Synthesizing Qualitative Research in Software Engineering: A Critical Review

Xin Huang[1], He Zhang[1], Xin Zhou[1], Muhammad Ali Babar[2], and Song Yang[1]
[1]State Key Laboratory of Novel Software Technology, Software Institute, Nanjing University, China
[2]School of Computer Science, University of Adelaide, Australia
njuhuangx@outlook.com,hezhang@nju.edu.cn,job@wetist.com,ali.babar@adelaide.edu.au,ysongpray@gmail.com

ABSTRACT

Synthesizing data extracted from primary studies is an integral component of the methodologies in support of Evidence Based Software Engineering (EBSE) such as System Literature Review (SLR). Since a large and increasing number of studies in Software Engineering (SE) incorporate qualitative data, it is important to systematically review and understand different aspects of the Qualitative Research Synthesis (QRS) being used in SE. We have reviewed the use of QRS methods in 328 SLRs published between 2005 and 2015. We also inquired the authors of 274 SLRs to confirm whether or not any QRS methods were used in their respective reviews. 116 of them provided the responses, which were included in our analysis. We found eight QRS methods applied in SE research, two of which, *narrative synthesis* and *thematic synthesis*, have been predominantly adopted by SE researchers for synthesizing qualitative data. Our study determines that a significant amount of missing knowledge and incomplete understanding of the defined QRS methods in the community. Our effort also identifies an initial set factors that may influence the selection and use of appropriate QRS methods in SE.

CCS CONCEPTS

• **General and reference → Empirical studies**;

KEYWORDS

Research synthesis; qualitative (synthesis) methods; systematic (literature) review; evidence-based software engineering;

ACM Reference Format:
Xin Huang[1], He Zhang[1], Xin Zhou[1], Muhammad Ali Babar[2], and Song Yang[1] . 2018. Synthesizing Qualitative Research in Software Engineering: A Critical Review. In *Proceedings of ICSE '18: 40th International Conference on Software Engineering , Gothenburg, Sweden, May 27-June 3, 2018 (ICSE '18),* 12 pages.
https://doi.org/10.1145/3180155.3180235

1 INTRODUCTION

'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.' [10] Synthesizing data is one of the most important tasks of a literature review, either a Systematic Literature Review (SLR) or an ad-hoc literature review. It has been recognized that an SLR is able to produce the highest strength evidence about a particular approach in Software Engineering (SE) [19]. With the increasing use of SLR, as a major research method in support of Evidence-Based Software Engineering (EBSE), SE researchers are expected to synthesize a variety of data extracted from the reviewed primary studies. Given the data reported in SE papers can broadly be classified in two categories: quantitative and qualitative, the data synthesis methods can also be classified as qualitative and quantitative. Compared to the approaches to synthesizing quantitative research data (e.g., meta-analysis), the use of Qualitative Research Synthesis (QRS) methods is relatively new in SE; the increasing trend of SLRs in SE has increased the importance of QRS as a large number of primary studies on many topics of SE incorporate qualitative data. For synthesizing qualitative data extracted from primary studies, SE researchers have been using the methods originally developed in other disciplines such as the social and behavioral sciences. It has been recognized in the community about the usefulness and importance of adoption of qualitative synthesis methods in SE research [58].

However, there exists much confusion about how the various QRS methods compare to each other and the factors that researchers should consider to determine which method best suits their needs, purposes, and ideological stance [31]. Since the quality of the results of a secondary study is largely dependent upon appropriate and reliable data synthesis, it is important to systematically study the use of QRS methods adopted in SE. We aim at rigorously reviewing the state-of-the-art of the use of the QRS methods in SE (through SLRs) and draw some conclusions that can be offered as support to develop methodological guidance for improving the selection, use, and reporting of synthesizing qualitative data in SE.

We carried out a tertiary study of 328 SLRs published in SE between 2005 and 2015. We followed up the initial findings from the SLR by an email-based survey inquiry from the authors of 274 SLRs. Each of the reviewed SLRs has used different methods for synthesizing qualitative data. We have identified that eight methods of QRS have been used in synthesizing qualitative data from the reviewed SLRs. However, most of the SLRs used one of the two QRS methods, *narrative synthesis* and *thematic synthesis*. Our study has identified several instances of lack of knowledge or incomplete understanding of the chosen QRS methods. We have also identified an initial set of factors that appear to have influenced the selection of the particular methods used in the reviewed SLRs. We discuss the findings in a way that can provide useful information about different aspects of selecting, using, and reporting QRS in SE. Our

ICSE '18, May 27-June 3, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5638-1/18/05…$15.00
https://doi.org/10.1145/3180155.3180235

findings can provide useful guidance to SE researchers, particularly PhD candidates and early career researchers, in selecting and using an appropriate QRS method.

The objective of our study is neither to criticize the reviewed SLRs, nor to criticize their authors by any means. Our effort aims to raise sufficient awareness about the problematic situation with regards to synthesizing qualitative data in SE as a result of missing guidance, incomplete understanding, and even misuse of QRS methods in the SE literature. Our contributions are consequently four-fold: we identify a large number of 328 SLRs that synthesized qualitative data and confirmed the use of QRS methods in 76 SLRs as the sample set for in-depth analysis; we report a critical review of the state-of-the-art of the adoption of QRS methods in SE; we investigate the SLR authors' understanding and experience with conducting qualitative data synthesis; we develop an initial set of factors that are likely to influence the choice of QRS methods in SE.

2 QUALITATIVE SYNTHESIS METHODS

The origin of QRS methods is debatable, however, Finfgeld[23] identified five QRS methods in the health and social sciences in 2003. Since then, at least a dozen more QRS methods have been developed [31]. Whilst the QRS methods can be used in various disciplines, there is not much known about how different QRS methods compare with each other despite several researchers have shared their observations and experiences of different QRS methods [23].

Hannes and Lockwood [31] have discussed 20 QRS methods. However, our study focuses on the QRS methods whose use is identified and confirmed in SE: *narrative synthesis, thematic synthesis, meta-ethnography, meta-summary, content analysis, grounded theory, comparative analysis*, and *case survey*. Table 1 presents a summary of the methods that can be categorized based on different attributes, such as *aggregative* as compared to *interpretive*, *epistemology* or *ontology*, *degree of iteration*, and *intended outcomes* [31].

The *aggregative* synthesis methods include obtaining findings of various primary studies and combining their themes with a general description. The *interpretive* synthesis approaches result in the aggregation of a new abstract model or theory of phenomena.The theory considers the research background, interpretations of the findings, and the reporting of the results [27]. The scope of *epistemological* positions extends from *idealism*, and one researcher assumes that all knowledge is based on the fact that a researcher assumes that he / she is looking at the world [59].

Some QRS methods, such as *grounded theory*, is iterative in their processes. Researchers are expected to modify initial decisions and processes while synthesizing data and meeting new data or reviewing primary studies. But *thematic synthesis* provides a highly structured approach to selecting, organizing, and tabulating the primary research data.

QRS methods also change based on their expected results. Some SLRs may intend to inform practice or policy; other may aim at forming a middle level theory and then test it [32, 36]. For example, the use of *grounded theory* is expected to lead to a new theory. Each QRS method fits in a different range between the two poles of these attributes. It is reasonable that some methods might hold a vague position for some attributes. Hence, due to page limit Table 1 just lists certain attributes for the methods to highlight their differences.

We briefly describe the eight QRS methods ever used in SE:

Narrative Synthesis (a.k.a *narrative summary*) features its defining characteristic that is summarized in narrative. It is a general framework of selected narrative descriptions and ordering of primary evidence with commentary. Meanwhile, it is normally combined with specific tools and techniques that help to increase transparency and trustworthiness [56, 59]. This method commonly has four main elements: making a theoretical model that explains how interventions work; developing a preliminary synthesis; exploring relationships within the data; then evaluating the robustness of the outcomes [54]. The top advantage of this method is that it offers high flexibility and is able to cope with diverse evidence types. Its disadvantages may include the lack of transparency and standards as well as the possible bias by prejudices of reviewer.

Thematic Synthesis is a defined method for identifying, analyzing, and reporting themes (patterns) in the data. It minimally organizes and describes the data sets in rich detail to describe data sets and also interprets diverse aspects of the research area. *Thematic synthesis* can be used in different theoretical frameworks. It can be a realist or essentialist inquiry that reports meanings, experiences, and reality. Also it can be a constructionist method to investigate the ways in which meanings, realities, experiences, events, and others impact the range of discourse. *Thematic synthesis* possesses limited interpretative power beyond mere description if it is not used in a theoretical framework [7, 9, 25]. The strengths of this method are the flexible procedures for reviewers that are able to cope well with a large, diverse body of evidence, and the support for theory building. Although its transparency is sometimes criticized, there exist many means to perform this method [9].

Meta-ethnography was first proposed by Noblit and Hare [51] with three different strategies: 1) *reciprocal translation*: translating the different results of the primary studies into a general themes, concepts, or metaphors; 2) *refutational translation*: clearing the contradictions and differences in various studies; 3) *line of argument*: developing a view of the overall phenomenon, through the various parts of the study. It is a translative process that distinguishes this method from others. By translating key concepts and metaphors between different studies, researchers are able to provide new interpretations of primary studies.

Meta-summary and Content Analysis are both quantitative oriented aggregation of qualitative methods and very similar in many aspects [41]. They distinguish the frequency of findings and present the results based on the data. The main difference is the level of data to be synthesized: *content analysis* pays more attention to the raw data; *meta-summary* focuses on higher-level findings [8].

Grounded Theory was proposed by Glaser et. al [26] in 1967. It is a descriptive method to describe qualitative sampling, data collection and data analysis. "*It includes simultaneous phases of data collection and analysis, the use of constant comparison method, the use of theoretical sampling, and the generation of new theory*" [48]. It treats research reports as a data form and can be analyzed to produce higher levels of topics and explanations [38]. The method builds theory from ground up, adjusts theory constantly to fit new data, and finally to get a mature theory. Though *thematic synthesis* also aims to identify, analyze, and report patterns or themes within the data, it lacks the process of adjustment theory that differs from *grounded theory*.

Table 1: Overview of qualitative synthesis methods applied in software engineering

Synthesis method	Features	Attributes	Aim	Example(s)
Narrative synthesis	Narrative description and ordering of primary evidence with commentary	•Interpretive •Epistemology of idealism	An overview of the findings of primary studies is presented, summarizes the main themes, findings and related issues.	[12] [37] [62]
Thematic synthesis	Identifying major or recurring themes in literature and summaries of results of primary studies under the headings of these themes	•Aggregative •Epistemology of realism •Highly structured in data organizing •Outcome utilitarian	Identify, analyze, and report themes or patterns within data	[4][5][64]
Meta-ethnography	"*Interpretations and explanations in the primary studies are treated as data, and are translated across several studies to produce a synthesis*"	•Interpretive •Epistemology of realism	The integration of data from the primary study by means of induction, interpretation, translation, helps to understand and transfer ideas and concepts	[14][57]
Meta-summary	Quantitative oriented aggregation of qualitative findings. Identify the frequency of each discovery, as well as the discovery of high frequency findings	•Aggregative •Epistemology of realism •Outcome theoretical	Discover a pattern or theme in qualitative research based on the higher frequency of findings	[16]
Content analysis	The evidence for each of the primary study is used under a wide range of thematic headings, designed to help with repetitive extraction tools	•Aggregative •Epistemology of realism	Count and tabulate on each occurrence of the theme	[30][39]
Grounded theory	Identifying patterns and relationships in primary data, sampling for analysis, exploring commonalities, and generating theories or models	•Interpretive •Epistemology of realism •Iterative and circular in processes •Outcome theoretical	Generates higher-order themes and interpretations	[50][52]
Comparative analysis	Using Boolean logic (based on specific results of truth tables) to analyze complex causal relationships	•Aggregative •Epistemology of realism	Analyzes complex causal connections	[13]
Case survey	Making closed questions to extract data and each primary study can be seen as a specific case	•Aggregative •Epistemology of realism	Extracted data can be used for further (statistical) analysis	[34]

Comparative Analysis "*requires the construction of a 'truth table', showing all logically possible combinations of the presence and absence of independent variables and the corresponding outcome variable*" [2, 15].

Case Survey proposed by Yin and Heald [65] is a process of systematically coding relevant data from a large number of qualitative cases. Making closed questions to extract data and each primary study can be seen as a specific case for quantitative analysis. The data is extracted from individual case studies using a set of structured close-ended questions. These data are transformed into quantitative forms that enable statistical analysis. The development of this method uses multiple coders to score the cases [15].

3 RESEARCH METHOD

This study was initiated in the middle of 2016 and conducted in two stages, an SLR and a survey inquiry. The review stage followed the SLR guidelines [44]. All the researchers involved in this study have prior experiences with SLRs. Two of the authors were PhD candidates with research topic on Empirical Software Engineering. They had received an extensive training in synthesizing quality data for about two months. This section describes the research method and the process of this study as depicted in Figure 1.

3.1 Research Questions

This study aims at addressing the following research questions:

- **RQ1.** *What methods have been used for synthesizing qualitative research in SE and how they have been used?*
- **RQ2.** *What is the current state of understanding on qualitative synthesis methods in SE?*
- **RQ3.** *What are the factors influencing the selection of qualitative synthesis methods in SE?*

*RQ*1 aims to portray an overview of the adoption of QRS methods in SE by identifying, confirming, and summarizing the SLRs that synthesize qualitative evidence. *RQ*2 looks into how SE researchers acknowledge the QRS methods by claiming and describing their methods of synthesizing qualitative data. The findings for *RQ*3 may contribute to the further methodological recommendations of qualitative synthesis for SE researchers.

3.2 Search Process

Since Kitchenham et al. introduced Evidence-Based Software Engineering (EBSE) in [45] at ICSE 2004 and reported the SLR methodology guidelines in [42], SE researchers have been reporting an increasing number of SLRs. We set 2005 as the starting year for our search of SLRs (similar to [10]) since the published SLRs with reference to the guidelines [42] first appeared in SE in 2005. As this study initiated in the middle of 2016, we did not include the SLRs published in 2016 yet. We did a comprehensive search of the SLRs published in SE from 2005 to 2015 using the 'Quasi-Gold Standard' method [66], which systematically integrates manual and automated search strategies and provides a relatively rigorous approach for search performance evaluation in terms of sensitivity and precision. According to the reported experience of SE community [66], we first chose the SE conferences and journals with emphasis on empirical research as the manual search venues, including TSE, IST, JSS, EMSE, ESEM (former ISESE and METRICS) and EASE, and correspondingly the follow-up automated search venues are IEEExplore, ACM DL, ScienceDirect, and SpringerLink. We borrowed and reused the following search string to search SLRs in SE that has been defined and evaluated in [66] for the automated search.

"*software AND (((systematic OR structured OR exhaustive OR comparative) AND (review OR survey OR map)) OR ((mapping OR scoping) AND study) OR (systematic map) OR (tertiary AND (review OR study) OR meta-analysis))*"

3.3 Study Selection

As shown in Table 2, we formulated the inclusion/exclusion criteria, which were expected to ensure that the identified SLRs are relevant to our research questions. We conducted two phases of study selection: first identifying all SLRs in SE published during the

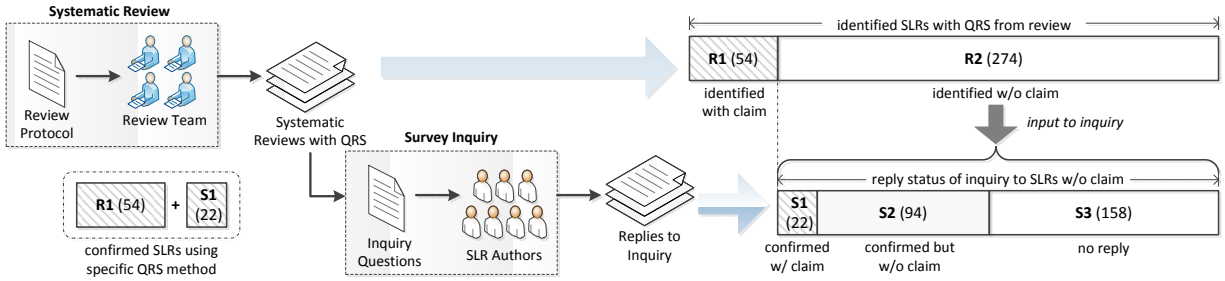

Figure 1: Research process of this study

Table 2: Study inclusion and exclusion criteria

	Inclusion criteria
1.	The authors either claim their study as an SLR in SE, or stated their review followed EBSE seminal papers [20, 45], or SLR guidelines in SE [6, 42, 44] or from other disciplines (e.g., [53])
2.	The papers were published in a refereed conference or journal
3.	The full-text of the paper can be accessible
4.	The paper includes a part of synthesis of qualitative data based on primary studies
5.	The paper either claims the use of a QRS method or the featured characteristic of a QRS method can be identified
	Exclusion criteria
1.	The papers are grey publications, e.g., technical reports, theses or dissertations
2.	The papers are not written in English
3.	The papers are explicitly short papers, position papers, and editorials

Table 3: Data items extracted from SLRs with QRS

RQ	Data item
1,2,3	The methods of qualitative synthesis used in the SLRs
1	Any table or graph used in qualitative synthesis
1,3	Whether the research questions are mainly interpretive or aggregative
1,2,3	Types of the SLRs
1	Published year
1,3	Sample size
1,3	The topics of the SLRs
1,3	Whether and how comparison among primary studies are presented
1,3	Whether and how primary studies are categorized
1,3	Whether and what statistical techniques are used
1,2	Whether and how the authors claim qualitative synthesis method
1,2,3	The data used in the part of qualitative synthesis

Table 4: Questions in survey inquiry

SQ1	*Did you adopt any method of qualitative synthesis in the abovementioned study? If YES, which method and to what extent you know the method?*
SQ2	*Had you ever learnt, and/or had any experience with any method of qualitative synthesis before the study? If YES, which method and from where did you learn it?*
SQ3	*If NO for the last question, have you learnt and adopted any synthesis method after that study? In which way?*

search time span, and then identifying the SLRs that synthesized qualitative research data. In the former phase, four researchers in two groups independently screened over 10,000 hits retrieved from the manual and automated searches through reading the titles, abstracts and keywords, then worked collectively with their selected studies, and finally reach a consensus on 803 published SLRs; in the latter phase, two of the researchers independently read the full text of all the identified SLRs (with special attention on the description of their data synthesis) and carefully checked if any SLRs with QRS. If an SLR does not explicitly claim any QRS methods applied in data synthesis, they checked if any featured characteristic of a defined QRS method could be identified. Such SLRs were marked and later discussed with domain experts in regular meetings until joint decisions could be made. The search and selection process took approximately four months.

3.4 Data Extraction

Table 3 shows the data items extracted from the included SLRs. The left column '*RQ*' indicates the research questions that are expected to be answered with the extracted data item on the right.

At the beginning of the data extraction phase, four researchers randomly selected 40 SLRs with QRS and performed independent reviews as pilot data extraction exercise. The research team held frequent meetings to discuss the extracted data and strove to reach consensus according to our review protocol. After the pilot data extraction, two researchers read the full text of all the identified SLRs with QRS and extracted the data from them independently using the data extraction form (stored into spreadsheets). The extracted data was cross-checked together after independent extractions. Any disagreements were brought into the discussion meetings within other team members. In some cases, expert advice was also sought. Sometimes it was quite difficult to extract the required information about the synthesis methods used from the SLRs as there was no explicit indication about the applied methods. Such cases were analyzed through a few iterations for identifying the most possible synthesis method(s) used by checking their featured characteristics distinct from others. In these cases, we carefully analyzed how the authors described their synthesis process and presented the outcomes from the synthesis, and then made our assessment based on the methodological definitions of different QRS methods. The entire data extraction phase took around four months, including iterations and rework.

3.5 Survey Inquiry

We also decided to survey the authors of the SLRs in which there was no explicit information about the use of QRS methods for synthesizing qualitative data. Note that this inquiry survey intended to merely collect further data for confirmatory purpose, rather than as a standalone research method that is able to generalize the findings from the sample set [24]. We chose to do email based survey consisting of three questions (Table 4): to determine whether or not any QRS method(s) used in the SLRs reported by the respondents, to figure out whether the authors of the identified SLRs possessed an understanding of and experience with the QRS methods when doing their SLRs, and to further check whether the authors knew any QRS methods after finishing their respective SLRs.

3.6 Data Synthesis and Analysis

In order to answer the research questions, we applied both quantitative and qualitative methods in this study for data synthesis (review) and analysis (survey). For *RQ*1, the data was synthesized using *descriptive statistics* and presented in chart and diagram formats for easy understanding. The *narrative synthesis* was also used to answer this research question. The narrative summaries of the findings were described for each of the identified QRS methods.

Given the large sample size and the page limit, we selected a limited representative set of SLRs for demonstrating how qualitative synthesis was performed and reported in SE research. For *RQ*2, *thematic analysis* was used to investigate how SE researchers understand the QRS methods based on the SLR authors' replies to the email inquiry. The descriptions on why and how the researchers adopted QRS methods were compared and analyzed between the identified SLRs and the replies to inquiry. The 'themes' (researchers' typical understanding and experience with QRS) emerged through comparison and categorization. We summarized the state-of-the-art in high-level. For *RQ*3, we used *grounded theory* to generate the factors that would have influenced the researchers' decisions about the selection and use of certain QRS method(s). The candidate factors from the extracted data were proposed, compared, merged, revised, and removed in an iterative synthesis process. The presented results are an initial set of four factors that appear to have influenced the decisions of the adoption of QRS methods in SE.

4 RESULTS

4.1 Literature Review

We identified 328 SLRs with QRS[1], published between 2005 and 2015. Figure 2 shows the number of SLRs with the QRS used and their respective percentages per year. It shows an increasing number of SLRs synthesizing qualitative evidence that were identified in our review (right bar). The number of SLRs claiming and confirming the use of QRS methods (in paper or by email) has also increased (left bar). We note that 30% to 50% of the SLRs published annually have synthesized qualitative data for answering certain research questions. Except 7 SLRs that do not report their sample sizes, the other 321 SLRs synthesize qualitative data from in total 19972 primary studies in SE, which indicates a significant demand of qualitative research in SE.

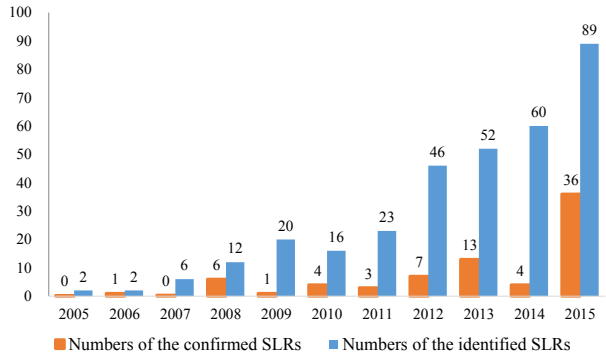

Figure 2: Distribution of SLRs with QRS per year

We found 54 SLRs that report the use of at least one QRS method, in which 30 SLRs explicitly claim the names of the QRS methods used and 24 SLRs report the details of their data synthesis processes that unambiguously indicate the signature practice(s) of certain QRS method(s). For example, the authors claimed the use of 'reciprocal translation' which is one of the three strategic approaches of *meta-ethnography*. There are SLRs whose authors claim the use of two or more QRS methods in their respective SLRs.

[1]The list the reviewed SLRs will be available upon request.

However, a majority of the reviewed SLRs (274 out of 328) do not clearly indicate the QRS methods used in those SLRs. We also found significant discrepancies between what the authors claimed and what they described in some cases. The authors of 6 SLRs claim the use of 'synthesis' methods that have not been defined as QRS methods in literature, such as *vote-counting* [55] and *mapping study*.

4.2 Survey Inquiry

We were unable to confirm the QRS methods used in 274 out of 328 reviewed SLRs. We emailed the authors of these SLRs to ask them about the approach or method they would have used in their respective SLRs. We followed a three phases process. First, we emailed to the first authors of the 274 SLRs with 65 replies received and 42 emails bounced back. Next, we searched the new email addresses of the 42 first authors to whom we resent the emails using their new email addresses. We were successful in delivering the emails to 36 of them, 2 of which replied to us. Last, for the SLRs whose first authors did not reply, we emailed all the other co-authors of those 207 SLRs with 49 replies received. In total, we received the replies to our inquiry from the authors of 116 SLRs.

Our analysis of the answers to the email survey questions enabled us to further determine whether or not the authors used one or more QRS methods in their respective SLRs. For *SQ*1, the authors of the 22 SLRs stated that they had used at least one QRS method in their SLRs: 9 of them indicated the use of *thematic analysis*, 6 claimed the use of *grounded theory* (including its specific techniques), 2 for *narrative synthesis*, 2 for *meta-ethnography*, 2 for *content analysis* and 1 for *meta-summary*. For *SQ*2, we received 19 responses, out which 7 mentioned the use of *grounded theory* and 4 mentioned the use of *thematic synthesis* (or *thematic analysis*). For *SQ*3, we received only 10 replies. Similar to the answers to *SQ*2, *grounded theory* (3 out of 10) and *thematic synthesis* (or *thematic analysis*) (4 out of 10) are the most adopted QRS methods by the respondents.

5 SYNTHESIS AND ANALYSIS

This section answers the research questions by analyzing and synthesizing the data collected from the reviewed papers and the replies to our inquiry respectively. Note that when there is divergence on the used QRS methods between our identification and author's confirmation (via email), we respect and accept the author's claim.

5.1 Qualitative Synthesis Methods (*RQ*1)

Based on the analysis of the data about the 76 SLRs, (R1+S1 in Figure 1), which report the use of certain QRS method(s), we identified the adoption of eight qualitative synthesis methods in SE (Table 1). The discussion on *RQ*1 is based on the synthesis of these 76 SLRs whose synthesis methods were confirmed. Figure 3 shows the frequencies of the use of different QRS methods in these SLRs, in which 8 SLRs used more than one QRS method (mixed-method).

We also classified 76 SLRs in terms of their SE topics. Figure 4 shows the most popular research topics (with 3 SLRs or more) of the confirmed SLRs and indicates the types of QRS methods used. It intends to highlight the distribution of the used QRS methods across the SE topics. It can be observed that the SLRs on '*Human Aspects*' adopted 6 different QRS methods, which indicates the need of using a diverse set of QRS methods for synthesizing qualitative evidence in SE, in particular for the human and social aspects [21].

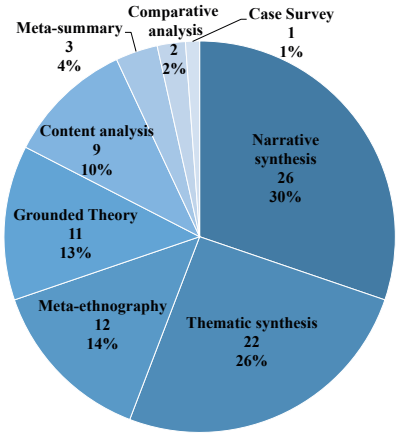

Figure 3: Frequencies of different QRS methods used in SLRs

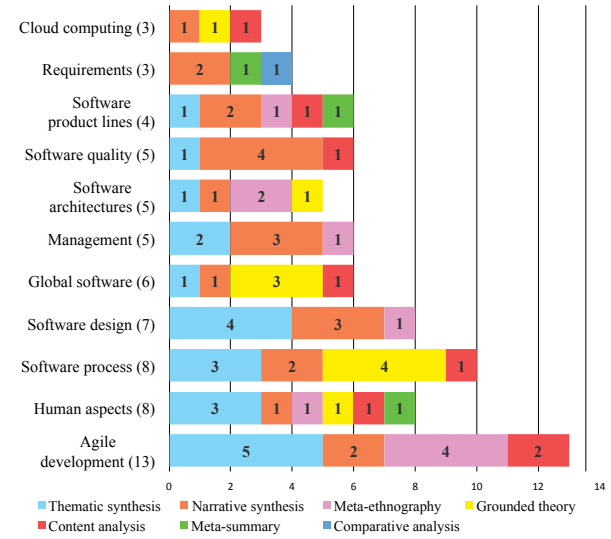

Figure 4: Distribution of QRS methods over SE topics

Of the 252 (S2+S3 in Figure 1) SLRs that could not be confirmed with authors, nearly half of them referenced the SLR guidelines[42, 44] only when reporting the synthesis method applied in them. However, these guidelines do not provide the required instructive information about synthesizing qualitative data.

5.1.1 Narrative Synthesis. We find that 26 (34%) of the SLRs adopted *narrative synthesis* in their QRS. The distinct feature that helps to identify the use of this method is that an SLR reports the findings from the reviewed primary studies in a narrative way as the evidence to support its conclusion. Some SLRs claim the use of *descriptive synthesis*, which is an interpretive means but not a recognized QRS method. As Kitchenham and her colleagues do not differentiate *descriptive synthesis* from *narrative synthesis* in the SLR guidelines [42, 44], we marked the SLRs with the claim of using *descriptive synthesis* and then merged them into *narrative synthesis* for analysis purpose.

There are two common styles of applying *narrative synthesis* in SLRs. The first is to present a narrative summary of the findings from every single primary study. Most of the SLRs using this style usually include a relatively smaller number (< 30) of primary studies than others. The reported details from primary studies vary significantly among these SLRs. In an SLR on aspect-oriented programming [3], the description of the findings from 22 primary studies was presented with rich details. In contrast, Khan et al. [40] reported an SLR on Chidamber and Kemerer metrics suite, where they presented the findings from 20 primary studies quite briefly with no more than a couple of sentences each, such as: "*WMC and WMC-McCabe were significant predictors of faulty classes for six versions of Mozilla Rhino.*"

The other style of using *narrative synthesis* in SLRs is to present the narrative summary of a selective or representative set of primary studies, which often come with a classification of the reviewed primary studies. For example, in an SLR on software analytics to software practice [1], 8 (out of 19) typical primary studies were selected to illustrate their findings in detail for their identified five research domains on software analytics.

Overall, the level of the details of the primary studies conveyed by the narrative synthesis varies a lot: from a couple of sentences for a selective samples to a 'thick' description of each primary study. Tabulation is a common format that SE researchers used to present the details while using *narrative synthesis*. The tabulation helps highlight the similarities and/or differences among the findings of the reviewed primary studies.

5.1.2 Thematic Synthesis. As the centric concept of this method, *themes* progressively emerge by comparing primary studies. *Thematic synthesis* is able to categorize studies and summarize the findings by identifying major or recurrent themes within studies. We find 22 (29%) SLRs using this method. For instance, *thematic synthesis* was adopted in an SLR on global software development [4] to establish a thematic map from the qualitative evidence on the relationship among context dispersion, team coordination and project performance, and further help comprehend the thematic map. With the aim to summarize the evidence to gain an overview of the relationship, the interpretations were based on the themes extracted from the empirical studies.

Whilst this method helps researchers aggregate qualitative evidence, the SLRs adopting this method can also appear interpretive at a detailed level. We find that with the themes emerging from studies, SE researchers elaborate the specifics of primary studies under the themes in a narrative manner. *Themes* can also be organized in a hierarchy or multi-dimension. In an SLR aimed at identifying the existing empirical evidence on the global dispersion dimensions' impact [5], *thematic synthesis* was used to find global dimensions and their measures from 46 primary studies. The relevant codes from the extracted data were merged into hierarchical themes to show the concepts and measures for dispersion dimensions. They further interpreted the findings narratively to ease the understanding.

5.1.3 Meta-ethnography. There are 12 (16%) SLRs that used *meta-ethnography* to understand and transfer ideas, concepts across different studies by induction, interpretation, and translational analysis of the primary studies. The translative process is the signature of this method that distinguishes itself from other QRS methods.

Meta-ethnography can be considered while the researchers aim at addressing conceptual data, such as definitions. For instance, Boer and Farenhorst examined how Architectural Knowledge (AK)

is defined and how different AK definitions are related to each other in [14] in order to obtain a comprehensive understanding of how different researchers view AK. The original definitions were extracted from the primary studies, and then translated into each other to deeply understand and compare these definitions. As a result 14 definitions of AK [14] were obtained from the translation.

It is observable that a few SLRs used *meta-ethnography* as a means to identify the benefits, limitations or challenges in a specific subject area. Dybå and Dingsøyr used this method to investigate what is currently known about the benefits and limitations of, and the strength of evidence for, agile methods [18]. Similarities and differences in the findings are examined in the process of translation and this process can help find the positive and negative dimensions in agile software development for all the key concepts of findings were compared and the opposites could become apparent.

5.1.4 Grounded Theory. Whilst there are 11 (14%) SLRs that claim to have adopted *grounded theory* for QRS, none of these SLRs appears or claims to generate any theory in their research despite the fact that *grounded theory* is a known method for theory generation, There are a few SLRs with the claim of using *grounded theory*; however, our analysis concludes that they have actually used some of the techniques of *grounded theory*, in particular *coding*, rather the complete method. For example, Dutra and Santos claimed they adopted *grounded theory*, specifically coding procedure, as part of the research method to identify and categorize existing risks of software process improvements in source material [17]. First, they marked the *codes* associated with the text propositions of the primary studies; then the relationships between these *codes* were identified by *constant comparison*. However, no theory, framework, or model was generated in their result. Similarly, some replies to our inquiry also claimed the use of the first phase of *grounded theory* to analyze and aggregate data (i.e. to identify codes, concepts and categories) but no theory was generated through those work.

5.1.5 Content Analysis and Meta-summary. These two methods were used in 9 (12%) and 3 (4%) of the SLRs respectively. They both are quantitative oriented aggregation of qualitative evidence as they help discover patterns or themes through discerning the raw data (*content analysis*) or *frequency* of each finding (*meta-summary*). As an example of *meta-summary*, Dutra et al. used this method to synthesize the data describing which teams can be a high performing team in SE by counting the frequencies of possible factors (as findings) which could affect the teamwork to figure out how a team can be a high performance team [16]. As an example of *content analysis*, the researchers counted the frequency of the defined themes from ISO 15939 model (as raw data) to obtain an overview of the measures and indicators on lean software development [22].

The SLRs applying these two methods pay little attention to the textual details of the evidence from a specific or representative set of primary studies, but they present the counted frequencies of qualitative data extracted from the studies. When these methods apply, the statistics are often used to assist the description and representation of the results in a quantitative format.

We also observed a small number of SLRs that applied two or more qualitative synthesis methods, i.e. *comparative analysis* and *case survey*, but each of them was used by no more than two SLRs.

5.1.6 Mixed-method Synthesis. An SLR may apply more than one QRS method when different research questions need to be addressed by applying different QRS methods. We identify 8 (11%) SLRs that have applied two or more QRS methods. These cases are *case survey + meta-ethnography*, *grounded theory + thematic synthesis* (2), *narrative synthesis + meta-ethnography* (2), and *narrative synthesis + meta-summary*. Two SLRs adopt three QRS methods: one with *thematic synthesis + meta-summary + content analysis*, and the other with *narrative synthesis + thematic synthesis + content analysis*. It is observed that narrative synthesis and thematic synthesis are common components of QRS using mixed-method.

Note that five SLRs [3, 33, 35, 47, 61] claim the use of *vote-counting* 'method' for synthesizing the data from the reviewed primary studies. *Vote-counting* is a standalone technique that does not depend on the actual effect size values and comparable metrics. It can be used with several synthesis methods and plays a supporting role. It is a quantitative oriented method to process qualitative data, in which each primary study casts a 'vote' in support of certain relationship and the numbers of votes are counted. With reference to the definitive literature of QRS in other disciplines, *vote-counting* is recognized as a specific practice rather a full-fledged QRS method.

5.2 Understanding of QRS Methods (*RQ2*)

To address the *RQ2*, we used the data collected from the replies to our survey questions compared with the analysis of the data extracted from the identified SLRs. Whilst we could not confirm the specific QRS methods used in the majority of the identified SLRs (S2+S3 in Figure 1), it becomes clear that these SLRs do synthesize qualitative data extracted from the primary studies. From the valid replies of 116 SLRs (S1+S2 in Figure 1), only 22 SLRs (19%) claim the adoption of certain QRS methods. This may imply that most of the SE researchers synthesize qualitative data by following their experience or intuition without guidance of any defined QRS method. The missing of such methodological guidance may result in systematic errors and/or researcher's subjective bias when performing QRS.

SE researchers tend to conduct qualitative data synthesis in their own styles that possibly match some characteristics of certain QRS method. As summarized in Table 5, over half of the replies claimed that they did synthesize qualitative data but did not follow any particular QRS method. For example, one typical reply is, *"In Result section and Conclusion section of the paper, the relevant studies have been thematically categorized and discussed. However, NO particular method has been adopted."*

Table 5: Categorized experiences with QRS from replies

Experience with QRS	#
Did adopt some QRS methods	22
Realize QRS methods, but never adopt any one for some reasons	3
Conducted qualitative synthesis in their own styles w/o following a defined QRS method	34
Only followed Kitchenham's guidelines [42, 44]	15
Adopted systematic mapping study (scoping study)	8
Just say '*No*' (when answering the questions to QRS methods)	31

We find that many respondents did not use specific QRS method(s) because they were not aware of QRS methods at all. Only a few respondents were able to confirm the use of specific QRS methods. Table 5 shows the categorized researchers' experiences that emerged from analyzing the authors' replies via email. In the responses from 94 SLRs that did not claim the use of QRS methods

(S2 in Figure 1), only the authors of 19 SLRs claimed they knew some of the QRS methods. This finding indicate that a large number of SE researchers lack the essential knowledge of QRS methods.

Table 5 indicates that some researchers do not have a good understanding of what a qualitative synthesis method is. Some of the respondents incorrectly perceive mapping study as a QRS method.

We also find that many of the SE researchers have incomplete understandings about the specific QRS methods. For example, some authors claimed that the qualitative synthesis method they adopted is *coding*, which is one common technique of *grounded theory*. According to [60], studies borrowing certain elements or techniques of *ground theory* are not *ground theory* studies.

Many of the SLR authors' knowledge about *meta-ethnography* is limited to what has been introduced in the SLR guidelines[44]. Whilst the guidelines do recommend the use of appropriate approaches like *meta-ethnography* for QRS, they do not intend to provide the detailed instructions on the correct use of this method. As stated in one email: "*The review we performed was based on the guidelines proposed by Kitchenham. They define how to proceed in a qualitative synthesis. I think our work fits in: Line of argument*".

5.3 Factors for Method Selection (*RQ*3)

It is critically important to select a suitable QRS method to ensure the quality of research synthesis for producing qualitative evidence. Based on the information emerged during the data extraction on the confirmed SLRs that applied any QRS method, we identified an initial set of the factors that might have influenced the choice of particular QRS method(s) used in SE in an iterative manner by following a *Ground Thoery* approach. This initial set contains four possibly influential factors that are concisely described below.

Implication of research question. Qualitative synthesis could be *aggregative* or *interpretive*, which is a defining attribute of QRS methods (cf. Table 1). Research is driven by research questions that always shape the evidence generated from primary studies. The research questions of the SLRs (R1+S1 in Figure 1) were coded with their types (e.g., '*how*', '*what*', or '*why*' question). We observe that most of the SLRs (51 out of 76) address '*what*' or '*which*' questions such as "*What creativity techniques for creating more creative requirements are used in industry?*", which indicates the trend of proposing interpretive research questions in these SLRs. This might be a possible reason why the adoption of interpretative QRS methods were more than aggregative methods[2] (as shown in Figure 3).

Evidence into synthesis. The data derived from primary studies can also make a difference in the choice of QRS methods. As known in the community, there are many different research methods for primary research in SE. It might be difficult to compare the studies implementing different research methods and the data in various formats extracted from these primary studies. As described in one email reply: "*This particular study had limited scope for synthesis, although some was performed where we had enough primary studies. However, we could only use narrative synthesis, largely because the variation on the forms of different primary studies made any other form of synthesis impossible. We did not explicitly identify this as having been used in our paper.*". In this case, the different forms of evidence play a great influence on the selection of a particular QRS method for use. For the primary studies with similar or comparable design, the comparison and aggregation of evidence (such as *grounded theory*) turns out to be more straight-forward.

Sample size. When comparing the two most frequently used QRS methods, i.e. *narrative synthesis* and *thematic synthesis*, some noticeable differences can be observed. The box-plots in Figure 5 shows the distributions of sample size over the SLRs using different QRS methods. For each given method (pair), the left plot shows the sample size of the SLRs about which we were able to confirm the used QRS methods (R1+S1 in Figure 1) and the right one shows the sample size of all the identified SLRs (R1+R2) for consistency check. The SLRs adopting *narrative synthesis* tend to have fewer primary studies than those using *thematic synthesis* on average. It can be inferred that the number of studies in an SLR can also influence the choice of QRS method. A small number with no much data to synthesize may lead to the choice of *narrative synthesis*. One email reply claimed that their sample size was relatively small so that *narrative synthesis* was the most suitable method though there were some other methods to be considered. In contrast, as aggregation-oriented methods, *content analysis* and *meta-summary* tend to work with an even larger sample size than other methods. They both discern the frequency of each finding of their primary studies, which require relatively more evidence than others. A small sample size may lower the confidence of an SLR's conclusion because of a high possibility of coincidence when using these methods.

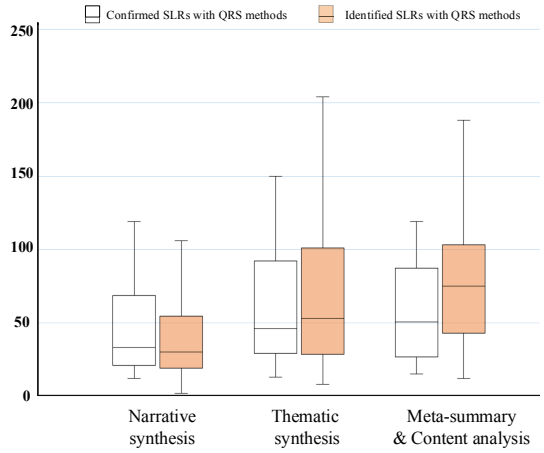

Figure 5: Sample size of SLRs with QRS

Researcher's experience. Given the apparent lack of knowledge of different aspects of established QRS methods among many SE researchers, their personal experience turns to be an important factor when they choose QRS methods. In other words, researchers tend to adopt a QRS method which they used before or they perceive they understand well. One of the replies to our inquiry claimed that they adopted *grounded theory* as they observed the use of this method in some other empirical studies and one of the authors is a specialist in *grounded theory*. In this case, the familiarity of certain method would encourage its reuse in more studies. Moreover, researcher's experience may also help control and secure a study's validity as there is no need to exercise the method in trials.

[2] *Thematic synthesis* can be either aggregative or interpretive.

6 DISCUSSION

" *Make-it-up-as-you-go-along may be OK, but then you have to say: 'I'm making-it-up-as-I-go-along, guys.'* " [49]

6.1 Claiming, Using, and Reporting QRS

SE researchers have been reporting an increasing number of studies with qualitative data, which are expected to be systematically extracted, interpreted, and synthesized for answering the relevant research questions. However, there are significant challenges in appropriately selecting and systematically using QRS methods. Our study has found that the authors of most of the SLRs with qualitative synthesis failed to claim the use of any QRS method or made incorrect claim of the use of a particular QRS method. This situation can have at least two interpretations. One is that there is a general lack of knowledge and understanding (or even lack of awareness) of different aspects of QRS methods. We have stated that all of the QRS methods used in SLRs have been recently adopted in SE from other disciplines such as social and behavioral sciences. The other possible interpretation is that SE researchers tend to perform and present their qualitative synthesis in an ad-hoc manner. We have also found that some SLRs have used more than one method for synthesizing the qualitative data without providing any reasonable justification. We assert that this may have been caused by presenting the synthesis with the authors' own style. When we discuss the overall state of the mentioned methods here, we focus on the synthesis part of these methods despite the fact that some of the QRS methods can cover the whole process of an empirical study. For example, *meta-ethnography* suggests seven phases for researchers to follow which cover the parts of design, analysis and synthesis. We could not find even a single study that had followed exactly the seven phases. When we identified the SLRs adopting this method, we focus on the translative process which is usually reported in the synthesis part of a study.

We have also noted that the quality of the reporting of the used QRS methods is quite poor. This makes the correct determination of the used QRS methods quite difficult. This finding is another clear indicator that there is a general lack of awareness about different aspects of the data synthesis approaches among SE researchers performing SLRs. Given a hugh amount of qualitative data have been ever synthesized from about 20,000 primary studies by SLRs by 2015 in SE (cf. Section 4.1), there is an urgent need of providing guidelines for SE researchers to appropriately select and use a suitable QRS method. We assert that it is important to place greater emphasis on appropriately selecting and correctly using QRS methods in the SLR guidelines for SE. We observe that the existing guidelines (e.g., [44]) for SLRs do not provide much discussion about the research synthesis (QRS in particular) and how to conduct and report it.

6.2 Incomplete Understanding and Misuses

This study has enabled us to conclude that a large number of SE researchers reporting SLRs have incomplete understanding of the used QRS methods in their SLRs. One of the most popular confounded issues is the difference between *narrative synthesis* and *thematic synthesis*. As described in Section 2, *narrative synthesis* is an interpretive approach while *thematic synthesis* is more an aggregative method. In general, *narrative synthesis* pays more attention to the details of the primary studies within a specific context; *thematic synthesis* focuses more on the overview of each *theme* generated from integrating individual studies.

One common issue with *narrative synthesis* is that the description of the findings from the primary studies is superficial. Researchers may describe the findings from primary studies on the high-level with the details and the context information missing. Some authors even just 'copy-paste' the work of their reviewed studies. For example, in [37] *descriptive synthesis* was claimed to reach their findings by synthesizing the available evidence about agile methods in embedded systems development. However, only the study topics were described without providing any details about the findings from the primary studies. *Narrative synthesis* is an interpretive method with more interest in the qualitative details of the findings from studies for a better understanding of the evidence. We can infer that one plausible reason for this issue can be the limited space available for reporting an SLR. In [63], the findings from all 48 primary studies were condensed into a small space. They did not provide much details about the QRS method they used.

In some cases, it is difficult to determine whether an SLR has used *meta-ethnography* or *narrative synthesis* as a result of a lack of clear understanding of the researchers reporting such SLRs. We observe that researchers may ignore the translative step when applying *meta-ethnography*. For example, in an SLR [28] using *meta-ethnography*, the primary studies were classified and synthesized into different definitions of 'Smart Product' literature. In each category, the evidence from the primary studies was presented in a narrative style with little new interpretation provided. When adopting *meta-ethnography*, an SLR has to concentrate on the new interpretation and combine the findings from the studies. We do not mean that all the authors did not conduct the translative process; we just report what we have found from the SLRs which had used one or more QRS method(s). It is possible that the authors using *meta-ethnography* might have conducted the translative process in their work but did not report the detail.

The inappropriate use of a QRS method may lead to serious threats to the validity of an SLR, such as construct, internal, external, and conclusion validities. For instance, as early mentioned, the superficial description of primary studies in an SLR adopting *narrative synthesis* can lead to a lack of evidence which makes the conclusion of an SLR less convincing.

6.3 Limitations

The main possible limitation of our study is the potential bias in the data extraction. Two researchers extracted the data independently in the review. There were some disagreements among researchers when identifying the qualitative synthesis methods used in certain SLRs; most of those disagreements can be attributed to the incomplete reporting or incorrect use of certain QRS methods in SLRs. In such cases, we decided about the used method by identifying the featured characteristics of the QRS methods based on the description provided in SLRs and the extensive discussions about the QRS method that would match the profile described in papers. The follow-up email confirmation with the SLR authors largely mitigated this threat to study validity.

Another limitation of this study might be the possibility of missing the SLRs with QRS during the search and selection phases of this review. It is impossible to identify and retrieve all the relevant

publications for any review study; hence, the solution is a balance and tradeoff between affordable effort and expected quality. However, our search strategy integrates manual and automated search together, and the digital libraries and publications venues we used are comprehensive for SE literature; hence, the number of possibly missing SLRs should be very small.

We acknowledge that the selection and use of qualitative synthesis methods should be put into the specific context containing a number of possible factors. The initial version of the factors does not cover all conditions of researchers, and should be continuously updated when aggregating more SLRs reported and observing new synthesis methods applied in SE in the future.

7 RELATED WORK

Given the increasing trend of SLRs, a few tertiary studies on the methodological aspects of SLRs have been published. Kitchenham et al. have reported two tertiary studies of SLRs in SE [43, 46], which were later extended by da Silva et al. [11]. Zhang and Ali Babar performed another tertiary study [67] to empirically investigate the adoption and use of SLRs in SE.

Cruzes and Dybå [10] reported a tertiary study of the research synthesis approaches used in the SLRs published between 2005 and the middle of 2010. They focused on the research synthesis methods in general (including both qualitative and quantitative methods); our work focuses on the QRS methods because they are diverse but less mature in EBSE compared to the quantitative data synthesis methods. We have reviewed much larger number (328) of SLRs than their work (49). We also used survey inquiry to collect the data from the SLR authors in this study. Another important difference is the viewpoint on mapping study (a.k.a scoping study), which is described as a synthesis method [10]. In the training and preparation sessions prior to this review, we consulted many definitive and seminal works of QRS but none of them identifies mapping study as a synthesis method except [10]. With reference to the SLR guidelines [44], a mapping study is used to gain a broad overview of a specific topic area of interest. Accordingly, mapping study is recognized as a study type instead of a QRS method in our work. Each of the SLRs was exclusively classified into one single synthesis method in [10]. We took a more realistic view that an SLR may use multiple QRS methods in terms of its research questions.

Some results from our review confirm the findings from Cruzes and Dybå's study. For instance, *narrative synthesis* and *thematic synthesis* were the two most frequently used QRS methods; almost half of the SLRs in [10] and a majority (77%) of the SLRs in our review provide no indication of a synthesis method being followed, and just a few SLRs explain the synthesis methods used in detail. In contrast, many authors explained how the extraction of the data was performed but describe little about their synthesis procedures. Thus, it can be concluded that there has been no significant improvements observed since the problematic issues with research synthesis in SE were reported in [10].

Guzmán et al. [29] performed an online survey to identify difficulties experienced when synthesizing evidence and reported their results in 2014. The sample of their survey includes 113 authors of the published SLRs and 49 ISERN members. They found that the state (i.e. availability, heterogeneity, and quality) of primary study and the low quality of reports are perceived as the most important difficulties by the respondents, and a methodological support for selecting and applying a suitable synthesis method is necessary.

Stol et al. [60] reported a review on *grounded theory* in SE which concludes that the improvements of the quality of both conducting and reporting *grounded theory* studies are necessary. Based on the data extracted from the 98 journal articles, they found that many studies present little to no details about the use of *grounded theory*; only a few studies describe or confirm the use of the key practices, many studies ignore the *grounded theory* variants that are significantly different from each other, and few "*grounded theory*" studies generated theory. Their findings imply that SE researchers may not have a good understanding of *grounded theory*, which conforms to the state-of-the-art of QRS reported in this study.

8 CONCLUSION

This paper reports an empirical investigation of the use of QRS methods in SLRs and an email-based survey inquiry (for data collection only). Based on a review of 328 SLRs, our study reveals (cf. Fig. 2) that there is a growing interest and exercise of QRS in SE over the past decade. We have observed that among the eight QRS methods used in the 76 SLRs confirmed with the authors, four QRS methods (*narrative synthesis*, *thematic synthesis*, *meta-ethnography* and *grounded theory*) were most used. It is disappointing to find that only part of SE researchers have a good knowledge and appropriate understanding of the QRS methods. We also found that some researchers incorrectly used certain methods for synthesizing qualitative data extracted from their reviewed studies. The common problem appears to be a confusion between *narrative synthesis* and *thematic synthesis*, *narrative synthesis* and *meta-ethnography*. Based on our findings, we summarize an initial set of four factors that may affect the selection of an appropriate QRS method for SE research: *implication of research question*, *evidence into synthesis*, *sample size* and *researcher's experience*. By combining the theoretical definitions and characteristics of the QRS methods, these factors discovered through our study will contribute to the methodological guidance for choosing an appropriate QRS method.

Whilst our findings can warn the community of the incomplete understanding and missing guidance on qualitative synthesis, as well as support for SE researchers in selecting an appropriate QRS method and correctly applying the selected method for synthesizing qualitative evidence. We do not claim that we have covered all the possible challenges in synthesizing qualitative data. Hence, our findings and recommendations may be more helpful to the novice and young researchers (e.g., PhD candidates or post-docs). We also note that the researchers have little knowledge and experience of the available synthesis methods in general and synthesizing qualitative data in particular. Since the qualitative data synthesis is becoming increasingly important in our discipline, SE researchers need to receive relevant materials and training in different QRS methods suitable for synthesizing qualitative research in SE. Hence, we can conclude that there is a critical need of further research to build the methodological body of knowledge about selecting, using, and reporting appropriate QRS methods for synthesizing qualitative research in SE.

REFERENCES

[1] Tamer Mohamed Abdellatif, Luiz Fernando Capretz, and Danny Ho. 2015. Software Analytics to Software Practice: A Systematic Literature Review. In *Proceedings of the First International Workshop on BIG Data Software Engineering*. 30–36.

[2] Peter Abell. 1984. Comparative Narratives: Some Rules for the Study of Action. *Journal for the Theory of Social Behaviour* 14, 3 (Oct. 1984), 309–331.

[3] Muhammad Sarmad Ali, Muhammad Ali Babar, Lianping Chen, and Klaas-Jan Stol. 2010. A systematic review of comparative evidence of aspect-oriented programming. *Information & Software Technology* 52, 9 (Sept. 2010), 871–887.

[4] Nguyen-Duc Anh, Daniela S. Cruzes, and Reidar Conradi. 2012. Dispersion, coordination and performance in global software teams: A systematic review. In *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*. ACM, 129–138.

[5] Nguyen-Duc Anh, Daniela S. Cruzes, and Reidar Conradi. 2014. The impact of global dispersion on coordination, team performance and software quality - A systematic literature review. *Information & Software Technology* 57, 1 (Jan. 2014), 277–294.

[6] Jorge Biolchini, Paula Gomes Mian, Ana Candida Cruz Natali, and Guilherme Horta Travassos. 2005. *Systematic Review in Software Engineering*. Technical Report. Universidade Federal do Rio de Janeiro.

[7] Virginia Brauna and Victoria Clarkeb. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* (2006), 77–101.

[8] Stephen Cavanagh. 1997. Content analysis: concepts, methods and applications. *Nurse Researcher* 4, 3 (May. 1997), 5–13.

[9] Daniela S. Cruzes and Tore Dybå. 2011. Recommended steps for thematic synthesis in software engineering. In *Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on*. IEEE, 275–284.

[10] Daniela S. Cruzes and Tore Dybå. 2011. Research synthesis in software engineering: A tertiary study. *Information & Software Technology* 53, 5 (May. 2011), 440–455.

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

[12] Ivonei Freitas da Silva, Paulo Anselmo da Mota Silveira Neto, Pádraig O'Leary, Eduardo Santana de Almeida, and Silvio Romero de Lemos Meira. 2011. Agile software product lines: a systematic mapping study. *Software: Practice & Experience* 41, 8 (2011), 899–920.

[13] Alan Davis, Oscar Dieste, Ann Hickey, Natalia Juristo, and Ana M. Moreno. 2006. Effectiveness of requirements elicitation techniques: empirical results derived from a systematic review. In *Requirements Engineering, 14th IEEE International Conference*. IEEE, 179–188.

[14] Remco C. de Boer and Rik Farenhorst. 2008. In search of 'architectural knowledge'. In *Proceedings of the 3rd international workshop on Sharing and reusing architectural knowledge*. ACM, Leipzig, Germany, 71–78.

[15] Mary Dixon-Woods, Shona Agarwal, David Jones, Bridget Young, and Alex Sutton. 2005. Synthesising qualitative and quantitative evidence: a review of possible methods. *Journal of health services research & policy* 10, 1 (2005), 45–53B.

[16] Alessandra Costa Smolenaars Dutra, Rafael Prikladnicki, and César França. 2015. What do we know about high performance teams in software engineering? Results from a systematic literature review. In *Software Engineering and Advanced Applications (SEAA), 2015 41st Euromicro Conference on*. IEEE, 183–190.

[17] Eliezer Dutra and Gleison Santos. 2015. Software process improvement implementation risks: A qualitative study based on software development maturity models implementations in Brazil. In *International Conference on Product-Focused Software Process Improvement*. Springer, 43–60.

[18] Tore Dybå and Torgeir Dingsøyr. 2008. Empirical studies of agile software development: A systematic review. *Information & Software Technology* 50, 9 (Aug. 2008), 833–859.

[19] Tore Dybå and Torgeir Dingsøyr. 2008. Strength of evidence in systematic reviews in software engineering. In *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement*. ACM, 178–187.

[20] Tore Dybå, Barbara Kitchenham, and Magne Jørgensen. 2005. Evidence-Based Software Engineering for Practitioners. *IEEE Software* 22, 1 (Jan. 2005), 158–165.

[21] Tore Dybå, Rafael Prikladnicki, Kari Rönkkö, Carolyn Seaman, and Jonathan Sillito. 2011. Special issue on qualitative research methods in software engineering. *Empirical Software Engineering* 16, 2 (2011).

[22] Markus Feyh and Kai Petersen. 2013. Lean Software Development Measures and Indicators-A Systematic Mapping Study. In *Lean Enterprise Software and Systems*. Springer, 32–47.

[23] Deborah L. Finfgeld. 2003. Metasynthesis: the state of the art–so far. *Qualitative Health Research* 13, 7 (Sept. 2003), 893–904.

[24] Floyd J. Fowler. 2013. *Survey Research Methods*. SAGE Publications.

[25] Jo Garcia, Leanne Bricker, Jane Henderson, Marie Anne Martin, Mira Mugford, Jim Nielson, and Tracy Roberts. 2002. Women's Views of Pregnancy Ultrasound: A Systematic Review. *Birth* 29, 4 (2002), 225–250.

[26] Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. 1967. The Discovery of Grounded Theory; Strategies for Qualitative Research. *Nursing Research* 17, 4 (Jul. 1967), 364.

[27] David Gough and Diana Elbourne. 2002. Systematic Research Synthesis to Inform Policy, Practice and Democratic Debate. *Social Policy & Society* 1, 03 (Jun. 2002), 225–236.

[28] César Gutierrez, Juan Garbajosa, Jessica Díaz, and Agustín Yagüe. 2013. Providing a Consensus Definition for the Term "Smart Product". In *Engineering of Computer Based Systems (ECBS), 20th IEEE International Conference and Workshops on the*. IEEE, 203–211.

[29] Liliana Guzmán, Constanza Lampasona, Carolyn B. Seaman, and Dieter Rombach. 2014. Survey on Research Synthesis in Software Engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14)*. ACM, 2:1–2:10.

[30] Jo E. Hannay and Magne Jørgensen. 2008. The role of deliberate artificial design elements in software engineering experiments. *Software Engineering, IEEE Transactions on* 34, 2 (Mar. 2008), 242–259.

[31] Karin Hannes and Craig Lockwood. 2011. *Synthesizing qualitative research: Choosing the right approach*. John Wiley & Sons.

[32] A Harden, J Garcia, S Oliver, R Rees, J Shepherd, G Brunton, and A Oakley. 2004. Applying systematic review methods to studies of people's views: an example from public health research. *Journal of Epidemiology & Community Health* 58, 9 (Sept. 2004), 794–800.

[33] Ali Idri, Fatima Azzahra Amazal, and Alain Abran. 2015. Analogy-based software development effort estimation: A systematic mapping and review. *Information & Software Technology* 58 (Feb. 2015), 206–230.

[34] Martin Ivarsson and Tony Gorschek. 2009. Technology transfer decision support in requirements engineering research: A systematic review of REj. *Requirements Engineering* 14, 3 (2009), 155–175.

[35] Ronald Jabangwe, Jürgen Börstler, Darja Šmite, and Claes Wohlin. 2015. Empirical evidence on the link between object-oriented measures and external quality attributes: a systematic literature review. *Empirical Software Engineering* 20, 3 (Mar. 2015), 640–693.

[36] Thomas James, Sutcliffe Katy, Angela Harden, Oakley Ann, Oliver Sandy, Rees Rebecca, Brunton Ginny, and Kavanagh Josephine. 2003. Children and healthy eating: a systematic review of barriers and facilitators. *EPPI-Centre, Institute of Education, University of London* (2003).

[37] Matti Kaisti, Ville Rantala, Tapio Mujunen, Sami Hyrynsalmi, Kaisa Könnölä, Tuomas Mäkilä, and Teijo Lehtonen. 2013. Agile methods for embedded systems development-a literature review and a mapping study. *EURASIP Journal on Embedded Systems* 2013, 1 (Dec. 2013), 1–16.

[38] Margaret H. Kearney. 1998. Ready-to-wear: Discovering grounded formal theory. *Research in Nursing & Health* 21, 2 (1998), 179–186.

[39] Siffat Ullah Khan, Mahmood Niazi, and Rashid Ahmad. 2011. Factors influencing clients in the selection of offshore software outsourcing vendors: An exploratory study using a systematic literature review. *Journal of systems & software* 84, 4 (Apr. 2011), 686–699.

[40] Yasser A. Khan, Mahmoud O. Elish, and Mohamed El-Attar. 2012. A systematic review on the impact of CK metrics on the functional correctness of object-oriented classes. In *Computational Science and Its Applications–ICCSA 2012*. Springer, 258–273.

[41] Cheryl Killion. 2008. Handbook for Synthesizing Qualitative Research. *Nursing Education Perspectives* 29, 3 (May. 2008), 176–177.

[42] Barbara Kitchenham. 2004. *Procedures for Undertaking Systematic Reviews*. Technical Report. Computer Science Department, Keele University and National ICT Australia.

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

[44] Barbara Kitchenham and Stuart Charters. 2007. *Guidelines for Performing Systematic Literature Reviews in Software Engineering (version 2.3)*. Technical Report. Keele University and University of Durham.

[45] Barbara Kitchenham, Tore Dybå, and Magne Jørgensen. 2004. Evidence-based Software Engineering. In *26th International Conference on Software Engineering (ICSE'04)*. IEEE Computer Society, Edinburgh, Scotland, UK, 273–281.

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

[47] Parastoo Mohagheghi and Reidar Conradi. 2007. Quality, productivity and economic benefits of software reuse: A review of industrial studies. *Empirical Software Engineering* 12, 5 (Oct. 2007), 471–516.

[48] Gareth Morgan. 1983. *Beyond method*. SAGE Publications.

[49] Janice M Morse. 1994. *Critical issues in qualitative research methods*. SAGE Publications.

[50] Srinivas Nidhra, Muralidhar Yanamadala, Wasif Afzal, and Richard Torkar. 2013. Knowledge transfer challenges and mitigation strategies in global software development?A systematic literature review and industrial validation. *International journal of information management* 33, 2 (2013), 333–355.
[51] George W. Noblit and R. Dwight. Hare. 1988. Meta-Ethnography: Synthesizing Qualitative Studies. *Sage Pubn Inc* (1988).
[52] Tanay Kanti Paul and Man Fai Lau. 2012. Redefinition of fault classes in logic expressions. In *Quality Software (QSIC), 2012 12th International Conference on*. IEEE, 144–153.
[53] Mark Petticrew and Helen Roberts. 2006. *Systematic reviews in the social sciences: A practical guide*. Wiley Blackwell.
[54] J Popay, H Roberts, A Sowden, M Petticrew, N Britten, Lisa Arai, K Roen, and M Rodgers. 2010. Developing methods for the narrative synthesis of quantitative and qualitative data in systematic reviews of effects. In *ESRC*. Centre for Review and Dissemination.
[55] J Richard and David B Pillemer. 1984. *Summing Up: The Science of Reviewing Research*. Harvard University Press.
[56] Mark Rodgers. 2009. Testing Methodological Guidance On The Conduct Of Narrative Synthesis In Systematic Reviews. *Evaluation* 15, 1 (Jan. 2009), 49–73.
[57] Fernando Selleri Silva, Felipe Santana Furtado Soares, Angela Lima Peres, Ivanildo Monteiro de Azevedo, Ana Paula L. F. Vasconcelos, Fernando Kenji Kamei, and Silvio Romero de Lemos Meira. 2015. Using CMMI together with agile software development: A systematic review. *Information & Software Technology* 58 (Feb. 2015), 20–43.
[58] Dag IK Sjøberg, Tore Dybå, and Magne Jørgensen. 2007. The future of empirical methods in software engineering research. In *Future of Software Engineering, 2007. FOSE'07*. IEEE, 358–378.
[59] Liz Spencer, Jane Ritchie, Jane Lewis, and Lucy Dillon. 2003. Quality in Qualitative Evaluation: A framework for assessing research evidence. *London United Kingdom Government Chief Social Researcher's Office Aug* (Aug. 2003).
[60] Klaas-Jan Stol, Paul Ralph, and Brian Fitzgerald. 2016. Grounded theory in software engineering research: a critical review and guidelines. In *Proceedings of the 38th International Conference on Software Engineering*. ACM, 120–131.
[61] Mark Turner, Barbara Kitchenham, Pearl Brereton, Stuart Charters, and David Budgen. 2010. Does the technology acceptance model predict actual use? A systematic literature review. *Information & Software Technology* 52, 5 (May. 2010), 463–479.
[62] Darja Šmite, Claes Wohlin, Tony Gorschek, and Robert Feldt. 2010. Empirical evidence in global software engineering: a systematic review. *Empirical Software Engineering* 15, 1 (Feb. 2010), 91–118.
[63] Igor Scaliante Wiese, Filipe Roseiro Côgo, Reginaldo Ré, Igor Steinmacher, and Marco Aurélio Gerosa. 2014. Social metrics included in prediction models on software engineering: a mapping study. In *Proceedings of the 10th International Conference on Predictive Models in Software Engineering*. ACM, 72–81.
[64] Charlie Wilson, Tom Hargreaves, and Richard Hauxwell-Baldwin. 2015. Smart homes and their users: a systematic analysis and key challenges. *Personal & Ubiquitous Computing* 19, 2 (2015), 463–476.
[65] Robert K. Yin and Karen A. Heald. 1975. Using the case survey method to analyze policy studies. *Administrative science quarterly* (Sept. 1975), 371–381.
[66] He Zhang and Muhammad Ali Babar. 2011. Identifying Relevant Studies in Software Engineering. *Information & Software Technology* 53, 6 (Jun. 2011), 625–637.
[67] He Zhang and Muhammad Ali Babar. 2013. Systematic reviews in software engineering: An empirical investigation. *Information & Software Technology* 55, 7 (2013), 1341–1354.