 Trondheim, Norway

ARTICLE INFO

Article history:
Received 30 September 2010
Received in revised form 5 January 2011
Accepted 8 January 2011
Available online 18 January 2011

Keywords:
Evidence-based software engineering
Empirical software engineering
Systematic review
Qualitative methods
Mixed-methods

ABSTRACT

Context: Comparing and contrasting evidence from multiple studies is necessary to build knowledge and reach conclusions about the empirical support for a phenomenon. Therefore, research synthesis is at the center of the scientific enterprise in the software engineering discipline.
Objective: The objective of this article is to contribute to a better understanding of the challenges in synthesizing software engineering research and their implications for the progress of research and practice.
Method: A tertiary study of journal articles and full proceedings papers from the inception of evidence-based software engineering was performed to assess the types and methods of research synthesis in systematic reviews in software engineering.
Results: As many as half of the 49 reviews included in the study did not contain any synthesis. Of the studies that did contain synthesis, two thirds performed a narrative or a thematic synthesis. Only a few studies adequately demonstrated a robust, academic approach to research synthesis.
Conclusion: We concluded that, despite the focus on systematic reviews, there is limited attention paid to research synthesis in software engineering. This trend needs to change and a repertoire of synthesis methods needs to be an integral part of systematic reviews to increase their significance and utility for research and practice.

Contents

* Corresponding author.
E-mail addresses: dcruzes@idi.ntnu.no (D.S. Cruzes), tore.dyba@sintef.no (T. Dybå).

0950-5849/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.infsof.2011.01.004

1. Introduction

Developing software engineering (SE) knowledge is a cooperative enterprise of accumulating empirical evidence in an orderly and accurate fashion. The evidence of a particular research study cannot be interpreted with any confidence unless it has been considered together with the results of other studies addressing the same or similar questions. Comparing and contrasting evidence is necessary to build knowledge and reach conclusions about the empirical support for a phenomenon. An accurate combination of study outcomes in terms of research syntheses is, therefore, at the center of the scientific enterprise in the SE discipline. Still, it was only half a decade ago when software researchers began to pay serious attention to how to systematically locate, evaluate, synthesize, and interpret the evidence of past research studies [18,32].

Research synthesis is a collective term for a family of methods that are used to summarize, integrate, combine, and compare the findings of different studies on a specific topic or research question [7,13,39]. These methods embody the idea of making a new whole out of the parts to provide novel concepts and higher-order interpretations, novel explanatory frameworks, an argument, new or enhanced theories, or conclusions. Such syntheses can also identify crucial areas and questions for future studies that have not been addressed adequately with past empirical research. Research synthesis is built upon the observation that no matter how well designed and executed, empirical findings from single studies are limited in the extent to which they may be generalized [5]. It is, thus, a way for drawing conclusions from a *collection* of studies [39].

The key objective of research synthesis is to analyze and evaluate multiple studies and select appropriate methods for integrating [7] or providing new interpretive explanations about them [39]. If the primary studies have similar interventions and quantitative outcome variables, it may be possible to aggregate them through meta-analysis, which uses statistical methods to combine effect sizes. However, in SE, primary studies are often too heterogeneous to permit a statistical summary and, in particular, for qualitative and mixed methods studies, different methods of research synthesis are needed [17].

Although research is underway in other disciplines (e.g., [13,41,50]), there is a number of methodological questions about the synthesis of qualitative and mixed-methods findings. There are technical challenges, such as inter-rater reliability in abstracting qualitative data from individual studies or from diverse study type analyses for producing a cross-study type synthesis. There are also challenges related to the epistemological and ontological commitments underlying qualitative research, the methods of qualitative synthesis, and to methods for integrating qualitative synthesis with meta-analysis.

The aim of this article is to contribute to a better understanding of these challenges and their implications for the progress of empirical and evidence-based SE research by examining the types and methods of research synthesis employed in systematic reviews (SRs) in SE. More specifically, we seek to answer the following research questions:

1. What is the basis, in terms of primary study types and evidence that is included, in SE systematic reviews?
2. How, and according to which methods, are the findings of systematic reviews in SE synthesized?
3. How are the syntheses of the findings presented?

The remainder of this article is organized as follows: Section 2 describes the theoretical background and examines the concept of research synthesis along with an overview of synthesis and appraisal methods. Section 3 provides an overview of the research methods that were used, while Section 4 presents findings related to the research questions. Finally, Section 5 provides a discussion of the findings and the implications for research and practice. Section 6 provides the conclusions of the study.

This article is an extension of a conference paper [10], which was extended in three respects. First, Section 2 is broadened and considerably expanded to provide a much fuller account of the concept of research synthesis and its role within systematic reviews. Additionally, there is extended coverage of emerging synthesis methods as well as new material on appraisal methods. The results in Section 4 are considerably expanded with new material related to the number of studies that were included and with respect to the topics that were covered. Finally, Section 5 is expanded with a deeper discussion of the findings, their implications for theory and practice, and opportunities for future research.

2. Theoretical background

In this section, we provide the theoretical background of SRs and their relationship to evidence-based software engineering (EBSE) by contrasting the reviews to traditional literature reviews and scoping studies. Furthermore, we present definitions of research synthesis and provide an overview of the most relevant methods for synthesis of qualitative and mixed-methods evidence, followed by an overview of different ways of appraising the quality of such evidence for inclusion in SRs.

2.1. The role and definition of systematic reviews

Along with several other domains, such as healthcare, public policy, education, and management, the evidence-based paradigm has also been proposed for SE research [32], practice [18], and education [25]. The goal of this paradigm is:

> to provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision-making process regarding the development and maintenance of software [32].

In this context, evidence is knowledge obtained from findings derived from analysis of data obtained from observational or experimental procedures that are potentially repeatable and that meet the currently accepted standards of design, execution, and analysis (e.g., [26,49]). Depending on how the evidence was obtained, it can vary greatly in terms of strength. The strongest empirical evidence is obtained from rigorous methods incorporated into a study designed to have a clear, unequivocal supporting or refuting outcome. However, the evidence can be weakened by the possibility of other explanations for the results or due to weaknesses in the methods. Because the opportunity for independent assessment of the strength of evidence is a key component in any empirical study, the methods used to obtain the evidence must

Table 1
Differences between traditional reviews and systematic reviews (adapted from [38]).

Feature	Traditional reviews	Systematic reviews
Question	Often broad in scope	Often a focused research question
Identification of research	Not usually specified, potentially biased	Comprehensive sources and explicit search strategy
Selection	Not usually specified, potentially biased	Criterion-based selection, uniformly applied
Appraisal	Variable	Rigorous critical appraisal
Synthesis	Often a qualitative summary	Qualitative and/or quantitative synthesis
Inferences	Sometimes evidence-based	Usually evidence-based

be described or referenced sufficiently. Currently, depending on the methods that were used, empirical evidence in SE varies from strong and useful to weak, wrong, or irrelevant [53].

A key element of EBSE is the SR, which is a concise summary of the best available evidence that uses explicit and rigorous methods to identify, critically appraise, and synthesize relevant studies on a particular topic. The individual studies that contribute to a SR are called primary studies, while the SR itself is a form of secondary study.

Typically, a SR focuses on a well-defined question aiming to provide an answer by synthesizing the findings from a relatively narrow range of quality-assessed studies. A fundamental distinction regarding the objective of such reviews is whether they attempt to provide knowledge support or decision support [44]. A SR directed at knowledge support will typically bring together and synthesize research evidence on a particular topic, while a SR aimed at decision support will be more specific and include analytical tasks to help make a decision within a particular context [35]. In reviews for knowledge support, approaches may be prioritized to avoid bias, whereas for supporting a decision, avoiding bias may be necessary but not sufficient and the reviewer must also be explicit about the basis of the judgments that are inevitably made [44]. Furthermore, when a review aims to provide decision support, it may need to include non-research evidence and possibly use various modeling and simulation methods, which will affect the methodological focus of the SR.

Both traditional literature reviews and SRs are retrospective, observational studies. Therefore, they are subject to systematic and random errors. The quality of a review depends on the extent to which scientific and transparent review methods were used to minimize error and bias. In addition to research synthesis, these methods are the key feature that distinguishes traditional reviews from SRs (see Table 1). There is, however, a discussion on how explicit and transparent a SR can be [23] as well as to what extent it is possible, or even relevant, to follow predefined procedures [15].

Another form of secondary study is the scoping study, which (unlike SRs) is less likely to attempt to address very specific research questions or to assess the quality of the included studies [2,12]. Typically, scoping studies address broader topics and are "designed to provide an initial indication of the size and location of the literature relating to a particular topic as a prelude to a comprehensive review" or "to establish how a particular term is used in what literature by whom and for what purpose" [1]. As a consequence, scoping studies tend to draw on a diverse range of qualitative and quantitative research, and non-research sources that cannot be easily appraised or synthesized. The lack of research synthesis and quality appraisal in scoping studies is what distinguishes these studies from SRs.

2.2. Synthesis of qualitative and mixed-methods evidence

Following the conventions in Mays et al. [35], we use the term "systematic review" to describe the whole process of bringing together evidence from a range of sources, and the term "synthesis" is used to describe the specific procedures within the SR that are used to combine the evidence from individual, primary studies.

There are three definitions of synthesis that were applicable to our purpose (Merriam-Webster's dictionary). According to the first definition, synthesis is the combination of parts or elements to form a whole. Synthesis can also be defined as the dialectic combination of thesis and antithesis into a higher stage of truth. Finally, synthesis can be defined as the combination of often diverse conceptions in a coherent whole.

Noblit and Hare employed these same three distinctions by stating that the first definition explains synthesis of directly comparable studies as 'reciprocal translations'; the second explains studies that stand in opposition to one another through 'refutational translations' (or forms of resolution); and the third explains the synthesis of studies that may represent a line of argument (or forms of reconceptualization) [39].

Noblit and Hare further distinguished between two synthesis approaches: integrative and interpretive [39]. Integrative synthesis combines or summarizes data to create generalizations [7]. It involves the quantification and systematic integration of data through techniques such as meta-analysis (which is concerned with the assembly and pooling of specific data) or less formal techniques (such as providing a descriptive account of the data).

Interpretive synthesis achieves, on the other hand, subsumes the concepts identified in the primary studies into a higher-order theoretical structure. The primary concern is with the development of concepts and of theories integrating those concepts. Therefore, an interpretive synthesis will avoid specifying concepts before the synthesis and ground the concepts in the data reported from the primary studies [13]. While most forms of synthesis could be characterized as being either primarily interpretive or primarily integrative, every integrative synthesis will include elements of interpretation, and every interpretive synthesis will include elements of integration.

The traditional view of research synthesis is the integrative, quantitative approach with an emphasis on the accumulation of data and analysis through meta-analysis. Meta-analysis is a form of additive synthesis that combines the numerical results of controlled experiments, estimates the descriptive statistics, and explains the inconsistencies of effects as well as the discovery of moderators and mediators in research findings [22,33,48]. The purpose of meta-analysis is to aggregate the results of studies to predict future outcomes for situations with analogous conditions. However, in order for meta-analyses to be convincingly performed, the experiments must represent results from a single underlying effect rather than a distribution of effects.

As empirical research has matured, there has been an increasing awareness that other research designs besides controlled experiments are necessary to understand more complex, and often more relevant, questions about what works, in which situations, and for whom. This has spurred a growing interest in qualitative research, which, in turn, has drawn attention to SRs and synthesis methods that can include evidence from diverse study types [17]. Therefore, contrary to the purely integrative, quantitative method we find several methods for conducting interpretive syntheses of

Table 2
Overview of methods for the synthesis of qualitative and mixed-methods evidence.

Synthesis method	Description
Narrative synthesis [47]	A defining characteristic of narrative synthesis is the adoption of a narrative (as opposed to statistical) summary of the findings of primary studies. It is a general framework of selected narrative descriptions and ordering of primary evidence with commentary and interpretation combined with specific tools and techniques that help to increase transparency and trustworthiness. Narrative synthesis can be applied to reviews of quantitative and/or qualitative research
Meta-ethnography [39]	Meta-ethnography resembles the qualitative methods of the primary studies. It aims to synthesize by induction, interpretation, and translational analysis of the primary studies to understand and transfer ideas, concepts, and metaphors across different studies. The product of a meta-ethnographic synthesis is the translation of studies into one another, synthesizing the translations to identify concepts that go beyond individual accounts to produce a new interpretation. Interpretations and explanations in the primary studies are treated as data, and are translated across several studies to produce a synthesis
Grounded theory [8,21]	Grounded theory is a primary research approach that describes methods for qualitative sampling, data collection, and data analysis. It includes simultaneous phases of data collection and analysis, the use of the constant comparison method, the use of theoretical sampling, and the generation of new theory. It treats study reports as a form of data on which analysis can be conducted to generate higher-order themes and interpretations
Cross-case analysis [36]	Cross-case analysis includes a variety of devices, such as tabular displays and graphs, to manage and present qualitative data, without destroying the meaning of it, through intensive coding. It includes meta-matrices for partitioning and clustering data in various ways. Evidence from each primary study is summarized and coded under broad thematic headings. Evidence is then summarized within themes across studies with a brief citation of primary evidence. Commonalities and differences between the studies are noted
Thematic analysis/synthesis [3]	Thematic analysis is a method for identifying, analyzing, and reporting patterns (themes) within data. It minimally organizes and describes the data set in rich detail and frequently interprets various aspects of the research topic. Thematic analysis can be used within different theoretical frameworks, and it can be an essentialist or realist method that reports experience, meanings, and the reality of participants. It can also be a constructionist method, which examines the ways in which events, realities, meanings, experience, and other aspects affect the range of discourses. Thematic analysis has limited interpretative power beyond mere description if it is not used within an existing theoretical framework
Content analysis [20]	Content analysis is a systematic way of categorizing and coding studies under broad thematic headings by using extraction tools designed to aid reproducibility. Occurrences of each theme are counted and tabulated. The frequencies of data are determined based on precise specifications of categories and the systematic application of rules. However, the frequency-counting techniques of content analysis may fail to reflect the structure or importance of the underlying phenomenon, and the results may be oversimplified and count components that are easy to classify and count rather than the components that are truly important [13]
Case survey [56]	The case survey method is a formal process for systematically coding relevant data from a large number of case studies for quantitative analysis. A set of structured closed-ended questions is used to extract data so that the answers can be aggregated for further analysis. Qualitative evidence is converted into a quantitative form, thereby synthesizing both qualitative and quantitative evidence. Each primary study is treated as a specific case. Study findings and attributes are extracted using closed-form questions for increased reliability, while survey analysis methods are used on the extracted data
Qualitative comparative analysis (QCA) [45]	The qualitative comparative analysis method is a mixed synthesis method that analyzes complex causal connections using Boolean logic to explain pathways to a particular outcome based on a truth table. The Boolean analysis of necessary and sufficient conditions for particular outcomes is based on the presence/absence of independent variables and outcomes in each primary study
Aggregated synthesis [19]	Aggregated synthesis is an interpretive process that contains elements of both grounded theory and meta-ethnography. It attempts to preserve the context of the original research while enhancing the generalizability of the original studies by building mid-range theories. The goal of aggregated synthesis is thus theory development and cumulative knowledge, which can explain as well as predict certain behaviors
Realist synthesis [43]	Realist synthesis is a theory-driven approach that encompasses quantitative and/or qualitative research from any kind of evidence by focusing on explaining how these interventions work and why they fail to work in particular contexts. Data extraction in realist synthesis takes the form of an interrogation of baseline inquiries for information on what works for whom under what circumstances. The theory that underlies a particular intervention is central to this method
Qualitative metasummary [50]	Qualitative metasummary is a quantitative oriented aggregation of qualitative findings. The goal is to discern the frequency of each finding and to find in higher frequency findings the evidence of replication foundational to validity in quantitative research and to the claim of having discovered a pattern or theme in qualitative research
Qualitative metasynthesis [50]	Qualitative metasynthesis is an interpretive integration of qualitative findings that are in the form of interpretive syntheses of data; either conceptual/thematic descriptions or interpretive explanations. Metasyntheses offer novel interpretations of findings derived from considering all the studies in a sample as a whole. Validity does not reside in replication logic, but rather in interpretation
Meta-study [41]	Meta-study involves the analysis of theories, methods, and findings in qualitative research as well as the synthesis of these insights into new ways of thinking about phenomena. The goal is to transform the accumulation of findings into a legitimate body of knowledge with the ultimate aim of both generating new theory and informing practice. The method is unique in the extent to which it focuses on the importance of understanding the findings in terms of the methods and theories that drive them

Table 3
Overview of approaches for the appraisal of qualitative and mixed-methods evidence.

Appraisal approach	Description
Critical appraisal skills programme [4]	A number of tools were developed under the CASP to help with the process of critically appraising articles of diverse types of research. The qualitative appraisal tool presents ten questions that deal very broadly with some of the principles or assumptions that characterize qualitative research. The following three broad issues are included: rigor, credibility, and relevance
Long and Godfrey [34]	A relatively lengthy tool that incorporates both descriptive and evaluative elements. It includes 34 questions across four key areas: characteristics of the study (study type, sampling and setting), how the study was done (rationale for the choice of setting, sample, data collection and analysis), research ethics, and policy and practice implications
Spencer et al. [54]	This framework was developed by the UK Cabinet Office to support work in Departments. It provides a guide for assessing the credibility, rigor, and relevance of individual research studies. The framework contains 18 appraisal questions related to nine key areas: findings, design, sample, data collection, analysis, reporting, reflexivity and neutrality, ethics, and auditability
Walsh and Downe [55]	Walsh and Down reviewed a number of frameworks for appraising qualitative research. They appraised and synthesized the resulting frameworks to make a practice-oriented checklist. The final checklist included 53 items related to eight key areas: scope and purpose, design, sampling strategy, analysis, interpretation, reflexivity, ethical dimensions, and relevance and transferability

qualitative and heterogeneous research [13], which has been presented in the form of a narrative.

There is, however, a debate about the appreciation and legitimization of qualitative synthesis. On one hand, proponents of qualitative synthesis view it as essential for achieving the goals of the evidence-based paradigm. On the other hand, proponents against the synthesis of qualitative research argue that there are epistemological and ontological commitments that are assumed to underlie qualitative research. They also argue that qualitative research is as resistant to synthesis as poems are [50].

A more pragmatic view was taken by Estabrooks et al. [19], who argued that the synthesis of multiple studies could result in a contribution to theory building that is more powerful than any single study. According to this view, the synthesis of qualitative evidence could allow for the construction of larger narratives and more general theories and, therefore, overcome the problem of isolation associated with qualitative research and allow for cross-study themes or higher-order analytical categories to be developed [13].

Accordingly, a number of different methods have been proposed for the synthesis of qualitative and mixed-methods findings, many of which were based on approaches used in primary research [13,44]. Some of the methods maintain the qualitative form of the evidence, such as meta-ethnography, and some methods involve converting qualitative findings into a quantitative form, such as content analysis and qualitative metasummary. Table 2 outlines some of these methods. This is by no means meant to be an exhaustive list; there are numerous other methods with slightly different terminology as well as different epistemological and ontological foundations. Ultimately, a number of factors, including the research question, the anticipated number of primary studies, and the knowledge and expertise of the team undertaking the review, will influence the choice of method.

2.3. Appraisal of qualitative and mixed-methods evidence

A final, but widely debated, concern is the issue of how, or whether, to appraise qualitative studies for inclusion in a SR. The process of quality appraisal is important because the quality of studies or other evidence may affect both the results of the individual studies and the conclusions derived from the synthesis. While some authors have argued against the use of standard quality criteria, or "criteriology" [52], for evaluating qualitative studies, others have accepted the need for transparent approaches for appraisal. Such quality appraisals are typically used to establish a minimum quality threshold for the inclusion of studies, to discriminate between overall contributions of studies, or to gain a better understanding of the strength of evidence [16].

Assessing the quality of a study is not straightforward, however, as there is no general, agreed upon definition of "quality" [29]. There are also common problems in appraising the quality of published research because journal articles and, in particular, conference papers rarely provide enough detail of the methods due to space limitations in journal volumes and conference proceedings. Despite these difficulties, and the lack of consensus regarding the quality appraisals of qualitative studies, there are a multitude of guidelines, tools, and checklists available that can be used to assess the quality of primary, qualitative studies. Table 3 outlines some of these methods.

When it comes to appraising the quality of SRs, several tools are also available (see [16] for examples). The York University, Centre for Reviews and Dissemination (CRD) Database of Abstracts of Reviews of Effects is particularly relevant (DARE, 2010). CRD undertakes high quality SRs that evaluate the research evidence on health and public health questions of national and international importance. The findings of CRD reviews are widely disseminated and have impacted on health care policy and practice both in the UK and internationally.

The CRD databases have become a key resource for health professionals, policymakers and researchers around the world. The databases assist decision makers by systematically identifying and describing SRs and economic evaluations, appraising their quality and highlighting their relative strengths and weaknesses. DARE is a CRD database that contains more than 15,000 abstracts of SRs, which includes over 6000 quality assessed reviews. Each month, thousands of citations are screened to identify potential SRs. These citations are independently assessed by two researchers for inclusion in DARE by using the following criteria:

1. Were inclusion/exclusion criteria reported?
2. Was the search adequate?
3. Were the included studies synthesized?
4. Was the validity of the included studies assessed?
5. Are sufficient details about the individual included studies presented?

Reviews included in DARE meet at least four of the five criteria (Criteria 1–3 are mandatory). Kitchenham et al. used the DARE criteria to evaluate the quality of SRs in the SE field [27,28], but did not include the mandatory Criterion 3, regarding synthesis. Likewise, Kitchenham and Charters did not include the synthesis requirement in their reference to DARE in the guidelines for performing systematic literature reviews in SE [31]. However, Criterion 3 is critically important for the DARE evaluation because it is mandatory for a SR to synthesize primary studies. Without such synthesis, the secondary study will be, at best, a scoping study.

3. Research methods

This study is a tertiary review [31] to assess the types and methods of research synthesis in systematic reviews in SE. Based on the research questions in Section 1, we used the ISI Web of Knowledge to conduct a search in the following databases:

- Science Citation Index Expanded (SCI-EXPANDED).
- Social Sciences Citation Index (SSCI).
- Arts and Humanities Citation Index (A&HCI).
- Conference Proceedings Citation Index-Science (CPCI-S).
- Conference Proceedings Citation Index-Social Science and Humanities (CPCI-SSH).

We searched within all the 'computer science' subject areas for all full proceedings papers and journal articles published from 1st January 2005 until 31st July 2010 that contained the term 'systematic review' in the title:

```
Title=(systematic review)
Refined by: Subject Areas=(COMPUTER SCIENCE,
   INFORMATION SYSTEMS OR COMPUTER SCIENCE,
   INTERDISCIPLINARY APPLICATIONS OR COMPUTER
   SCIENCE, SOFTWARE ENGINEERING OR COMPUTER
   SCIENCE, THEORY & METHODS)
Timespan=2005-2010. Databases=SCI-EXPANDED, SSCI,
   A&HCI, CPCI-S, CPCI-SSH
```

Due to previously reported inadequacies with the ACM Digital Library, and the fact that ISI Web of Knowledge does not index ACM proceedings papers [17], we performed a separate search in the ACM Digital Library for such papers. Also, because Kitchenham et al. reviewed the status of EBSE since 2004 in two tertiary

Table 4
Publication year.

Year	#	Percent (%)
2005	2	4.08
2006	5	10.20
2007	7	14.29
2008	12	24.49
2009	12	24.49
2010	11	22.45

Table 5
Publication venues.

Publ.	#	Percent (%)
IST	22	44.9
Conference	12	24.5
TSE	5	10.2
Workshop	4	8.2
EMSE	2	4.1
Others	4	8.2

reviews that focused on articles describing secondary studies [27,28], we examined the articles included in those reviews for possible inclusion in the current study as well.

We restricted the start of our search to the beginning of 2005 because we would not expect earlier papers to be influenced by the seminal papers on EBSE [18,32] or the procedures for undertaking systematic reviews [30]. We also restricted the search to secondary studies that, themselves, claimed to be SRs; either by stating so in the title or by explicitly referencing Kitchenham and Charters's guidelines for conducting SRs [31], in the case they were identified by the tertiary reviews of Kitchenham et al. but did not include the search term in the title.

Our search procedure retrieved 84 articles of which two were duplicates and two were conference paper versions of later journal articles, thus, resulting in 80 unique studies. Of these, we excluded 40 studies that were either short papers, were clearly outside the subject area of SE (e.g., studies within medical informatics), or that were tertiary reviews or lessons learned reports on conducting systematic reviews. Of the remaining 40 studies, we were not able to retrieve one of the papers. In addition, we added 10 more SRs from Kitchenham et al.'s tertiary reviews, thus leaving 49 articles for data extraction and analysis (see Appendix A).

We extracted the following data from each study:

- The source and full bibliographic reference.
- The main topic area, overall aim and research questions.
- How the authors perceived synthesis within the context of a systematic review.
- The databases used to search for primary studies.
- The number and time span of the primary studies included.
- Whether the authors mentioned the types of primary studies included, and if so, which types.
- Whether a separate section on synthesis method(s) was included and whether they explicitly mentioned a method of synthesis with a corresponding reference.
- Quality assessment approach and its use.
- Whether the authors synthesized findings according to the types of primary studies included or according to the quality of the studies.
- The types and methods of synthesis used.
- How the synthesis was performed and presented.

The first author (Cruzes) extracted and categorized the data while the second author (Dybå) checked the extraction. Whenever we had a disagreement, we discussed the issues until we reached an agreement. To answer our research questions, we analyzed the extracted data both qualitatively and quantitatively. Although we included a short narrative description of the results, the majority of the results were tabulated to present an overview of the findings as well as basic information about each study. As such, our study is a scoping study [2] that 'maps' out the SR literature in the SE field based on the types and methods of research synthesis employed.

4. Findings

The number of SRs in SE has grown from one published study in 2005 (S25), to 12 studies in 2009, and to 11 until the middle of 2010, for a total of 49 studies (Table 4 and Fig. 1). The journal *Information and Software Technology* (IST) was the first to introduce systematic reviews to its readers and to publish them. This journal published 22 of the 49 reviews in our sample (Table 5 and Fig. 1). Papers from conferences represented one fourth of the SRs (12/49), while *IEEE Transactions on Software Engineering* (TSE) published 5 of the 49 SRs. The remaining 10 SRs were published in various SE and SE-related journals and proceedings.

There was a diversity of topics addressed in the SRs. As shown in Table 6, SRs in SE can be classified into 21 broad research areas, reflecting topics in which empirical research in SE has increased during the last years thus making systematic reviews possible. Requirements engineering (6/49), software design (5/49) and experimental methods in SE (5/49) were the three topic areas with most studies.

Most authors claimed that the rationale behind their SRs was that it is a formalized, repeatable process for systematically searching a body of literature to document the state of knowledge on a particular subject. They also claimed that the benefit of performing a SR is that it provides researchers with more confidence that they have located as much relevant information as possible. Although search strategies and data extraction methods were typically described in detail, few studies mentioned synthesis methods. Some studies did not even mention the synthesis part at all, but often

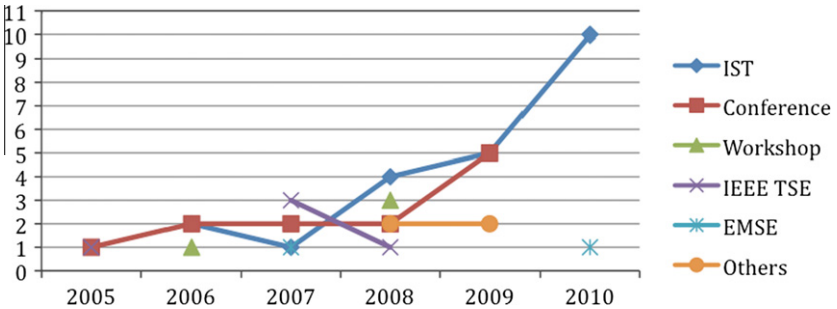

Fig. 1. Publication years and venues.

Table 6
Main topic areas in SE systematic reviews.

Main topic area	Studies
Agile software development	S5, S12
Aspect-oriented programming	S45
Distributed software development	S15, S19, S47
Domain analysis	S22
Estimation models	S16, S21, S35, S43
Experimental methods in SE	S6, S17, S18, S38, S44
Global software engineering	S31
Knowledge management in SE	S3
Motivation in SE	S2, S28
Product lines software development	S20, S25, S33
Requirements engineering	S4, S14, S26, S29, S32, S48
Reuse	S36
Software design	S8, S23, S30, S39, S42
Computer games	S49
Software maintainability	S34
Software measurement	S9, S40
Software process	S27, S41
Technology acceptance model	S46
Testing	S1, S7
Theory use in SE	S10, S11
Web development	S13, S24, S37

covered only the selection and extraction processes. Other studies used terms such as 'interpretation', 'summarize', 'inferences', 'aggregation', and 'analyze' to refer to the synthesis. In the remainder of this section, we present the findings of our study according to the research questions stated in Section 1.

4.1. What was the basis for the reviews?

Sample size, publication year, and the types of primary studies included varied among the studies. The identification process of appropriate publications was typically described. Manual searches were included in more than 60% of the studies. About 70% of the searches were based on retrievals from IEEE eXplore and the ACM Digital Library. Other databases that appeared in around 20–50% of the papers were ISI, Google Scholar, Inspec, Ei Compendex and Science Direct. A few studies did not describe their search procedures or searches in detail.

All of the studies had information on the sample size. The number of primary studies included in the SRs ranged from 10 to 304 with a median of 54. However, eight articles were not clear about the number of primary studies that were examined when searching for studies.

The majority of the SRs, 63.3%, classified the primary studies based on the type of intervention, but only 20.4% of the SRs used this classification for the synthesis of primary studies (Table 7).

In addition to empirical research studies, and contrary to the aims of systematic review research [24], most of the SRs also included non-empirical primary studies. Some SRs even based their findings on statements and the author's own experience. For example, in S12, 16 out of 20 studies were lessons learned reports based on expert opinion, while more than 90% of the primary studies in S20 were based on claims and expert opinions without any corresponding empirical data. Furthermore, 117 out of the 173 primary studies included in S25 were advocacy research, proof of concept, or experience reports, whereas half of the studies included in S13 did not conduct any empirical evaluations.

Table 7
Basis for systematic reviews in SE.

	Yes (%)	No (%)
Classified the types of studies (e.g., case studies, experiments, surveys)?	63.3	36.7
Synthesis based on study types?	20.4	79.6
Quality appraisal?	34.7	65.3
Quality appraisal used for the exclusion of papers?	6.1	93.9
Quality appraisal used for synthesis?	14.3	83.7

The extent of quality appraisal of primary studies included in the SRs in SE was very low (Table 7). Only 34.7% of the SRs appraised the quality of the primary studies, while as few as 6.1% used such appraisals to exclude the primary studies with low quality from entering the SR. Furthermore, only 14.3% of the SRs used the quality appraisal in the synthesis of the studies or to weigh the evidence from different sources.

Since Kitchenham et al. did not include the mandatory criterion regarding synthesis in their appraisal of the SRs in SE [27,28], we reappraised the studies by using all five DARE criteria to ascertain whether the studies were in accordance with the DARE inclusion criteria or not (Table 8). From this appraisal, we observed that 63.3% of the SRs did not satisfy the DARE criteria. Of the three mandatory criteria, Criterion 3, which was omitted by Kitchenham et al., was the main criterion that would exclude the studies from inclusion in the DARE repository.

Finally, if the validity of the studies was not assessed (Criterion 4), there should be sufficient details in the report of the SR that allow for independent assessment to be completed. Normally, this requirement would imply including a full list of bibliographic information for the included primary studies. As seen in Table 8, 20.4% of the SRs did not provide this information.

4.2. How were the findings synthesized?

In this section, we divided the analysis into three perspectives: (1) the methods of synthesis as described by the authors, (2) the methods of synthesis as we classified them (based on the original literature and the actual synthesis described in the papers), and (3) the appropriateness of the synthesis methods according to the goal of the SR.

4.2.1. Methods of synthesis as described by the authors of SRs

We classified the SRs according to the methods of synthesis in their study as indicated by the authors (Table 9). In almost half of the SRs, there was no indication of a synthesis method being followed. In a few SRs, the synthesis methods were explained in detail, occasionally with illustrating descriptions (e.g., S2, S5, S6, and S18). Some studies provided little information on their procedures, while others were more detailed. In some SRs, instead of explaining the synthesis procedures, the authors explained how the extraction of the data was performed. In addition, although half of the papers contained a synthesis section, only ten (20.4%) of the SRs included a reference for the method of synthesis (Table 9), although the authors did not always follow the method described in these references.

We found only six original references to synthesis methods in which there was a detailed explanation and definition of the method. Examples of these methods references included Noblit and Hare's meta-ethnography [39], Ragin's qualitative comparative method [45], Miles and Huberman's methods [36], Strauss and Corbin's constant comparison method [8], and Cohen's post hoc power calculations [6]. Four of the references we found for methods of synthesis were not adequate, as they referred to papers that did not define the methods of synthesis. For example, in one SR, the authors provided a reference on narrative synthesis that was a paper on systematic mapping studies that did not discuss narrative synthesis methods. In another SR, we found a reference to what the authors called a "grounded approach." We checked the reference and could not find a definition of the method mentioned by the authors.

Table 8
Quality appraisal of systematic reviews in SE using the DARE criteria.

SR	(1) Were inclusion/exclusion criteria reported?	(2) Was the search adequate?	(3) Were the included studies synthesized?	(4) Was the validity of the included studies assessed?	(5) Are sufficient details about the individual studies presented?	Inclusion in DARE?
S1	Yes	Yes	Yes	Yes	Yes	Yes
S2	Yes	Yes	Yes	Yes	Yes	Yes
S3	Yes	Yes	Yes	No	Yes	Yes
S4	Yes	Yes	Yes	No	Yes	Yes
S5	Yes	Yes	Yes	Yes	Yes	Yes
S6	Yes	Yes	Yes	No	No	No
S7	Yes	Yes	Yes	Yes	Yes	Yes
S8	Yes	Yes	No	No	Yes	No
S9	Yes	Yes	No	No	No	No
S10	Yes	Yes	No	No	Yes	No
S11	Yes	Yes	No	No	Yes	No
S12	Yes	Yes	Yes	No	Yes	Yes
S13	Yes	No	No	No	No	No
S14	Yes	Yes	Yes	Yes	Yes	Yes
S15	Yes	No	No	No	Yes	No
S16	Yes	Yes	No	No	Yes	No
S17	Yes	Yes	No	Yes	Yes	No
S18	Yes	Yes	Yes	No	No	No
S19	Yes	Yes	Yes	No	No	No
S20	Yes	Yes	Yes	Yes	Yes	Yes
S21	Yes	Yes	Yes	Yes	Yes	Yes
S22	Yes	Yes	No	No	Yes	No
S23	Yes	No	No	No	Yes	No
S24	Yes	No	No	No	No	No
S25	Yes	Yes	No	No	No	No
S26	Yes	No	Yes	No	Yes	No
S27	Yes	Yes	Yes	No	Yes	Yes
S28	Yes	Yes	Yes	No	Yes	Yes
S29	Yes	No	Yes	Yes	Yes	No
S30	Yes	No	Yes	No	Yes	No
S31	Yes	Yes	Yes	No	Yes	Yes
S32	Yes	Yes	No	Yes	Yes	No
S33	Yes	Yes	No	No	Yes	No
S34	Yes	Yes	Yes	Yes	Yes	Yes
S35	Yes	Yes	Yes	Yes	Yes	Yes
S36	Yes	No	Yes	No	Yes	No
S37	Yes	Yes	No	Yes	No	No
S38	Yes	Yes	No	No	Yes	No
S39	Yes	Yes	Yes	No	Yes	Yes
S40	Yes	Yes	No	No	Yes	No
S41	Yes	No	No	Yes	No	No
S42	Yes	Yes	No	No	Yes	No
S43	Yes	No	No	No	Yes	No
S44	Yes	Yes	No	No	No	No
S45	Yes	Yes	Yes	Yes	Yes	Yes
S46	Yes	Yes	Yes	Yes	Yes	Yes
S47	Yes	Yes	No	No	Yes	No
S48	Yes	Yes	No	Yes	Yes	No
S49	No	Yes	No	No	Yes	No
% Yes	98.0%	79.6%	51.0%	34.7%	79.6%	36.7%
% No	2.0%	20.4%	49.0%	65.3%	20.4%	63.3%

4.2.2. Methods of synthesis according to the original references

Table 10 shows which synthesis methods the authors actually performed according to the original description of the methods (see Table 2). When the authors attempted to describe their synthesis method, they were mostly correct, but many authors did not use the appropriate terminology for the method. We categorized the studies that did not describe a synthesis method as scoping studies. In addition, the "classification analysis" that was mentioned by some authors also fell into the scoping study category.

Twenty-four of the 49 studies were categorized as scoping studies (Table 10). These studies involved the analysis of a wide range of research and non-research material to provide an overview or mapping about a specific topic or field of interest. The main reason for a study to be in this category was that the study did not synthesize evidence from the area in focus, but provided an overview of the subject area. S44 is a good example of such a scoping study. This study reported how controlled experiments in SE are conducted and the extent to which relevant information is reported. Based on this study, other secondary studies could be performed, such as S11 (scoping study), S17 (scoping study), S18 (meta-analysis), and S38 (scoping study).

Another example of a scoping study is S8. In this study, ten domain design approaches were selected from the literature with a brief chronological description of the selected approaches. The authors performed a mapping of the completeness of the domain design approaches and evaluated the key points and drawbacks of the approaches that were reviewed, but they did not synthesize any findings from the primary studies. S10 is also a scoping study; they referenced Robson [46] for the synthesis method, and claimed that the study was a "grounded approach." However, we did not find the steps for this method described in the cited book or in the literature on grounded theory (e.g., [8]). The authors identified

Table 9
Methods of Synthesis as Described by the Authors of SRs.

Method as described by the authors	Studies	Examples of Ref to the method	Original reference to the method?
Mapping	S8	None	
Classification analysis	S3, S7, S9, S11, S13, S16, S17, S26, S15.	According to ISO/IEC 12207 (S15)	No
Thematic synthesis	S2, S12, S28	None	No
Descriptive evaluation	S1	None	No
Vote counting	S45, S36, S46	L.M. Pickard, B.A. Kitchenham, P.W. Jones, Combining empirical results in software engineering, IST 40 (1998) 811–821; P. Mohagheghi, R. Conradi, An empirical investigation of software reuse benefits in a large telecom product, TOSEM 17(3) (2007) 1–31 (S36)	No
Narrative synthesis	S31	K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson, Systematic mapping studies in software engineering, Proc. EASE'08, 2008, pp 71–80 (S31)	No
Reciprocal translational analysis	S39	Noblit and Hare [39] (S39)	Yes
Grounded approach	S10	Robson [46] (S10)	No
Quantitative approach	S35	None	No
Comparative analysis	S4	Ragin [45] (S4)	Yes
Content analysis	S38, S40	K. Krippendorff, Content Analysis: An Introduction to Its Methodology, second ed., Sage, 2004; U.H. Graneheim, B. Lundman, The Challenge of qualitative content analysis, Nurse Education Today, 24 (2004) 105–112 (S38)	Yes
Quantitative aggregation based on standardized effect sizes	S18	R.B. Kline, Beyond Significance Testing, Reforming Data Analysis Methods in Behavioral Research, American Psychological Association, Washington DC, 2004; Rosenthal and DiMatteo [48] (S18)	Yes
Post-hoc power calculations	S6	Cohen [6] (S6)	Yes
Meta-ethnography	S5	Miles and Huberman [36]; Noblit and Hare [39] (S5)	Yes
Not explicit about the method	S34, S27, S14, S21, S23, S43, S48, S22, S24, S32, S33, S37, S41, S42, S44, S47, S49, S19, S20, S29, S30, S25		No

Table 10
Methods of Synthesis as described by the authors (rows) vs. by the literature (columns).

	Scoping study	Thematic synthesis	Narrative synthesis	Comparative analysis	Meta-analysis	Case survey	Meta-ethnography
Mapping (Scoping)	S8						
Classification analysis	S9, S11, S13, S15, S16, S17	S3	S7, S26				
Thematic synthesis		S2, S12, S28					
Descriptive evaluation			S1				
Vote counting			S45, S36	S46			
Narrative synthesis			S31				
Reciprocal translational analysis			S39				
Grounded approach	S10						
Quantitative approach				S35			
Comparative analysis				S4			
Content analysis	S38, S40						
Quantitative aggregation based on standardized effect sizes					S18		
Post-hoc power calculations					S6		
Meta-ethnography							S5
Not explicit about the method	S22, S23, S24, S25, S32, S33, S37, S41, S42, S43, S44, S47, S48, S49	S19, S20, S29, S30	S27, S34	S21		S14	
Percentage	49.0%	16.3%	18.4%	8.2%	4.1%	2.0%	2.0%

the number of articles using a theory in specific ways then applied a second categorization process to analyze how each theory was used. However, they did not synthesize any of their findings.

We classified eight SRs as thematic syntheses (Table 10), but none of the studies referenced any methodological paper for the method used for synthesis. One example of thematic synthesis is S12 in which the authors identified the themes emanating from the findings reported in each of the primary studies included in their SR. The authors presented frequencies for the number of times each theme was identified in different studies. Subsequently, the authors synthesized the findings from the primary studies in order to find the answers to their research questions according to the themes.

In another example of thematic synthesis, the authors of S29 conducted a SR with the aim of identifying and classifying types of requirement errors into a taxonomy to support the prevention and detection of errors. The authors first described the errors and their characteristics based on the research questions posed for the SR. The authors then organized the errors into a taxonomy with the intent of addressing limitations in the existing quality

improvement approaches. Finally, they synthesized and described each error class along with the specific errors that made up that class as well as the references that backed up the findings.

Nine SRs were categorized as narrative syntheses (Table 10). In S1, the authors described the evidence from each study in a chronological way and then discussed some differences and possible explanations for the differences in the results. The authors performed the narrative synthesis according to the categories found for non-functional search-based software testing. In S7, the authors analyzed the empirically evaluated relationships between regression test selection techniques by showing the results of the studies in graphs combined with narrative synthesis. In S36, the authors performed what they described as a modified approach for vote-counting and used the results from the SR to describe how the intervention worked, why it worked, and for whom it worked; approaching a realist review.

Four SRs were categorized as comparative analyses. One of them referenced Ragin [45], while the others did not explicitly mention their approach. Neither of them fully applied the method as described by Ragin because they did not use the concept of a truth table and Boolean algebra. In study S4, the authors provided a table in which they compared the studies providing evidence for and against a certain result as well as any relevant aggregation-related issues. In study S21, the authors provided tables in which they identified a variety of options for performing a comparative study of cross-company and within-company estimation models. They considered the pros and cons of each option and identified which primary studies (if any) used that option. Based on the results of this study, and on their own experience, they provided a comparison table with a summary of advice based on the evidence in favor of and against each item. In S46, the authors performed vote-counting to compare the studies because they could not undertake an effect-size meta-analysis in their sample of studies.

Two SRs were categorized as meta-analyses. The two reviews were from the same research group, and they used the same set of primary studies (103 experimental papers). One of the reviews (S6) analyzed the statistical power, and the other (S18) analyzed the effect size in SE experiments. Following the post hoc method, S6 aggregated the power of each test in the primary studies in relation to Cohen's definitions of small, medium, and large effect sizes [6]. Study S18 cited various meta-analysis references (e.g., [48]) and performed a meta-analysis using Hedges' *g* as the standardized effect-size measure.

Study S14 was the only example of a case survey. The goal of the SR was to provide an objective view of what technologies were present in requirements engineering research and to what extent papers describing these technologies provided decision support for practitioners seeking to adopt such technologies. Each of the research questions in S14 was mapped to a data extraction form in form of a closed-ended questionnaire. The questionnaire was concerned with the credibility of the evidence and the degree to which practitioners could use the evidence to guide decisions for adopting specific technologies in industrial practice. The evidence was then synthesized considering its strength.

One study (S5) performed a meta-ethnographic study of agile software development. The authors described the evidence from each study according to the themes found in the primary studies. In the discussion section, the authors synthesized the findings according to the research questions and the themes identified in the literature. Meta-ethnographic methods that were used to synthesize the data extracted from the primary studies referenced Noblit and Hare [39].

4.2.3. SR goals and the use of synthesis methods

We classified the goals and research questions of each paper as decision support, knowledge support, or scoping. A scoping goal is one that leads to a mapping of the studies selected, as for example, S46, which stated the following goal: *"...identify and characterize approaches used to evaluate architectural documents..."*, which led to the following research questions: *"How is software architecture or an architectural document evaluated? Which are the approaches and what is the context that they are executed*?" The goals of 23 SRs indicated that they were scoping studies (46.9%). Of these, 16 were ultimately classified as a scoping study, four as thematic analysis, and three as narrative synthesis.

About 10% of the studies had research questions and aims for a synthesis that would produce a SR for decision support. One example is S21, which had the following goal: "*The main aim of our systematic review is to assist software companies with small data sets in deciding whether or not to use an estimation model obtained from a benchmarking data set. The secondary aim is to provide advice to researchers intending to investigate the potential value of cross-company models.*" Of the SRs designed for decision support, only one was defined as a scoping study in our classification because it did not fulfill the goals of the study.

As shown in Table 11, knowledge support was the aim for 21 SRs (42.9%). However, seven of these studies were defined as scoping studies in our classification. The remaining fourteen studies used various methods of synthesis in the review process. One example of a study with a knowledge support goal is S35, which did a comparative analysis of estimation models to arrive at conclusions in the review. The goal of this review was to perform *"a review of the effectiveness of within and between company software effort estimation models"*, and the research question for this review was: "*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?*"

4.3. How were the syntheses presented?

At the center of the findings sections of the SRs, there were always a narrative about the discoveries that were made. Sometimes, these sections included a compelling narrative of the topic under investigation, and other times there was just a brief description of tables. In some cases, we could recognize some logical structure in the text, such as a narrative in a chronological order of the evidence (S1), while in other cases, we could not recognize a logical structure.

Table 11
Goals vs. synthesis methods.

	Scoping study	Thematic analysis	Narrative synthesis	Comparative analysis	Meta-analysis	Case survey	Meta-ethnography
Scoping	S11, S13, S15, S16, S17, S22, S24, S25, S32, S33, S37, S41, S42, S44 S47, S49	S12, S20, S29, S30	S26, S27, S39				
Decision support	S9		S7	S4, S21			S5
Knowledge support	S8, S10, S23, S38, S40, S43, S48	S2, S3, S19, S28	S1, S31, S34, S36, S45	S35, S46	S6, S18	S14	

Tables provide the simplest type of graphic presentation. This form of data representation was found in almost all the SRs that performed a synthesis. Tables provided important structure and sequencing that made logical trends easier for readers to follow. But comparisons of the findings and results were not very common in the studies we examined; only 25% of the studies had a table comparing the findings of the primary studies (e.g., S18, S19, S21, S28).

When the topic or concept that is being studied, or a topic that emerged during the review, includes a specific process, a flowchart can be a useful data organizing tool. We identified one case in which a flowchart was developed during the SR (S30). This SR was about software changes, and it provided comprehensive insight into the architecture change process as well as a framework for assessing change characteristics and their impact on a system.

Other visual representations can also be effective for presenting findings, particularly complex findings, and for showing relationships between concepts. One example is S7 (Fig. 2), where the authors used graphs to show connections among the primary studies' findings and then used the graphs to drive the synthesis of their findings. Alternative representations, such as timelines (Fig. 3) and illustrations of hierarchies, were also found. These representations were usually useful for obtaining an overview of the studies, and especially important to show relationships between findings or when findings become difficult to view in a table format (e.g., S11). The challenge of these types of charts is that all findings may appear to be of equal weight from a visual perspective. The authors in S11 attempted to counter this challenge by assuring that the studies included in the findings were all of the same type (in this case; experiments).

5. Discussion

This study revealed that there is a growing interest and an increasing number of SRs for a wide range of topics in SE. However, this study has also shown that synthesizing the evidence from a set of studies encompassing many countries and years, and incorporating a wide variety of research methods and theoretical perspectives, is probably the most challenging task of performing a SR. These challenges involve which studies to include, how to synthesize the findings, and how to present the results of the synthesis.

In disciplines where there is likely to be a large amount of literature, it is important to consider the breadth and quality of the evidence. The Cochrane Collaboration has emphasized the inclusion, as much as possible, of all RCTs in a specific area, even unpublished ones, to overcome publication bias. There is clearly a tension between minimizing the potential for publication bias and making reviews manageable. SRs that involve the transformation of raw data, or that include large numbers of primary studies, require greater resources, and where the review question and/or range of evidence is very broad, it may be necessary to sample [44]. In the present study, we observed that this tension is occurring in the SE discipline, and that it affects both the potential and the quality of research synthesis in SE.

The fact that most of the SRs included non-empirical primary studies, including expert opinion and advocacy research, clearly indicates that current SRs in SE lack the necessary basis to synthesize results for knowledge support as well as decision support. It is also challenging in itself to synthesize such diverse study types [17]. Therefore, as a basis for the synthesis, it is also important that future reviewers decide on the types of interventions and studies that their SRs will include.

Although two-thirds of the SRs classified their primary studies by the type of intervention, only one-fifth of the SRs used this classification as a basis for synthesis. This trend indicates that the authors of SRs considered intervention as an important aspect for classifying primary studies. At the same time, the authors did not seem to attribute the same importance to the choice of synthesis methods.

A striking result with respect to the quality of the SRs included in the current review was that two-thirds of them did not satisfy the DARE criteria, even though these criteria have been referred

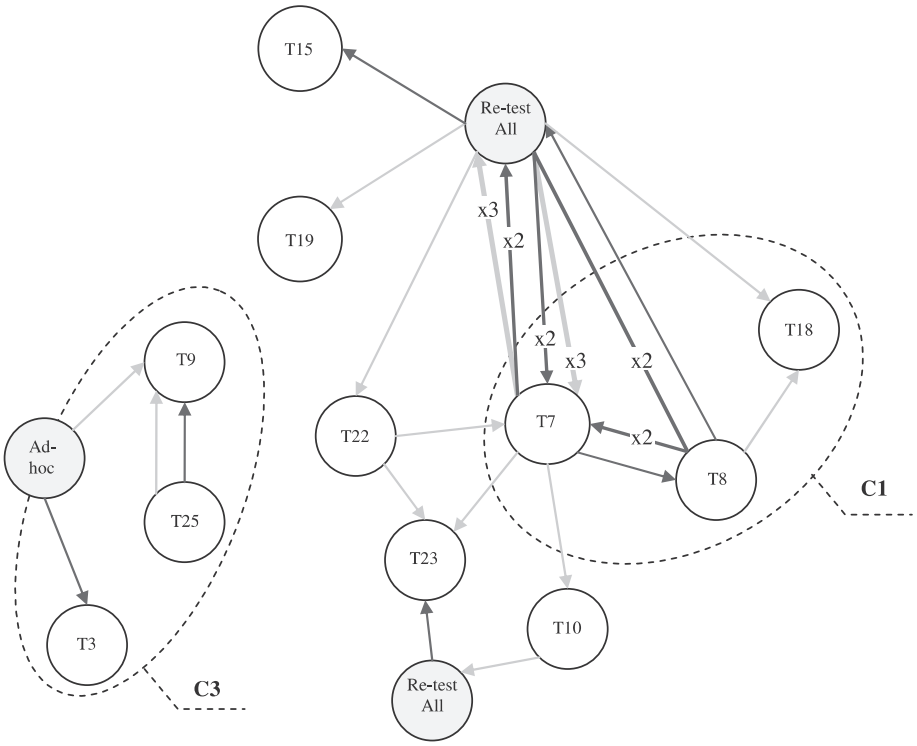

Fig. 2. Empirical results for total time variable in studies included in S7. Grey arrows indicate lightweight empirical result while black arrows indicate medium weight result. A line means that the studies have similar effect; an arrow points to a better technique. Thicker lines represent more studies.

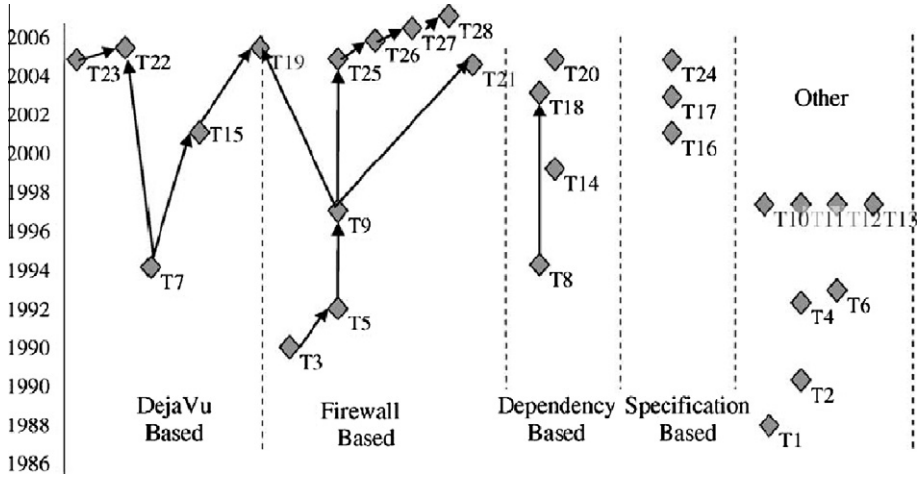

Fig. 3. Evolution of techniques included in the primary studies in S7. There are many variants of techniques, gradually evolved over time, e.g., beginning with T3, T7 and T8.

to and used extensively in influential SE sources on SRs. For example, in two tertiary studies performed by Kitchenham et al. [27,28], each SR was evaluated according to the DARE criteria. However, despite the critical importance of synthesis in SRs, both of these tertiary studies skipped the mandatory DARE criterion on synthesis. The fact that leading researchers within SR methods in SE have not included the mandatory criterion on synthesis in their evaluations of SRs, might explain why so many of the secondary studies in SE that claim to be SRs, are actually scoping studies without synthesis.

Table 12 compares the results of the current study with the results of a tertiary review of health and healthcare studies. Dixon-Woods et al. extracted the topics of the SRs and reported methods for searching, appraisal, and synthesis [14]. The table abridges some of the findings from Dixon-Woods et al.'s study along with the corresponding findings from the current study.

A striking difference between the two disciplines is that the explicitness of searches in SE was much higher than in healthcare and that the median of primary studies included in SE was much higher than the median in healthcare. In healthcare, there is a debate about whether to sample when there is a large body of literature and there are too many sources of evidence to be reviewed feasibly [44]. A similar debate was not raised in the SRs included in this study.

Concerning the quality appraisal of the primary studies, it seems that both disciplines struggle with a lack of consensus for methods and criteria, especially on how to employ appraisals in the synthesis process. In SE only one-third of the SRs appraised their studies, and almost none used the appraisal to exclude low quality primary studies, used the appraisal in the synthesis, or to weigh the evidence. While half of the SRs in healthcare appraised their primary studies, they failed to give an account of whether the judgments of quality were used to exclude papers from the review. They also failed to mention how the outcomes of the appraisals were taken into account in the synthesis.

The disciplines of healthcare and SE seem to be at different stages of development with respect to synthesis and the use of synthesis methods. The differences lead to the assumption that the scope of SRs in SE is broader, but less rigorous with respect to quality and the analysis of primary studies. While all of the healthcare SRs performed a synthesis, as many as half of the claimed SRs in SE did not. Of the SRs in SE that did synthesize their primary studies, two-thirds of them performed narrative synthesis and thematic analysis. In the healthcare discipline, meta-ethnography was the method of synthesis for half of the SRs, while this method was only used by one of the SRs in SE. In addition, some of the SRs in healthcare reported attempts to innovate or adapt methods for synthesis, while no innovation was evident in the SE SRs.

For both disciplines, several SRs lack an explicit description about the methods for searching, appraisal, and synthesis, and there is little evidence of an emerging consensus on many of these issues. Based on the evidence gathered in the current study, we conclude that continued methodological progress and improved reporting are needed.

In the remainder of this section, we discuss the implications of our findings for theory and practice, the limitations of the review, and provide some suggestions for future research.

5.1. Implications for theory and practice

The findings of this review have raised a number of issues that have implications for research and practice. It shows that it is

Table 12
Comparison of tertiary studies in health and healthcare to software engineering.

Criteria	Health and healthcare [14]	Software engineering (current study)
Time span	1994–2004	2005–2010
No. of SRs	42	49
Descriptions of search	The databases that were searched to identify candidate studies for inclusion in reviews were specified by 64.3% (27/42) of the papers	The databases that were searched to identify candidate studies for inclusion in reviews were specified by 95.9% (47/49) of the papers
No of studies synthesized	Ranged from 3 to 292 (median 15)	Ranged from 10 to 304 (median 54)
Appraisal of studies	Precisely, 50% (21/42) of SRs described appraisal of candidate studies in their reviews	Approximately, 35% (17/49) SRs described appraisal of candidate studies in their reviews
Synthesis	All SRs performed some type of synthesis. Almost 50% (19/42) used meta-ethnography as the method of synthesis	Only 50% (24/49) of the SRs performed synthesis of primary studies. Narrative synthesis (nine SRs) and thematic analysis (eight SRs) were the two most common synthesis methods

possible to synthesize primary studies in ways that are somewhere between the extremes of meta-analyses and narrative reviews, and which, we believe, generate new valuable insights for both research and practice. Also, this review shows that most of the claimed SRs in SE are not actually SRs but rather scoping studies of the literature. A key strength of such scoping studies is that they can provide a rigorous and transparent method for mapping current areas of research in terms of the volume, nature, and characteristics of the primary research [2]. Scoping studies make it possible to identify the gaps in the evidence base and to disseminate research findings. However, because the overwhelming majority of SE scoping studies includes non-empirical work, the identification of the actual evidence base underlying these studies is not always straightforward.

In our view, the potential of empirical research will not be realized if individual primary studies are merely listed in the absence of some sort of synthesis. The usefulness of a secondary study will be very limited without such synthesis and without solid empirical studies to base it upon.

Closely related issues, therefore, are the evaluation of the quality of primary studies [53], the decisions of which studies should be included, and how the evidence should be weighted according to quality and suitability for the SR. Although some of the SRs did a quality assessment of their primary studies, these assessments were basically used to characterize the studies, not as a basis for decisions regarding inclusion or exclusion, or to support the synthesis of the evidence. Lessons learned and experience reports or other non-empirical based findings are unlikely to add much value and confidence to the final conclusions of a SR. We would, therefore, encourage SE researchers to be much more restrictive with respect to the included primary studies, or, at a minimum, to factor the low quality studies into the presentation or discussion of the findings.

With respect to study types, two SRs (S6 and S18) included only controlled experiments in their (meta-analytical) synthesis; all of the other SRs included both empirical and non-empirical studies from a variety of perspectives and research methods. Although explicit guidelines are available on how to synthesize quantitative studies in SE (e.g., [37]), there is much less advice on how to synthesize primary studies that incorporate qualitative and mixed-methods approaches. However, as is shown in Section 2, there are several well-known methods for synthesizing evidence from diverse study types; these methods are equally as relevant for SRs in SE as for SRs in any other discipline.

Although the epistemological and ontological foundations of the primary studies might be important for the choice of synthesis method, at present, we would rather call for a more pragmatic approach. In our view, SE researchers would benefit the most from using the SRs' research questions and the primary studies' designs, data collection methods, and methods for data analysis as the drivers for choosing their synthesis methods. However, as this review shows, this issue is currently not being appropriately addressed by the majority of the SRs in SE. Surprisingly, only a small number of SRs describe their methods of synthesis and even fewer cite a recognized method.

Synthesis of findings across diverse study designs is far from simple and is likely to be aimed at identifying recurring themes in the studies or common contextual factors [9,11]. SRs conducted with respect to the determination of why study results differ (as they are likely to do), and the evaluation of the potentially contrasting insights from qualitative and quantitative studies will generally be more helpful in SE than those that focus on identifying average effects. Seemingly unpatterned and disagreeing findings from quantitative studies may have underlying consistency when study design, study settings, developer types, customer and domain characteristics, application details, and the nature of the organizational culture are taken into account. Qualitative data can also be useful in capturing developers' subjective evaluations of organizational- or project-level interventions and outcomes. In addition, qualitative findings can be used to develop theories and to identify relevant variables to be evaluated in future quantitative studies.

The synthesis and presentation of findings are best thought of as parts of the same process, particularly when there are many primary studies included in the review. In this case, it will not be possible to reach a synthesis until an approach for the presentation of the findings has been developed. In addition to a narrative, most SRs in our review provided descriptive tables covering aspects of each study, such as authors, year, and a detailed description of the intervention, theoretical basis, design, quality assessment, outcomes, and main findings. The advantage of such tables is that they make the SR more transparent. However, authors of SR articles must go beyond the presentation of large tables listing large amounts of data from individual studies to create a more useful tabular synthesis that combines the key findings in a more accessible way. Establishing a logical structure of the narrative with supporting recommendations and visual representations will improve the readability of the article and support the decision-making process of practitioners.

5.2. Recommendations and future research

Although our findings may be indicative of the status on the types and methods of synthesis used in systematic reviews in SE, further tertiary studies are needed to evaluate the rigor of such syntheses. Based on the limited attention given to research synthesis that we have identified in current secondary studies, we offer some recommendations and possible future directions for SE researchers.

Given the increased interest in synthesis of research evidence, it is challenging to investigate the ideas inherent to the methods and to explore the facts behind the doubts and warnings that researchers in the field have put forth. SE researchers must incorporate the synthesis methods already defined in other areas such as medicine, nursing, and social science and adapt these to our context in order to improve the synthesis methods used in our discipline. For example, the issue of combining analysis and interpretation of studies with markedly different approaches and intentions presents a particular challenge that may not be surmountable in all cases through the process of synthesis as it has been originally described.

As SRs evolve and continue to gain popularity, awareness is needed to ensure greater transparency and methodological rigor, which will increase the legitimacy of findings and relevance for practice [40]. Despite the fact that standardized quality criteria have yet to be defined, several innovative methods are being developed to address the issue of quality evaluation for qualitative data and synthesis of data from mixed sources of evidence in other areas, such as medicine and nursing. These include approaches that combine syntheses of different pieces of evidence or different types of evidence brought together under a single overarching synthesis [13,41,44,50,51]. These approaches serve to counter much of the criticism of the synthesis of qualitative and mixed methods research in terms of the overall quality, suitability, and the legitimacy of its findings. Such methods would not only provide a reasonable combination of evidence that can be considered trustworthy and relevant, but would also provide a basis for confidence among researchers and practitioners in the use of that evidence.

A particularly relevant method for future research synthesis in SE is Pawson's theory-driven approach of realist synthesis [42]. The core principle of such synthesis is that one should make explicit the underlying assumptions regarding the method in which an intervention is supposed to work and should then gather evidence

in a systematic way to test and refine this theory. Rather than seeking generalizable lessons or universal truths, this approach recognizes and directly addresses the fact that the "same" intervention is impossible to implement in an identical manner and, therefore, never has the same impact because of differences in its context, setting, process, stakeholders, and outcomes. Rather, the aim of realist synthesis is explanatory: "what works for whom, in what circumstances, in what respects, and how?" [43].

Finally, as with other approaches to research and evidence synthesis, a more rigorous approach is required. The researchers in SE must be more consistent when performing SRs. There is a good consistency thus far with respect to the definition of the research questions and search strategies for primary studies. However, SE researchers must be more consistent in the methods with which they select, characterize, analyze, and synthesize the primary studies. We suggest that at all levels of inquiry, a quality SR is one that demonstrates procedural and methodological rigor in all steps. In addition, explicit identification of practical, methodological, and theoretical limitations of the approach undertaken should be described, to ensure that its usefulness and the value of its findings can be appropriately interpreted and used by others.

5.3. Limitations

The main limitations of this review are bias in the selection of publications, inaccuracy in data extraction, and potential author bias. As for the selection of studies, we implemented a simple search for "systematic review" in the title of publication in the ISI Web of Knowledge. We also performed a separate search in the ACM Digital Library for proceedings papers not indexed by ISI. In addition, we examined the articles included in the tertiary reviews by Kitchenham et al. [27,28] for possible inclusion in the current study. Therefore, any studies in publication venues not indexed or included by these sources were not retrieved. Also, because our focus was on systematic reviews and not on meta-analyses, we did not include "meta-analysis" as a search term, and our review would, therefore, not be comprehensive with respect to the total number of secondary studies using meta-analysis as the synthesis method.

Several articles lacked sufficient information regarding the included primary studies and their methods of synthesis to allow us to document them satisfactorily in the extraction form. There is, therefore, a possibility that the extraction process may have resulted in some inaccuracy in the data.

Finally, a potential bias lies in the fact that one of the authors (Dybå) has written papers that were included in the review. In these cases, however, the other author (Cruzes) decided whether or not to include them and judged the extraction, categorization, and analysis of their findings.

6. Conclusion

Our tertiary review of the types and methods of synthesis in systematic reviews shows that there is limited attention paid to research synthesis in SE. We identified few studies that adequately demonstrated a robust academic approach to such synthesis. Half of the studies that referred to themselves as systematic reviews, did not include synthesis, and were, rather, scoping studies that merely mapped out and categorized the primary studies. Furthermore, many of the reviews included primary studies that were either conceptual or that did not base their findings on empirical evidence. In addition, as many as two-thirds of the studies did not use synthesis methods specific for the types of the evidence included in the primary studies.

Synthesis of empirical research is at the heart of systematic reviews, and future attention must be directed toward synthesis methods that increase our ability to find methods in which to compare and combine that which is seemingly incomparable and uncombinable. Such methods will pave the way to increased significance and utility for research and practice of future systematic reviews in SE.

Appendix A. Studies included in the review

[S1] W. Afzal, R. Torkar, R. Feldt, A systematic review of search-based testing for non-functional system properties, Inf. Softw. Technol. 51(6) (2009) 957–976.

[S2] S. Beecham, N. Baddoo, T. Hall, H. Robinson, H. Sharp, Motivation in software engineering: a systematic literature review, Inf. Softw. Technol. 50(9–10) (2008) 860–878.

[S3] F.O. Bjørnson, T. Dingsøyr, Knowledge management in software engineering: a systematic review of studied concepts, findings and research methods used, Inf. Softw. Technol. 50(11) (2008) 1055–1068.

[S4] A. Davis, O. Dieste, A. Hickey, N. Juristo, A.M. Moreno, Effectiveness of requirements elicitation techniques: empirical results derived from a systematic review, Proc. 14th IEEE International Conference on Requirements Engineering, 11–15 September 2006, pp. 179–188.

[S5] T. Dybå, T. Dingsøyr, Empirical studies of agile software development: a systematic review. Inf. Softw. Technol. 50(9–10) (2008) 833–859.

[S6] T. Dybå, V.B. Kampenes, D.I.K. Sjøberg, A systematic review of statistical power in software engineering experiments. Inf. Soft. Technol. 48 (2006) 745–755.

[S7] E. Engström, P. Runeson, M. Skoglund, A systematic review on regression test selection techniques. Inf. Softw. Technol. 52(1) (2010) 14–30.

[S8] E.D. Souza Filho, R. Oliveira Cavalcanti, D.F. Neiva, T.H. Oliveira, L.B. Lisboa, E.S. Almeida, S.R. Lemos Meira, Evaluating domain design approaches using systematic review, Proc. 2nd European Conference on Software Architecture (Paphos, Cyprus, September 29–October 01, 2008) LNCS, Springer-Verlag, Berlin, Heidelberg, 50–65.

[S9] Oswaldo Gómez, Hanna Oktaba, Mario Piattini, Félix García: a systematic review measurement in software engineering: state-of-the-art in measures. 224–231, ICSOFT 2006, First International Conference on Software and Data Technologies, Setúbal, Portugal, September 11–14, 2006.

[S10] T. Hall, N. Baddoo, S. Beecham, H. Robinson, H. Sharp, A systematic review of theory use in studies investigating the motivations of software engineers, ACM Trans. Softw. Eng. Methodol. 18(3) (2009) 1–29.

[S11] J.E. Hannay, D.I.K. Sjøberg, T. Dybå, A systematic review of theory use in software engineering experiments, IEEE Trans. Softw. Eng. 33(2) (2007) 87–107.

[S12] E. Hossain, M.A. Babar, H. Paik, Using scrum in global software development: a systematic literature review. In: Proc. of the 4th International Conference on Global Software Engineering, IEEE Press, Los Alamitos, 2009.

[S13] E. Insfran, A. Fernandez, A systematic review of usability evaluation in web development. In: Proc. International Workshops on Web information Systems Engineering (Auckland, New Zealand, 1–4 September 2008). LNCS. Springer-Verlag, Berlin, Heidelberg, 81–91.

[S14] M. Ivarsson, T. Gorschek, Technology transfer decision support in requirements engineering research: a systematic review of REj, Requir. Eng. 14(3) (2009) 155–175.

[S15] M. Jimenez, M. Piattini, Problems and Solutions in Distributed Software Development: A Systematic Review. Software Engineering Approaches for Offshore and Outsourced Development. SEAFOOD 2008, Zurich, Switzerland, July 2–3, 2008.
[S16] M. Jørgensen, M. Shepperd, A Systematic review of software development cost estimation studies, IEEE Trans. Softw. Eng. 33(1) (2007) 33–53.
[S17] V.B. Kampenes, T. Dybå, J.E. Hannay, D.I.K. Sjøberg, A systematic review of quasi-experiments in software engineering. Inf. Softw. Technol. 51(1) (2009) 71–82.
[S18] V.B. Kampenes, T. Dybå, J.E. Hannay, D.I.K. Sjøberg, A systematic review of effect size in software engineering experiments. Inf. Softw. Technol. 49(11–12) (2007) 1073–1086.
[S19] S.U. Khan, M. Niazi, R. Ahmad, Critical success factors for offshore software development outsourcing vendors: a systematic literature review, Proc. 4th IEEE ICGSE (July 13–16, 2009). IEEE Computer Society, Washington, DC, 207–216.
[S20] M. Khurum, T. Gorschek, A systematic review of domain analysis solutions for product lines, J. Syst. Softw. 82(12) (2009) 1982–2003.
[S21] B.A. Kitchenham, E. Mendes, G.H. Travassos, Cross versus within-company cost estimation studies: a systematic review, IEEE Trans. Softw. Eng. 33(5) (2007) 316–329.
[S22] L.B. Lisboa, V.C. Garcia, D. Lucrédio, E.S. de Almeida, S.R. de Lemos Meira, R.P. de Mattos Fortes, A systematic review of domain analysis tools. Inf. Softw. Technol. 52(1) (2010) 1–13.
[S23] F.J. Lucas, F. Molina, A. Toval, A systematic review of UML model consistency management. Inf. Softw. Technol. 51(12) (2009) 1631–1645.
[S24] E. Mendes, A systematic review of Web engineering research, Proc. International Symposium on Empirical Software Engineering, Noosa Heads, Australia, (17–18 November 2005) pp. 498–507.
[S25] Y. Morais, T. Burity, G. Elias, A systematic review of software product lines applied to mobile middleware, Proc. 6th Int. Conf. on Inf. Technology, New Generations, USA, April, 2009.
[S26] J. Nicolás, A. Toval, On the generation of requirements specifications from software engineering models: a systematic literature review. Inf. Softw. Technol. 51(9) (2009) 1291–1307.
[S27] F.J. Pino, F. García, M. Piattini, Software process improvement in small and medium software enterprises: a systematic review, Soft. Qual. Control 16(2) (2008) 237–261.
[S28] M. Staples, M. Niazi, Systematic review: systematic review of organizational motivations for adopting CMM-based SPI, Inf. Softw. Technol. 50(7–8) (2008) 605–620.
[S29] G.S. Walia, J.C. Carver, A systematic literature review to identify and classify software requirement errors, Inf. Softw. Technol. 51(7) (2009) 1087–1109.
[S30] B.J. Williams, J.C. Carver, Characterizing software architecture changes: a systematic review, Inf. Softw. Technol. 52(1) (2010) 31–51.
[S31] D. Smite, C. Wohlin, T. Gorschek, R. Feldt, Empirical evidence in global software engineering: a systematic review. Empirical Softw. Eng. 15(1) (2010) 91–118.
[S32] M. Svahnberg, T. Gorschek, R. Feldt, R. Torkar, S.B. Saleem, M.U. Shafique, A systematic review on strategic release planning models, Inf. Softw. Technol. 52(3) (2010) 237–248.
[S33] R. Rabiser, P. Grünbacher, D. Dhungana, Requirements for product derivation support: Results from a systematic literature review and an expert survey, Inf. Softw. Technol. 52(3) (2010) 324–346.
[S34] M. Riaz, E. Mendes, E. Tempero, A systematic review of software maintainability prediction and metrics, Proc. 3rd international Symposium on Empirical Software Engineering and Measurement (15–16 October 2009) 367–377.
[S35] S.G. MacDonell, M.J. Shepperd, Comparing local and global software effort estimation models - reflections on a systematic review, Proc. of the First International Symposium on Empirical Software Engineering and Measurement (20–21 September 2007), IEEE Computer Society, Washington, DC, 401–409.
[S36] P. Mohagheghi, R. Conradi, Quality, productivity and economic benefits of software reuse: a review of industrial studies. Empirical Softw. Engg. 12(5) (2007) 471–516.
[S37] A.P. Freire, R. Goularte, R.P. de Mattos Fortes, Techniques for developing more accessible web applications: a survey towards a process classification, Proc. 25th Annual ACM international Conference on Design of Communication (El Paso, Texas, USA, 2–24 October 2007), SIGDOC '07. ACM, New York, NY, 162–169.
[S38] J. Hannay, M. Jørgensen, The role of deliberate artificial design elements in software engineering experiments, IEEE Trans. Softw. Eng. 34(2) (2008) 242–259.
[S39] R.C. de Boer, R. Farenhorst, In search of 'architectural knowledge'. In: Proceedings of the 3rd International Workshop on Sharing and Reusing Architectural Knowledge (Leipzig, Germany, 13 May 2008), SHARK '08. ACM, New York, NY, 71–78.
[S40] C.G. Bellini, R.D.C.D.F. Pereira, J.L. Becker, Measurement in software engineering from the roadmap to the crossroads, Int. J. Softw. Eng. Knowledge 18(1) (2008) 37–64.
[S41] H. Zhang, B. Kitchenham, D. Pfahl, Reflections on 10 years of software process simulation modeling: a systematic review, Proc. International Conference on Software Process, Leipzig, Germany, 10–11 May 2008, 345–356.
[S42] R. Barcelos, G.H. Travassos, Evaluation approaches for software architectural documents: a systematic review, Proc. Workshop Iberoamericano de Ingenieria de Requisitos y Ambientes de Software, La Plata Argentina, vol. 1, 2006, pp. 433–446.
[S43] S. Grimstad, M. Jørgensen, K. Moløkken-Østvold, Software effort estimation terminology: the tower of Babel, Inf. Softw. Technol. 48(4) (2006) 302–310.
[S44] D.I.K. Sjøberg, J.E. Hannay, O. Hansen, V.B. Kampenes, A. Karahasanovic, N. Liborg, A.C. Rekdal, A survey of controlled experiments in software engineering, IEEE Trans. Softw. Eng. 31(9) (2005) 733–753.
[S45] M.S. Ali, M.A. Babar, L. Chen, K. Stol, A systematic review of comparative evidence of aspect-oriented programming, Inf. Softw. Technol. 52(9) (2010) 871–887.
[S46] M. Turner, B. Kitchenham, P. Brereton, S. Charters, D. Budgen, Does the technology acceptance model predict actual use? A systematic literature review, Inf. Softw. Technol. 52(5) (2010) 463–479.
[S47] R. Prikladnicki, J.L. Audy, Process models in the practice of distributed software development: a systematic review of the literature, Inf. Softw. Technol. 52(8) (2010) 779–791.
[S48] V. Alves, N. Niu, C. Alves, G. Valença, Requirements engineering for software product lines: a systematic literature review, Inf. Softw. Technol. 52(8) (2010) 806–820.
[S49] A. Ampatzoglou, I. Stamelos, Software engineering research for computer games: a systematic review, Inf. Softw. Technol. 52(9) (2010) 888–901.

References

[1] Anderson S, Allen P, Peckham S, Goodwin N. (2008) "Asking the right questions: scoping studies in the commissioning of research on the organisation and delivery of health services," *Health Res Policy Syst.* Jul 9; 6:7.
[2] H. Arksey, L. O'Malley, Scoping studies: towards a methodological framework, Int. J. Social Res. Methodol. 8 (1) (2005) 19–32.

[3] V. Braun, V. Clarke, Using thematic analysis in psychology, Qual. Res. Psychol. 3 (2006) 77–101.
[4] CASP, Critical Appraisals Skills Programme, NSH, UK, 2006. <http://www.sph.nhs.uk/what-we-do/public-health-workforce/resources/critical-appraisals-skills-programme>.
[5] B.P. Cohen, Developing Sociological Knowledge: Theory and Method, second Ed., Nelson-Hall, Chicago, 1989.
[6] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, second ed., Laurence Erlbaum, 1988.
[7] H. Cooper, L.V. Hedges, J.C. Valentine (Eds.), The Handbook of Research Synthesis and Meta-Analysis, second ed., Russell Sage Foundation, 2009.
[8] J. Corbin, A. Strauss, Basics of Qualitative Research, third ed., Sage, 2007.
[9] D.S. Cruzes, M.G. Mendonça, V.R. Basili, F. Shull, M. Jino, Extracting Information from Experimental Software Engineering Papers, Proc. SCCC'07, 2007, pp. 105–114.
[10] D.S. Cruzes, T. Dybå, Synthesizing Evidence in Software Engineering Research," Proceedings of the 4th International Symposium on Empirical Software Engineering and Measurement (ESEM 2010), Bolzano-Bozen, Italy, 16–17 September, 2010.
[11] D.S. Cruzes, V.R. Basili, F. Shull, M. Jino, Using Context Distance Measurement to Analyze Results across Studies, Proc. ESEM'07, 2007, pp. 235–244.
[12] K. Davies, N. Drey, D. Gould, What are scoping studies? A review of the nursing literature, Int. J. Nurs. Stud. 46 (2009) 1386–1400.
[13] M. Dixon-Woods, S. Agarwal, D. Jones, B. Young, A. Sutton, Synthesising qualitative and quantitative evidence: a review of possible methods, J. Health Ser. Res. Policy 10 (1) (2005) 45–53.
[14] M. Dixon-Woods, A. Booth, A. Sutton, Synthesizing qualitative research: a review of published reports, Qual. Res. 7 (3) (2007) 375–422.
[15] T. Dybå, Improvisation in small software organizations, IEEE Soft. 17 (5) (2000) 82–87.
[16] T. Dybå, T. Dingsøyr, Strength of Evidence in Systematic Reviews in Software Engineering, Proceedings of the 2nd International Symposium on Empirical Software Engineering and Measurement (ESEM'08), Kaiserslautern, Germany, 9–10 October, ACM Press, 2008, pp. 178–187.
[17] T. Dybå, T. Dingsøyr, G.K. Hanssen, Applying Systematic Reviews to Diverse Study Types: An Experience Report, Proc ESEM'07, 2007, pp. 225–234.
[18] T. Dybå, B.A. Kitchenham, M. Jørgensen, Evidence-based software engineering for practitioners, IEEE Soft. 22 (1) (2005) 58–65.
[19] C.A. Estabrooks, P.A. Field, J.M. Morse, Aggregating qualitative findings: an approach to theory development, Qual. Health Res. 4 (4) (1994) 503–511.
[20] R. Franzosi, Quantitative Narrative Analysis, Sage, 2010.
[21] Glaser, Strauss, The Discovery of Grounded Theory: Strategies for Qualitative Research, Aldine Transaction, 1967.
[22] G.V. Glass, Primary, secondary, and meta-analysis of research, Educ. Res. 5 (10) (1976) 3–8.
[23] M. Hammersley, Systematic or Unsystematic, is that the Question? Some Reflections on the Science, Art and Politics of Reviewing Research Evidence, London: Health Development Agency Public Health Steering Group, 2002.
[24] J.P.T. Higgins, S. Green (Eds.), Cochrane Handbook for Systematic Reviews of Interventions, Version 5.0.2, The Cochrane Collaboration, 2009. <http://www.cochrane-handbook.org> (updated September).
[25] M. Jørgensen, T. Dybå, B. Kitchenham, Teaching Evidence-Based Software Engineering to University Students, Proceedings of the 11th International Software Metrics Symposium (Metrics 2005), Como, Italy, 19–22 September, 2005.
[26] B.A. Kitchenham, S.L. Pfleeger, L.M. Pickard, P.W. Jones, D.C. Hoaglin, K.E. Emam, J. Rosenberg, Preliminary guidelines for empirical research in software engineering, IEEE Trans. Softw. Eng. 28 (8) (2002) 721–734.
[27] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering – a systematic literature review, Inf. Softw. Technol. 51 (1) (2009) 7–15.
[28] B. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, S. Linkman, Systematic literature reviews in software engineering – a tertiary study, Inf. Softw. Technol. 52 (8) (2010) 792–805.
[29] B. Kitchenham, D.I. Sjøberg, O.P. Brereton, D. Budgen, T. Dybå, M. Höst, D. Pfahl, P. Runeson, Can we evaluate the quality of software engineering experiments?. In: Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement (Bolzano-Bozen, Italy, September 16–17, 2010). ESEM '10. ACM, New York, NY, 2010, 1–8.
[30] B.A. Kitchenham, Procedures for Performing Systematic Reviews, Keele University, Technical Report TR/SE-0401 and NICTA Technical Report 0400011T.1, 2004.
[31] B.A. Kitchenham, S. Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering, Version 2.3, Keele University, EBSE Technical Report, EBSE-2007-01, 2007.
[32] B.A. Kitchenham, T. Dybå, M. Jørgensen, Evidence-based Software Engineering, Proc. ICSE'04, Edinburgh, Scotland, 23–28 May, 2004, pp. 273–281.
[33] M.W. Lipsey, D.B. Wilson, Practical Meta – Analysis, Sage, 2001.
[34] A.F. Long, M. Godfrey, An evaluation tool to assess the quality of qualitative research studies, Int. J. Soc. Res. Meth. 7 (2004) 181–196.
[35] N. Mays, C. Pope, J. Popay, Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field, J. Health Serv. Res. Policy 10 (Supplement 1) (2005) 6–20.
[36] M.B. Miles, A.M. Huberman, Qualitative Data Analysis: An Expanded Source Book, Sage, 1994.
[37] J. Miller, Applying meta-analytical procedures to software engineering experiments, J. Syst. Soft. 54 (2000) 29–39.
[38] C. Mulrow, D. Cook (Eds.), Systematic Re-views: Synthesis of Best Evidence for Health Care Decisions, Am. College of Physicians, Philadelphia, 1998.
[39] G.W. Noblit, R.D. Hare, Meta-ethnography: Synthesizing Qualitative Studies, Sage, 1988.
[40] R.T. Ogawa, B. Malen, Towards rigor in reviews of multivocal literatures: applying the exploratory case study method, Rev. Educ. Res. 61 (3) (1991) 265–286.
[41] B.L. Paterson, S.E. Thorne, C. Canam, C. Jillings, Meta-study of Qualitative Health Research: A Practical Guide to Meta-analysis and Meta-synthesis, Sage, 2001.
[42] R. Pawson, Evidence-based Policy: A Realist Perspective, Sage, 2006.
[43] R. Pawson, T. Greenhalgh, G. Harvey, K. Walshe, Realist review – a new method of systematic review designed for complex policy interventions, J. Health Serv. Res. Policy 10 (1) (2005) 21–34.
[44] C. Pope, N. Mays, J. Popay, Synthesizing Qualitative and Quantitative Health Evidence: A Guide to Methods, Open University Press, 2007.
[45] C.C. Ragin, The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies, University of California Press, 1987.
[46] C. Robson, Real World Research, second ed., Blackwell, 2002.
[47] M. Rodgers, A. Sowden, M. Petticrew, L. Arai, H. Roberts, N. Britten, J. Popay, Testing methodological guidance on the conduct of narrative synthesis in systematic reviews, Evaluation 15 (1) (2009) 49–74.
[48] R. Rosenthal, D.M. DiMatteo, Meta-analysis: recent development in quantitative methods for literature reviews, Ann. Rev. Psychol. 52 (2001) 59–82.
[49] P. Runeson, M. Höst, Guidelines for conducting and reporting case study research in software engineering, Empirical Softw. Eng. 14 (2) (2009) 131–164.
[50] M. Sandelowski, J. Barroso, Handbook for Synthesizing Qualitative Research, Springer, 2007.
[51] M. Sandelowski, S. Docherty, C. Emden, Qualitative metasynthesis: issues techniques, Res. Nurs. Health 20 (4) (1997) 365–371.
[52] T.A. Schwandt, Farewell to criteriology, Qual. Inquiry 2 (1) (1996) 58–72.
[53] D.I.K. Sjøberg, T. Dybå, M. Jørgensen, The Future of Empirical Methods in Software Engineering Research, Proc. FOSE'07, 2007, pp. 358–378.
[54] L. Spencer, J. Ritchie, J. Lewis, L. Dillon, Quality in Qualitative Evaluation: A Framework for Assessing Research Evidence, Cabinet Office, London, 2003. monograph online.
[55] D. Walsh, S. Downe, Appraising the quality of qualitative research, Midwifery 22 (2006) 108–119.
[56] R.K. Yin, K.A. Heald, Using the case survey method to analyze policy studies, Admin. Sci. Quart. 20 (1975) 371–381.