Synthesizing Evidence in Software Engineering Research

Daniela S. Cruzes
NTNU
NO-7491 Trondheim, Norway
Tel. +47 73 59 34 40
dcruzes@idi.ntnu.no

Tore Dybå
SINTEF
NO-7465 Trondheim, Norway
Tel. +47 73 59 29 47
tore.dyba@sintef.no

ABSTRACT
Synthesizing the evidence from a set of studies that spans many countries and years, and that incorporates a wide variety of research methods and theoretical perspectives, is probably the single most challenging task of performing a systematic review. In this paper, we perform a tertiary review to assess the types and methods of research synthesis in systematic reviews in software engineering. Almost half of the 31 studies included in our review did not contain any synthesis; of the ones that did, two thirds performed a narrative or a thematic synthesis. The results show that, despite the focus on systematic reviews, there is, currently, limited attention to research synthesis in software engineering. This 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.

Categories and Subject Descriptors
D.2.0 [**Software Engineering**]:General

General Terms
Management, Measurement, Experimentation, Theory

Keywords
Evidence-Based Software Engineering, Systematic Reviews, Research Synthesis, Qualitative, Quantitative, Mixed Methods

1. INTRODUCTION
The progress of Software Engineering (SE) as a discipline hinges on our abilities to cumulate empirical evidence about phenomena in an orderly and accurate fashion. Comparing and contrasting such evidence is necessary to reach conclusions about the empirical support for a phenomenon. Accurate combination of study outcomes in terms of research synthesis is, therefore, at the center of the scientific enterprise of the SE discipline.

Research synthesis is a collective term for a family of methods for summarizing, integrating, combining, and comparing the findings of different studies on a topic or research question [5], [9], [19]. Such synthesis can also identify crucial areas and questions that have not been addressed adequately with past empirical research. It 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 [3]. Research synthesis is, thus, a way of making sense of what a *collection* of studies is saying [19].

The key objective of research synthesis is to evaluate the included studies for heterogeneity and select appropriate methods for integrating [5] or providing interpretive explanations about them [19]. If the primary studies are similar enough with respect to interventions and quantitative outcome variables, it may be possible to synthesize them by 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 required [10].

Although research is underway in other disciplines (e.g., [9], [21], [27]), there remain 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 and from intra-study type syntheses to produce a cross-study type synthesis. There are also challenges related to the methods of qualitative synthesis, as well as to ways of integrating qualitative synthesis with meta-analysis.

The aim of this paper is to contribute to a better understanding of these challenges and their implications for the progress of empirical SE research by examining the types and methods of research synthesis employed in systematic reviews (SR) 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 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 paper is organized as follows: Section 2 describes the theoretical background and examines the concept of research synthesis, Section 3 provides an overview of the research method used in our study, Section 4 presents the findings of our study. Finally, Section 5 provides a discussion of the findings and their implications for research and practice, as well as for future research, while Section 6 provides a conclusion of the study.

ESEM'10, September 16-17, 2010, Bolzano-Bozen, Italy.

2. THEORETICAL BACKGROUND

According to Merriam-Webster's dictionary, there are three definitions of synthesis that are applicable to our purpose. First, synthesis is the combination of parts or elements so as to form a whole. Second, synthesis is the dialectic combination of thesis and antithesis into a higher stage of truth. Third, synthesis is the combining of often diverse conceptions into a coherent whole.

Noblit and Hare made these same three distinctions 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 as 'refutational translations' (or forms of resolution), while the third explains synthesis of studies that taken together may represent a line of argument (or forms of reconceptualization) [19].

Noblit and Hare further distinguished between integrative and interpretive syntheses [19]. Integrative syntheses are concerned with combining or summarizing data for the purpose of creating generalizations [5]. It involves the quantification and systematic integration of data through techniques, such as meta-analysis (that are concerned with assembling and pooling well specified data) or less formal techniques (such as providing a descriptive account of the data). Interpretive syntheses, on the other hand, achieve synthesis through subsuming the concepts identified in the primary studies into a higher-order theoretical structure. The primary concern is with the development of concepts and theories that integrate those concepts. An interpretive synthesis will therefore avoid specifying concepts in advance of the synthesis, but rather ground concepts in the data reported in the primary studies [9]. While most forms of synthesis can 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.

In the remainder of this section, we present, and briefly comment upon, a selection of synthesis methods (Table 1). Based on the methods' epistemological and ontological foundations, synthesis methods can be characterized as primarily integrative, primarily interpretive or they can combine elements of both integration and interpretation.

Table 1. Selected methods for synthesis of qualitative and quantitative evidence (adapted from Dixon-Woods et al., 2005)

Method	Outline	Strengths	Challenges
Narrative synthesis (Rodgers et al. [25])	Selected narrative description and ordering of primary evidence with commentary and interpretation.	• Can cope with large evidence base, comprising diverse evidence types. • Flexibility. • Can be used for theory-building.	• Lack of transparency. • Many variants and lack of procedures/standards. • May be dependent on prejudices of reviewer.
Meta-ethnography (Noblit & Hare [19])	Induction, interpretation and translational analysis of primary studies to understand and transfer ideas, concepts and metaphors across different studies.	• Seeks higher order (generalized) theories. • Can potentially deal with diverse evidence types.	• Lack of transparency. • May be order-of-synthesis effects.
Cross-case analysis (Miles, Huberman [18])	Evidence from each primary study is summarized and coded under broad thematic headings. Evidence is then summarized within themes across studies, with brief citation of primary evidence. Commonalities and differences between studies are noted.	• Highly systematic method. • Potentially allows inclusion of diverse evidence types. • Could be used for theory-building.	• Can be seen as unnecessarily and inappropriately stifling interpretive processes.
Thematic Analysis (Braun, Clarke [2])	Identification of major/recurrent themes in literature; summary of findings of primary studies under these thematic headings.	• Flexible procedures for reviewers. • Copes well with diverse evidence types. • Could be used for theory-building.	• Lack of transparency. • Largely descriptive/data-driven basis to groupings.
Content Analysis (Franzosi [12])	Evidence from primary studies coded under broad thematic headings, using extraction tools designed to aid reproducibility. Occurrences of each theme counted and tabulated.	• Software available. • Transparent and auditable. • Could be used for theory-building.	• Emphasis on frequency compared with importance and interpretation of evidence. • Tends to diminish complexity and context.
Case Survey (Yin, Heald [31])	Each primary study treated as a 'case'. Study findings and attributes extracted using closed-form questions, for reliability. Survey analysis methods used on extracted data.	• Can incorporate diverse evidence types. • Can cope with large numbers of primary studies. • Could be used for theory-building.	• Applicable to outcomes, but less adequate for process. • Lacks sensitivity to interpretive aspects of evidence.
Qualitative Comparative Analysis (Ragin [23])	Boolean analysis of necessary and sufficient conditions for particular outcomes to be observed, based on presence/absence of independent variables and outcomes in each primary study.	• Transparent. • Can incorporate diverse forms of evidence. • Allows competing explanations to be explored and retained and permits theories about causality. • Does not require as many cases as the case survey method.	• Focused on causality determination, not interpretive aspects of qualitative data.
Meta-analysis (Glass [13])	Quantitative synthesis of data, in which evidence is pooled using statistical techniques.	• Transparent and auditable. • Can cope with large numbers of primary studies. Could be used for theory-building.	• May be difficult to obtain an effect size for a given study. • Heterogeneous studies.

The traditional view of research synthesis is the integrative, quantitative approach with its emphasis on accumulation of data, and analysis in terms of a meta-analysis. Meta-analysis is a form of additive synthesis that combines the numerical results of primary studies, estimates the descriptive statistics and explains the inconsistencies of effects as well as the discovery of moderators and mediators in bodies of research findings [13],[17],[26]. The purpose of meta-analysis is to aggregate the results of studies to predict future outcomes for situations with analogous conditions.

Contrary to the purely integrative, quantitative method, we find several methods for conducting interpretive syntheses of qualitative and heterogeneous research. Traditionally, the interpretive synthesis has been presented in the form of a narrative, as opposed to statistical, summary of the findings of studies [25]. Its form may vary from the simple recounting and description of findings through to more interpretive and explicitly reflexive accounts that include commentary and higher levels of abstraction [9].

Meta-ethnography is a more formal method that involves induction and interpretation and, thus, resembles the qualitative methods of the studies it aims to synthesize. The product of a meta-ethnographic synthesis is the translation of studies into one another. It is this aspect of mutual translation that distinguishes meta-ethnography from more traditional methods. Interpretations and explanations in the primary studies are treated as data, and are translated across several studies to produce a synthesis [19].

Miles and Huberman identified some possible methods within an overall approach to qualitative data analysis that they classified as cross-case analysis [18]. Their approach includes a variety of devices such as tabular displays and graphs to manage and present qualitative data without destroying the meaning it through intensive coding. These methods include meta-matrices for partitioning and clustering data in various ways. However, Miles and Huberman's emphasis on highly disciplined procedures has been criticized as an unnecessarily and inappropriately overwhelming interpretive process [9].

A recent evolution within the synthesis literature involves the application of synthesis methods to combine evidence of both qualitative and quantitative methods to obtain a more complete picture of a phenomenon. Thematic analysis is one such method for identifying, analyzing and reporting patterns (themes) within data [2]. However, an important problem with respect to research synthesis is that a thematic analysis has limited interpretative power beyond mere description if it is not used within an existing theoretical framework.

Content analysis is a more systematic way of categorizing and determining the frequencies of data that relies on precise specifications of categories and systematic application of rules [12]. It is a well-developed, fairly transparent and widely used method in the social sciences. However, the frequency-counting techniques of content analysis may fail to reflect the structure or importance of the underlying phenomenon, and results may be over-simplified and count what is easy to classify and count, rather than what is truly important [9].

The case survey method is an example of 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 [31]. It converts qualitative evidence into a quantitative form, thereby synthesizing both qualitative and quantitative evidence. An important limitation to the case survey method, however, is the difficulty in coping with the interpretive properties of qualitative data; and contextual factors that might be important in explaining the features of particular cases may be 'stripped out' as the data are reduced to quantitative form [9].

Finally, the qualitative comparative analysis (QCA) method is an example of a mixed synthesis method that analyzes complex causal connections using Boolean logic to explain pathways to a particular outcome based on a 'truth table' [23]. However, as with any approach based on converting qualitative evidence into quantitative form, qualitative researchers are likely to argue about ontological and epistemological assumptions [9].

The synthesis methods discussed above is by no means meant to be an exhaustive list; there exists numerous other methods with slightly different terminology as well epistemological and ontological foundations. Besides, although combined approaches are methodologically challenging, in that they combine both integrative and interpretive synthesis, they may help move SE beyond the currently competing polarities of quantitative and qualitative empirical research.

3. RESEARCH METHODS

This study is a tertiary review [15] 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 & Humanities Citation Index (A&HCI)
- Conference Proceedings Citation Index- Science (CPCI-S)
- Conference Proceedings Citation Index-Social Science & 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 January 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

We restricted the start of our search to the beginning of 2005 since we would not expect papers published earlier to have been influenced by the seminal papers on evidence-based SE [11], [16] or the procedures for undertaking systematic reviews [14].

This search procedure retrieved 75 articles of which two were duplicates and two were conference paper versions of later journal articles, thus, resulting in 71 unique studies. Of these, we excluded two short papers and 37 studies that were clearly outside the subject area of SE (e.g., studies within medical informatics) or that were lessons learned reports on conducting systematic reviews. In addition, we were not able to retrieve one of the

papers, thus leaving 31 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 perceive synthesis within the context of a systematic review.
- Databases used to search for primary studies.
- The number, and time-span, of 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 and a corresponding reference.
- Whether the authors synthesized findings according to the types of primary studies included.
- Which types and methods of synthesis were 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 agreement. To answer our research questions, we analyzed the extracted data both qualitatively and quantitatively. Although we include a short narrative description of the results, the majority of the results are tabulated to present an overview of the findings and basic information about each study. As such, our study is a scoping study [1] that 'maps' out the systematic review literature in SE 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 for a total of 31 studies until the beginning of 2010 (Table 2). The journal *Information and Software Technology* (IST) was the first to introduce systematic reviews to its readers and to publish them, and it has taken the lead in the number of systematic reviews (15/31 SRs) (Table 3). The second place is taken by Conferences in general (9/31), and the third place by TSE with 3 SRs. The remaining three are published in various SE and SE related journals and proceedings.

There is a diversity of topics addressed in the SRs. As shown in Table 4, systematic reviews in SE can be classified into 15 broad research areas, reflecting the topics in which empirical research in SE has increased during the last years thus making systematic reviews possible. Experimental Methods in SE (4/31), Requirements Engineering (4/31), and Software Design (3/31) are the three areas with most systematic reviews. Of the 15 topics, four were addressed by only one systematic review.

Most authors claim that the rational behind their SR is that it is a formalized, repeatable process to systematically search a body of literature to document the state of knowledge on a particular subject, and that the benefit of performing a SR is that it provides the researchers with more confidence that they have uncovered as much relevant information as possible. Although search strategies and data extraction methods are typically described in detail, few mention synthesis methods. Some doesn't even mention the synthesis part at all, but basically only describe the selection and extraction. Others use 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 review according to the three research questions in Section 1.

Table 2. Publication year

Year	#	Percent
2010	3	9.7%
2009	12	38.7%
2008	8	25.8%
2007	4	12.9%
2006	3	9.7%
2005	1	3.2%

Table 3. Publication venues

Publ.	#	Percent
IST	15	48.4%
Conference	9	29.0%
TSE	3	9.7%
TOSEM	1	3.2%
SQJ	1	3.2%
Workshop	1	3.2%

Table 4. Main topic areas in SE systematic reviews

Main Topic Area	Studies
Agile Software Development	S5, S12
Cost Estimation	S16, S22
Distributed Software Development	S15, S19
Domain Analysis	S23
Experimental Methods in SE	S17, S18, S21, S6
Knowledge Management in SE	S3
Motivation in SE	S2, S29
Product Lines Software Development	S20, S26
Requirements Engineering	S4, S14, S27, S30
Software Design	S8, S24, S31
Software Measurement	S9
Software Process	S28
Testing	S1, S7
Theory Use in SE	S10, S11
Web Development	S13, S25

4.1 What Is the Basis for the Reviews?

Sample size, publication years, and types of primary studies included, varied among the studies. The identification process of appropriate publications was almost always described (Table 5). 64.5% of the searches were based on retrievals from IEEE eXplore and the ACM Digital Library. Manual searches were included in 51.6% studies. Other databases appearing in around 20 to 35 percent of the papers were ISI, Google Scholar, Inspec, Ei Compendex and Science Direct.

Five studies did not describe their search procedures or searches in detail. One study (S8) included books, chapters, dissertations, and theses. In addition to empirical research studies, and contrary to the aims of systematic review research, most of the SRs also included non-empirical primary studies.

All studies had information on the sample size, which ranged from 10 to 304 primary studies. The average is 74 papers. Four articles were not clear about the amount of primary studies that were scanned through when searching for studies.

The majority of the SRs were clear about what type of primary studies were included. But, ten (32.3%) of the SRs didn't classify their primary studies, meaning that the synthesis didn't consider the types of studies included. Of the 17 (54.8%) SRs that synthesized findings only eight (25.8%) did some analysis considering the type of primary study. Six (19.4%) SRs explicitly included articles without empirical validation, while seven (22.6%) explicitly included experience reports in their reviews. 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.

Table 5. Overview of systematic reviews in SE, January 2005 to February 2010

ID	Main Topic Area	#Primary Studies	Classified the Types of Studies (Ex, case studies, Experiments, Surveys?)	Types of Studies Included	Synthesis according to study types?
Thematic Analysis					
S3	Knowledge Management in SE	64 out of 2102	Yes	Diverse	No
S12	Agile Software Development	20 out of 366	Yes	Diverse including Experience Reports	No
S19	Distributed Software Development	122 out of 1300	No	Diverse including non-empirical	No
S20	Product Lines Software Development	62 out of 843	Yes	Diverse including non-empirical	Yes
S29	Motivation in SE	43 out of 634	Yes	Diverse including Experience Reports	Yes
S30	Requirements Engineering	149 out of 7838	Yes	Diverse	No
S31	Software Design	130 out of 2752	No	Diverse	No
Narrative Synthesis					
S1	Testing	35 out of 415	Yes	Diverse	Yes
S2	Motivation in SE	92 out of > 2000	Yes	Diverse including non-empirical	No
S7	Testing	27 out of 2923	Yes	Diverse	Yes
S27	Requirements Engineering	23 out of 795	Yes	Diverse including non-empirical	No
S28	Software Process	45 out of 743	Yes	Diverse	No
Comparative Analysis					
S4	Requirements Engineering	26 out of 564	No	Diverse	No
S22	Cost Estimation	10 out of 1344	No	Diverse	No
Meta-Analysis					
S6	Experimental Methods in SE	103 out of 5423	Yes	Experiments Only	Yes
S18	Experimental Methods in SE	103 out of 5423	Yes	Experiments Only	Yes
Case Survey					
S14	Requirements Engineering	97 out of 181	Yes	Diverse including non-empirical	Yes
Meta-Ethnography					
S5	Agile Software Development	36 out of 1996	Yes	Diverse	Yes
Scoping Study					
S8	Software Design	17 out of ?	No	Diverse including non-empirical	N/A
S9	Software Measurement	78 out of 8912	Yes	Diverse including non-empirical	N/A
S10	Theory Use in SE	92 out of > 2000	No	Diverse including non-empirical	N/A
S11	Theory Use in SE	103 out of 5423	Yes	Experiments Only	N/A
S13	Web Development	51 out of 410	Yes	Diverse including non-empirical	N/A
S15	Distributed Software Development	69 out of 2224	Yes	Diverse including non-empirical	N/A
S16	Cost Estimation	304 out of ?	Yes	Diverse including non-empirical	N/A
S17	Experimental Methods in SE	103 out of 5423	Yes	Experiments Only	N/A
S21	Experimental Methods in SE	20 out of ?	No	Systematic Literature Reviews	N/A
S23	Domain Analysis	19 out of ?	No	Diverse including non-empirical	N/A
S24	Software Design	54 out of 907	No	Diverse including non-empirical	N/A
S25	Web Development	173 out of 343	Yes	Diverse including non-empirical	N/A
S26	Product Lines Software Development	18 out of 3003	No	Diverse including non-empirical	N/A

Furthermore 117 out of the 173 primary studies included in S25 were advocacy research, proof of concept or experience report, whereas half of the studies included in S13 did not conduct any type of empirical evaluation. These findings clearly indicate that the synthesis of primary studies in current SRs may lack rigor on the procedures to synthesize results. In addition, with such diversity in the primary studies included in current SRs, it will be very challenging to perform tertiary studies since the basis for synthesis is very heterogeneous.

4.2 How Are the Findings Synthesized?

As shown in Table 5 and Table 6, a large proportion of the studies (13 SRs; 41.9%) did not perform any synthesis and should therefore be termed scoping (or mapping) studies rather than SRs. Thematic analysis (7 SRs; 22.6%) and narrative synthesis (5 SRs; 16.1%) were the two most common synthesis methods, followed by comparative analysis and meta-analysis with two SRs (6.5%) each. Finally, there were one case survey and one meta-ethnography in our sample.

Table 6. Synthesis methods in SE systematic reviews

Synthesis Method	Studies	#	Percent
None (Scoping)	S8, S9, S10, S11, S13, S15, S16, S17, S21, S23, S24, S25, S26	13	41.9%
Thematic Analysis	S3, S12, S19, s20, S29, S30, S31	7	22.6%
Narrative Synthesis	S1, S2, S7, S27, S28	5	16.1%
Comparative Analysis	S4, S22	2	6.5%
Meta-analysis	S6, S18	2	6.5%
Case Survey	S14	1	3.2%
Meta-ethnography	S5	1	3.2%

In some SRs, the research methodology was explained in detail, sometimes 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 papers, instead of explaining the synthesis procedures, the authors explained how the extraction of the data was done. In addition, although half of the papers contained some synthesis section, only five (16.1%) of the papers included a reference to a method of synthesis, even though not always following the reference in the synthesis. Examples of these methods references included Noblit and Hare's meta-ethnography [19], Ragin's qualitative comparative method [23], Miles and Huberman's [18], and Strauss and Corbin's constant comparison method [30], and Cohen's post hoc power calculations [4].

Fourteen studies were categorized as scoping studies (Table 6). 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. Scoping studies are commonly intended to guide more focused lines of research and development [1], [8]. The main reason for a study to be in this category is that the study does not synthesize evidence of the area in focus, but rather provides an overview of the subject area. One example of a scoping study is S16, in which the authors classified the estimation studies according to pre-specified properties and categories such as estimation topics, estimation approach, research approach, study context and data set.

Another example is S8 in which ten domain design approaches were selected from the literature with a brief chronological description of the selected approaches. The authors do a mapping of the completeness of domain design approaches and evaluate the key points and drawbacks of the approaches reviewed, but they do not synthesize any findings from the primary studies considering these approaches. S10 is also a scoping study in which they reference Robson [24] for the method of synthesis as they affirm that the study is a "grounded approach", although we didn't find the steps described in the book or in the literature about grounded theory (e.g., [30]). The authors analyzed each article in terms of classic theory use and also identified other theories of human behavior such as Taylor's scientific management. They identified the number of articles using a theory in specific ways and then applied a second categorization process to analyze how each classic theory was used. However, they did not synthesize any of their findings.

Seven SRs did a thematic analysis of the primary studies (Table 6), but none of them reference any methodological paper for supporting the method used for the analysis. One example of thematic analysis 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 present frequencies of the number of times each theme is identified in different studies. The respective frequencies reflect the number of times a particular challenge has been mentioned in different papers. Exploring potential strategies to deal with those challenging factors were also another research focus. The authors presented the findings in two stages: first, initial quantitative data about the number of published papers in each year starting from 2003, the types of studies reported in the reviewed papers and the contextual factors of the projects were reported. In the second stage, 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 analysis, the authors of S30 show a SR with the aim to identify and classify types of requirement errors into a taxonomy to support the prevention and detection of errors. The authors first describe 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 in the following way: the errors identified from the SE and psychology fields were collected, analyzed for similarities, and grouped into the taxonomy. The thematic analysis was based on errors that had similar characteristics (symptoms) and were grouped into an error class. The authors then synthesized and described each error class along with the specific errors that make up that class as well as the references that backed up the findings.

Five SRs were categorized as narrative synthesis (S1, S2, S7, S27, S28). In S1, the authors described the evidence from each study in a chronological way and then discussed some differences and possible explanations for them in the results. The authors do the narrative synthesis according to the categories found for non-functional search-based software testing. In S7, the authors analyzed the empirically evaluated relations between the regression test selection techniques by showing the results of the studies in graphs combined with narrative synthesis.

Two SRs were categorized as comparative analysis. One of them references to Ragin [23], while the other one doesn't mention the approach explicitly. However, neither of them applied fully the method as described by Ragin, since they do not use the concepts of a 'truth table' and Boolean algebra. In study S4 the authors provide a table in which they compare the studies providing evidence for and against a certain result, as well as any relevant aggregation-related issues. In study S22 the authors provide tables in which they identify a variety of options for performing a comparative study of cross-company and within-company estimation models. They consider the pros and cons of each option and identify which primary studies (if any) used that option. Based on the results of this study and in their own experience they provide a comparison table with a summary of advice on factors to consider when considering a cross-company model describing the evidences in favor and against each item.

Two SRs were categorized as meta-analysis. The two reviews are from the same research group, using the same set of primary studies (103 experimental papers). One of the reviews (S6) analyzed the statistical power and the other (S18) the effect size in SE experiment. 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 [4]. Study S18 cited various meta-analytic references (e.g., [26]) and performed

meta-analysis based on Hedges' *g* as the standardized effect-size measure.

The systematic review, S14 is the only example of case survey. The goal of the SR was to provide an objective view of what technologies are present in requirements engineering research and to what extent papers describing these technologies provide 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 concern with the credibility of the evidence and the degree to which practitioners can use the evidence to guide decisions for adopting the technologies in industrial practice. This is done by evaluating the type of research performed in terms of research methodology, who performs the evaluations, where the evaluations are performed, and to what extent the design, context, and validity are described. The evidence for the support for technology transfer is synthesized considering the strength of the evidence gathered.

One study (S5) performed a meta-ethnographic study in agile software development. The authors of the SR described the evidence from each study according to the themes found in the primary studies. In the discussion section the authors synthesize the findings according to the research questions and the themes identified in the literature. Meta-ethnographic methods were used to synthesize the data extracted from the primary studies referencing Noblit and Hare [19]. The first stage of the synthesis was to identify the main concepts from each primary study, using the original author's terms. The key concepts were then organized in tabular form to enable comparison across studies and the reciprocal translation of findings into higher-order interpretations. When the authors identified differences in findings, they investigated whether these could be explained by the differences in methods or characteristics of the study settings. This process is analogous to the method of constant comparison used in qualitative data analysis and they referenced Miles and Huberman [18] and Strauss and Corbin [30] for the method.

4.3 How Are the Syntheses Presented?

The purpose of the findings section in a SR is to present what has been discovered through the process of synthesis. At the heart of the findings section, there is always a narrative of the discoveries made (See Table 7). Sometimes these are a compelling narrative of the topic under investigation; other times 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 such logic.

Table 7. Presentation of SE systematic reviews

Presentation of the Synthesis	#	Percent
Narrative	31	100.0%
Characterization of primary studies	28	90.3%
Frequency tables or graphics	21	67.7%
Findings organized according to primary studies	17	54.8%
Comparison of findings or results	11	35.5%
Illustrations of hierarchies, timelines, etc	6	19.4%
Graphs	2	6.5%

In addition to the narrative, tables and charts represent the simplest types of graphic presentations. Tables allow for condensing and organizing data about many themes into rows or columns. They provide important structure and sequencing that makes the logic trail easier for readers to follow. As already noticed, several of the SRs had incomplete analysis and most of them actually lacked a synthesis altogether. So, about 50% of the SRs ended the discussion when they showed the frequency tables about the primary studies. Only 54.8% of the SRs had tables or figures showing the findings and references to the primary studies in which the findings were supported. The number of studies that had comparison of the findings or results is even lower (35.5%), while the number of studies that containing other types of visual representations are just 19.4%.

Other visual representations can also be effective in presenting findings, particularly complex findings, and in showing connections between concepts. Several different graphic organizers can help make narrative information visually appealing and easily understood. One example is S7, where the authors used graphs to show connections among the primary studies' findings and then using the graphs to drive the synthesis of their findings. Alternative representations, such as timelines figures, illustrations of hierarchies were also found. These representations were usually useful to get an overview of the studies and especially important to show relationships between findings or when findings become unwieldy in a table format (e.g., S11). The challenge of these types of chart is that visually, all findings may appear to be of equal weight. 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 has revealed that there is a growing interest and an increasing number of SRs within a wide range of topics in SE. However, the study has also shown that synthesizing the evidence from a set of studies spanning 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. In the remainder of this section, we discuss the implications of our findings, the limitations of our review, and 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 possible to synthesize papers in ways that are somewhere between the extreme points of meta-analysis and narrative reviews, and which, we believe, produces new valuable insights for both research and practice. Also it shows that most of the SRs in SE are not really systematic reviews 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 [1]. Scoping studies makes it possible to identify the gaps in the evidence base and disseminating research findings. However, since the overwhelming majority of SE scoping studies includes non-empirical work, it is not always straightforward to identify the actual evidence base underlying these studies.

In our view, the potential of empirical research will not be realized if individual primary studies are merely listed and some kind of synthesis is not carried out. It is this synthesis of empirical evidence that is at the heart of a SR. Without such synthesis, and without solid empirical studies to base it upon; the usefulness of SRs will be very limited.

A closely related issue, therefore, is the evaluation of the quality of primary studies [29] and the decisions on which studies to include and how to weight the evidence 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 and not as a basis for decisions of inclusion or exclusion, or to support the synthesis of the evidence. Lessons learned and experience report or other non-empirical based finding are unlikely to add much value and confidence to the final conclusions of a systematic review We would, therefore, encourage SE researchers to be much more restrictive with respect to which primary studies they include, but as a minimum factoring low quality studies into the presentation or discussion of the findings.

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

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 driver for choosing their synthesis methods. However, as this review shows, this is currently not being properly addressed by most of the SRs in SE. It is a surprising small number of SRs that describe their methods of synthesis and even fewer that cites 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 [6], [7]. SRs of why study results differ (as they are likely to do), and of the potentially contrasting insights from qualitative and quantitative studies, will generally be more helpful in SE than identifying average effects. Seemingly unpatterned findings in different directions from quantitative studies may be shown to have underlying consistency when study design, study settings, developer types, customer and domain characteristics, application details, as well as 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 as well as identifying 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, especially when there are many studies in the review. In this case, it will not be possible to reach a synthesis until an approach to presenting 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, detail of the intervention, theoretical basis, design, quality assessment, outcomes, and main findings. The advantage of such tables is that the SR gets more transparent. However, authors of SRs articles have to go beyond the presentation of large tables listing a lot of data from individual studies to produce a more useful tabular synthesis, combining the key findings in a more accessible way. Establishing a logical structure of the narrative with supporting recommendations and visual representations for example, will aid the readability of the paper and support the decision-making process of practitioners.

5.2 Recommendations and Future Research

Although our findings may be indicative for 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 to research synthesis that we have identified in current systematic reviews, 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 in the methods and explore the facts behind the doubts and warnings that eminent researchers in the field have put forth. SE researchers need to incorporate the synthesis methods already defined in other areas like medicine, nursing and social science and adapt to our context improving the use the synthesis methods in our area. For example, the issue of combining analysis and interpretation from 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 that will increase the legitimacy of findings and relevance for practice [20]. However, standardized quality criteria have yet to be defined although several innovative methods are being developed to address the issue of quality of qualitative data and synthesis of data from mixed sources of evidence in other areas such as medicine and nursing. These include approaches that combine different evidence syntheses or different types of evidence brought together under a single overarching synthesis [9], [21], [22], [27], [28]. These approaches serve to counter much of the criticism leveled at the synthesis of qualitative and mixed method research in terms of the overall quality, appropriateness, and legitimacy of findings. Such methods would not only provide a reasonable combination of evidence considered trustworthy and relevant, but also provide a basis for confidence among researchers and practitioners in the utility of that evidence.

Finally, as with other approaches to research and evidence synthesis a more rigorous approach is required. The researchers in SE need to be more consistent when performing systematic reviews. There is a good consistency so far in terms of definition of the research questions and search strategies for primary studies. However, SE researchers need to be more consistent in the way they select, characterize, analyze, and synthesize the primary

studies. We suggest that at whatever level 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 the usefulness and value of the findings could 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 ISI Web of Knowledge. This means that any such studies in publication venues not indexed by ISI would not be retrieved, as is the case for, e.g., ACM proceedings papers. Although, a search in the ACM Digital Library to check for this possible bias returned only one additional paper that would have been included in our review. This single paper, which was a scoping study, would not affect the results of the review or change the conclusions. Also, as our focus was on systematic reviews and not on meta-analysis, we did not include 'meta-analysis' as a search term, and 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 about the included primary studies and their methods of synthesis for us to be able 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, there is a potential bias in that one of 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 findings.

6. CONCLUSION

Our tertiary review of the types and methods of synthesis in systematic reviews shows that there is limited attention to research synthesis in SE. Half of the studies that called themselves 'systematic reviews' did not include synthesis, and were rather 'scoping studies' that just 'mapped' out and categorized the primary studies. Furthermore, many of the reviews included primary studies that were either conceptual or did not base their findings on empirical evidence. In addition, as much of two thirds of the SRs did not use synthesis methods tailored to 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 ways of comparing and combining what 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.

7. APPENDIX A: STUDIES INCLUDED

S1. Afzal, W., Torkar, R. & Feldt, R. (2009) A systematic review of search-based testing for non-functional system properties. IST, 51(6): 957-976.

S2. Beecham, S. et al. (2008) Motivation in Software Engineering: A systematic literature review. IST, 50(9-10): 860-878.

S3. Bjørnson, F.O. & Dingsøyr, T. (2008) Knowledge management in software engineering: A systematic review of studied concepts, findings and research methods used. IST, 50(11): 1055-1068.

S4. Davis, A. et al. (2006) Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review. RE'06, pp. 179-188.

S5. Dybå, T. & Dingsøyr, T. (2008) Empirical studies of agile software development: A systematic review. IST, 50(9-10): 833-859.

S6. Dybå, T., Kampenes, V.B. & Sjøberg, D.I.K. (2006) A systematic review of statistical power in software engineering experiments. IST, 48(8), 745-755.

S7. Engstrom, E., Runeson, P. & Skoglund, M. (2010) A systematic review on regression test selection techniques. IST, 52(1): 14-30.

S8. Filho, E. et al. (2008). Evaluating Domain Design Approaches Using Systematic Review. Software Arch. (Vol. 5292, pp. 50-65).

S9. Gomez, O., Oktaba, H., Piattini, M. & Garcia, F. (2006) A systematic review measurement in software engineering - State-of-the-art in measures. ICSOFT 2006, CCIS 10, pp. 165–176.

S10. Hall, T. et al. (2009) A Systematic Review of Theory Use in Studies Investigating the Motivations of Software Engineers. TOSEM, 18(3): 10.1-10.29.

S11. Hannay, J.E., Sjøberg, D.I.K. & Dybå, T. (2007) A systematic review of theory use in software engineering experiments. IEEE TSE, 33(2): 87-107.

S12. Hossain, E., Babar, M.A., & Paik, H. (2009) Using Scrum in Global Software Development: A Systematic Literature Review. ICGSE, 175-184.

S13. Insfran, E., & Fernandez, A. (2008) A systematic review of usability evaluation in Web development. Web Information Systems Engineering, Proceedings (Vol. 5176, pp. 81-91).

S14. Ivarsson, M., & Gorschek, T. (2009) Technology transfer decision support in requirements engineering research: a systematic review of REj. Requirements Engineering Journal, 14(3): 155-175.

S15. Jimenez, M., & Piattini, M. (2008) Problems and Solutions in Distributed Software Development: A Systematic Review. SEAFOOD , Vol. 16, Springer (2008) , p. 107-125.

S16. Jørgensen, M., & Shepperd, M. (2007) A systematic review of software development cost estimation studies. TSE, 33(1): 33-53.

S17. Kampenes, V.B., Dybå, T., Hannay, J.E. & Sjøberg, D.I.K. (2009) A systematic review of quasi-experiments in software engineering. IST, 51(1): 71-82.

S18. Kampenes, V.B., Dybå, T., Hannay, J.E. & Sjøberg, D.I.K. (2007) A systematic review of effect size in software engineering experiments. IST, 49(11-12): 1073-1086.

S19. Khan, S. U., Niazi, M. & Ahmad, R. (2009) Critical Success Factors for Offshore Software Development Outsourcing Vendors: A Systematic Literature Review. ICGSE'09, 207-216.

S20. Khurum, M. & Gorschek, T. (2009) A systematic review of domain analysis solutions for product lines. JSS, 82(12): 1982-2003.

S21. Kitchenham, B.A. et al. (2009) Systematic literature reviews in software engineering – A systematic literature review. IST 51(1): 7-15.

S22. Kitchenham, B.A., Mendes, E. & Travassos, G. (2007) Cross versus within-company cost estimation studies: A systematic review. IEEE TSE, 33(5): 316-329.

S23. Lisboa, L.B. et al. (2010) A systematic review of domain analysis tools. IST, 52(1): 1-13.

S24. Lucas, F., Molina, F. & Toval A. (2009) A systematic review of UML model consistency management. IST, 51(12): 1631-1645.

S25. Mendes, E. (2005) A systematic review of web engineering research. ISESE'05, pp. 498–507.

S26. Morais, Y., Burity, T. & Elias, G. (2009) A Systematic Review of Software Product Lines Applied to Mobile Middleware. Sixth Int. Conf. on Information Technology, pp. 1024-1029.

S27. Nicolas, J. & Toval, A. (2009) On the generation of requirements specifications from software engineering models: A systematic literature review. IST, 51(9): 1291-1307.

S28. Pino, F.J., Garcia, F. & Piattini, M. (2008) Software process improvement in small and medium software enterprises: a systematic review. Software Quality Journal, 16(2): 237-261.

S29. Staples, M. & Niazi, M. (2008) Systematic review of organizational motivations for adopting CMM-based SPI. IST, 50(7-8): 605-620.

S30. Walia, G.S. & Carver, J.C. (2009) A systematic literature review to identify and classify software requirement errors. IST,51(7): 1087-1109.

S31. Williams, B.J. & Carver, J.C. (2010) Characterizing software architecture changes: A systematic review. IST, 52(1): 31-51.

8. REFERENCES

[1] Arksey, H. and O'Malley, L. (2005) "Scoping Studies: Towards a Methodological Framework," *Int. Journal of Social Research Methodology*, 8(1): 19–32.

[2] Braun, V. and Clarke, V. (2006) "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology*, 3: 77–101.

[3] Cohen, B.P. (1989) *Developing Sociological Knowledge: Theory and Method*, 2nd Ed., Chicago: Nelson-Hall.

[4] Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed., Laurence Erlbaum.

[5] Cooper, H., Hedges, L.V., and Valentine, J.C. (Eds.) (2009) *The Handbook of Research Synthesis and Meta-Analysis*, 2nd Ed., Russell Sage Foundation.

[6] Cruzes, D.S, Mendonça, M.G., Basili, V.R., Shull, F., and Jino, M. (2007) Extracting Information from Experimental Software Engineering Papers, *Proc. SCCC'07*, pp. 105–114.

[7] Cruzes, D.S., Basili, V.R., Shull, F., and Jino, M. (2007) "Using Context Distance Measurement to Analyze Results across Studies," *Proc. ESEM'07*, pp. 235–244.

[8] Davies, K., Drey, N., and Gould, D. (2009) "What are Scoping Studies? A Review of the Nursing Literature," *Int. Journal of Nursing Studies*, 46: 1386–1400.

[9] Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., and Sutton, A. (2005) "Synthesising Qualitative and Quantitative Evidence: A Review of Possible Methods," *Journal of Health Services Research & Policy*, 10(1): 45–53.

[10] Dybå, T., Dingsøyr, T., and Hanssen, G.K. (2007) "Applying Systematic Reviews to Diverse Study Types: An Experience Report," *Proc ESEM'07*, pp. 225–234.

[11] Dybå, T., Kitchenham, B.A., and Jørgensen, M. (2005) "Evidence-based Software Engineering for Practitioners," *IEEE Software*, 22(1): 58–65.

[12] Franzosi, R. (2010) *Quantitative Narrative Analysis*, Sage.

[13] Glass, G.V. (1976) "Primary, Secondary, and Meta-Analysis of Research," *Educational Researcher*, 5(10): 3–8.

[14] Kitchenham, B.A. (2004) *Procedures for Performing Systematic Reviews*, Keele University, Technical Report TR/SE-0401 and NICTA Technical Report 0400011T.1.

[15] Kitchenham, B.A. and Charters, S. (2007) *Guidelines for performing Systematic Literature Reviews in Software Engineering*, Version 2.3, Keele University, EBSE Technical Report, EBSE-2007-01.

[16] Kitchenham, B.A., Dybå, T., and Jørgensen, M. (2004) Evidence-based Software Engineering, *Proc. ICSE'04*, Edinburgh, Scotland, 23-28 May, pp. 273–281.

[17] Lipsey, M.W. and Wilson, D.B. (2001) *Practical Meta-Analysis*, Sage.

[18] Miles M.B. and Huberman, A.M. (1994) *Qualitative Data Analysis: An Expanded Source Book*, Sage.

[19] Noblit, G.W. and Hare, R.D. (1988) *Meta-ethnography: Synthesizing Qualitative Studies*, Sage.

[20] Ogawa, R.T., and Malen, B. (1991) "Towards Rigor in Reviews of Multivocal Literatures: Applying the Exploratory Case Study Method, *Review of Educational Research*, 61(3): 265–286.

[21] Paterson, B.L., Thorne, S.E., Canam, C., and Jillings, C. (2001), *Meta-study of Qualitative Health Research: A Practical Guide to Meta-analysis and Meta-synthesis*, Sage.

[22] Pope, C., Mays, N., and Popay, J. (2007) *Synthesizing Qualitative and Quantitative Health Evidence: A Guide to Methods*, Open University Press.

[23] Ragin, C.C. (1987) *The Comparative Method: Moving Beyond Qualitative and Quantitative Strategies*, University of California Press.

[24] Robson, C. (2002) *Real World Research*, 2nd Ed., Blackwell.

[25] Rodgers, M., Sowden, A., Petticrew, M., Arai, L., Roberts, H., Britten, N., and Popay, J. (2009) "Testing Methodological Guidance on the Conduct of Narrative Synthesis in Systematic Reviews," *Evaluation*, 15(1): 49–74.

[26] Rosenthal, R. and DiMatteo, D.M. (2001) "Meta-analysis: Recent Development in Quantitative Methods for Literature Reviews," *Annual Review of Psychology*, 52, 59–82.

[27] Sandelowski, M. and Barroso, J. (2007) *Handbook for Synthesizing Qualitative Research*, Springer.

[28] Sandelowski, M., Docherty, S., & Emden, C. (1997) "Qualitative Metasynthesis: Issues and Techniques," *Research in Nursing and Health*, 20(4): 365–371.

[29] Sjøberg, D, Dybå, T., and Jørgensen, M. (2007) "The Future of Empirical Methods in Software Engineering Research," *Proc. FOSE'07*, pp. 358–378.

[30] Strauss, A. and Corbin, J. (1998) *Basics of Qualitative Research*, 2nd Ed., Sage.

[31] Yin, R.K. and Heald, K.A. (1975) "Using the Case Survey Method to Analyze Policy Studies," *Administrative Science Quarterly*, 20: 371–381.