An Empirical Investigation of Systematic Reviews in Software Engineering

He Zhang
National ICT Australia
University of New South Wales, Australia
Email: he.zhang@nicta.com.au

Muhammad Ali Babar
IT University of Copenhagen
Denmark
Email: maba@itu.dk

***Abstract*—BACKGROUND: Systematic Literature Reviews (SLRs) have gained significant popularity among software engineering (SE) researchers since 2004. Several researchers have also been working on improving the scientific and technological support for SLRs in SE. We argue that there is also an essential need for evidence-based body of knowledge about different aspects of the adoption of SLRs in SE.**
OBJECTIVE: The main objective of this research is to empirically investigate the adoption and use of SLRs in SE research from various perspectives.
METHOD: We used multi-method approach as it is based on a combination of complementary research methods which are expected to compensate each others' limitations.
RESULTS: A large majority of the participants are convinced of the value of using a rigorous and systematic methodology for literature reviews. However, there are concerns about the required time and resources for SLRs. One of the most important motivators for performing SLRs is new findings and inception of innovative ideas for further research. The reported SLRs are more influential compared to the traditional literature reviews in terms of number of citations. One of the main challenges of conducting SLRs is drawing a balance between rigor and required effort.
CONCLUSIONS: SLR has become a popular research methodology for conducting literature review and evidence aggregation in SE. There is an overall positive perception about this methodology. The findings provide interesting insights into different aspects of SLRs. We expect that the findings can provide valuable information to readers on what can be expected from conducting SLRs and the potential impact of such reviews.

***Keywords*-Systematic (literature) reviews; evidence-based software engineering; methodology adoption;**

I. Introduction

Systematic Literature Reviews (SLRs, also referred as systematic reviews) aim to identify, assess and combine the evidence from primary research studies using an explicit and rigorous method. This methodology is widely implemented in some disciplines, such as medicine and sociology. Since Kitchenham et al. published the seminal paper [14] of Evidence-Based Software Engineering (EBSE) in 2004, systematic review has become an important research methodology of EBSE, and many SLRs have been conducted and reported in Software Engineering (SE).

This paper focuses on a multi-perspective of SLRs in SE over the past seven years, and offers the readers a broad but real perspective of its adoption. It shows the 'solid work' done through SLR methodology and, more importantly, shares the perceptions and experiences from the methodology users, *systematic reviewers*, as well as intends to enhance the communications among SE researchers regarding SLRs and EBSE. To be specific this study makes the following contributions:

- It provides a holistic status report on the adoption of SLRs in SE research from various perspectives after seven years' practice.
- It presents a systematic reflection of the methodology adoption based on the outcomes (SLRs) and their producers.
- It also reports an initial comparison of the use and potential value of two literature review methodologies in SE: SLR and Traditional Literature Review (TLR).

In order to minimize the potential limitations of applying single research method, this research combined multiple empirical research methods, and integrated them systematically. For the reported research, we employed *semi-structured interviews*, *tertiary study* (using SLR methodology), and *questionnaire based surveys*. The use of multi-method approach enabled us to effectively build the evidence upon each other for our goals.

Unlike other 'roadmap' or 'overview' papers, this paper presents and summarizes the evidence we collected from the real state of SLR's adoption in SE and the responses from the methodology users without author's subjective inputs. Hence, it is expected to enable readers to make their own conclusion based on these reflections with a minimized bias.

The rest of this paper is structured as follows. Section 2 sets the research context related to EBSE and SLR, and defines our research scope and intents. We enumerate the multi-perspective research questions, and describe our research design and implementation in Section 3. Section 4, 5 and 6 address the preliminary answers to the research questions about SLR's adoption in SE respectively. They are followed by the reflections and discussions in Section 7. In the end, we conclude our study in Section 8.

978-0-7695-4604-9/11 $26.00 © 2011 IEEE
DOI 10.1109/ESEM.2011.17

II. Research Context and Intents

This section introduces the context and motivation of the reported research by discussing the role of EBSE and SLRs in SE. We also define the scope of the reported research.

A. EBSE and Systematic Reviews

The main objective of EBSE is reported to be improving decision making for selecting software development technologies (i.e., methods, approaches, and tools) by gathering, evaluating, and synthesizing current available evidence from research. EBSE is expected to narrow the gap between research and practice by placing emphasis on scientific rigor as well as practical relevance of research [9].

Systematic reviews are predominantly used for following EBSE. An SLR is "*a means of evaluating and interpreting all available research relevant to a particular research question, topic area or phenomenon of interest*" [9], [13]. One of the main goals of an SLR is to ensure that the review is methodical, repeatable, and thorough. A systematic review also attempts to minimize the level of bias that can be prevalent in Traditional (ad hoc) Literature Reviews (TLRs).

There are an increasing number of SLRs being performed in SE since 2004. SE researchers have also provided methodological support by developing guidelines for performing SLRs ([13], [3]) and have reported lessons learned in order to share knowledge and experiences ([4], [8]). Several researchers have also identified the areas for improving the published guidelines and the needs for supportive techniques [25]. The increasing trend to use SLRs highlights the need for providing appropriate knowledge and training in different aspects of SLRs [20]. That means SE researchers need to allocate significantly more resources to develop suitable support system for guiding SE researchers on how to design, conduct, and report high quality systematic reviews in SE and practitioners on how to assess the quality and results of published SLRs on a topic that interests them. At the same time, there is also a need for understanding the use and adoption of systematic reviews in SE and any challenges that researchers are facing while using this methodology.

B. Research Scope and Intents

SLRs promise to provide the mechanism needed to assist practitioners to adopt appropriate technologies [14], [9]. However, this research is not aimed at exploring the use and adoption of systematic reviews by SE practitioners. Rather the focus of the reported research at this stage is the SE research community.

There are two main reasons that exclude the external validation of SLRs from our research scope at this stage. First, as SLR was introduced to SE as a new research methodology in 2004, many practitioners in software industry have yet to know this research methodology well (this is also true even for many SE researchers). Second, though many SLRs have been reported in SE in the recent years, the number of explored topics are still limited, and the distribution of SLRs over these topics are not oven. The technology practitioners may be unable to freely find an SLR, which exactly matches their own interests or questions at present. Given these two constraints, we are not in a position to properly examine the SLRs' external relevance at this stage.

Hence, the objective of the reported research is to carry out an internal validation of SLR methodology within SE research community. Unlike in medical discipline, researchers are expected to be the main users of SLR methodology and evidence-based practice in SE. Therefore, an internal validation of this methodology should necessarily start within SE research community, particularly the literature reviewers irrespective of whether or not they have used SLR.

This research has been motivated by an increasing recognition of the need and importance of providing methodological and technological infrastructures for maximizing the exploitation of potential benefits of EBSE in general and SLRs in particular. We assert that any such effort would greatly benefit from a good understanding of the use, value, and challenges involved in performing SLRs in SE. Thus, we decided to systematically study the adoption and use of SLR in SE using multiple research methods such as tertiary study and surveys (interviews and questionnaire based).

C. Tertiary Studies on Systematic Reviews in SE

Kitchenham et al. conducted a continuous tertiary study to provide an overview of the secondary studies (i.e., SLRs and meta-analyses) related to EBSE in 2007 [12] and 2009 [15]. Cruzes and Dybå performed another tertiary study to assess the types and methods of research synthesis in systematic reviews in SE [5]. However, these tertiary studies merely focused on the SLRs published in SE. Our research sought both systematic reviews (SLRs) and traditional reviews (TLRs) as a basis for the methodology adoption and impact analysis. We also covered the methodology users' perceptions and experiences, which are also important for a new methodology adoption. Hence, our study is able to provide a more holistic view of SLR methodology adoption in SE.

In addition, compared to the SLRs identified in the other tertiary studies ([12], [15], [5]), our research systematically produced a more recent and comprehensive list of SLRs available to SE researchers and practitioners.

III. Research Design and Implementation

A. Research Questions

As the intended post-mortem review of the past seven years' adoption of SLR in SE, we investigated this methodology from a multi-perspective. The research questions discussed in this paper are as below.

- RQ1. What is the value of SLR for SE? Why did SE researchers do (not) SLRs?
- RQ2. What SE topics have been targeted by SLRs? What has the influence of SLRs been in SE research?

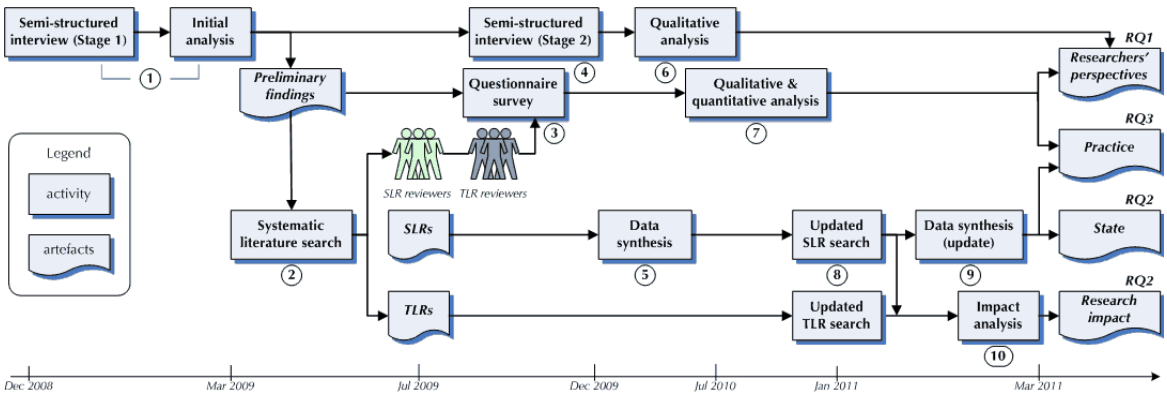

Figure 1. Research design and implementation process

RQ3. How did SE researchers perform SLRs in terms of, for example, rigor and effort?

After the implementation of the research design (described in the rest of this section), the above questions are answered in the following sections.

B. Research Design

This research explicitly distinguishes two types of literature reviews: *systematic reviews* and *traditional reviews*, which are carried out in an informal or ad hoc style. Accordingly, the literature review performers can be grouped as *systematic reviewers*, who use the SLR methodology for performing literature reviews; *traditional reviewers*, who undertake the traditional (narrative) style literature reviews.

Our research design was based on one important assumption that each of the (co-)authors of an SLR has participated in at least one phase of the reported SLR (e.g., *reporting review*) as a systematic reviewer. A similar assumption has been made about the traditional reviewers.

Figure 1 illustrates the used research methods and process. In order to systematically investigate the use of SLRs in SE, our research design incorporated three empirical research methods: **semi-structured interview**, **systematic review** (tertiary study), and **questionnaire-based survey**.

According to the research design, this study was composed of a number of activities, each activity was supported by one empirical method. They were sequentially connected to form the research process. One activity was based on its former activities, and the outputs from one activity might become inputs to the following activities. In the rest of this paper, we use the label numbers in Firgure 1 to indicate individual research activity of the research process.

In planning this research, we decided to start with exploring the perceptions and experiences of systematic reviewers using interview research method. The primary findings from the analysis of the data gathered through the first stage of interviews were reported in [1].

Afterward, we conducted a tertiary study with the research questions based on the '*preliminary findings*'. This phase was a *systematic review* of SLRs reported by that time (the middle of 2009). Different from the tertiary studies reported by Kitchenham et al. [12], [15], the search performed in this study had a broader scope (seeking both SLRs and TLRs) and were later updated till the end of 2010. Both types of literature reviews and their corresponding performers, i.e. *systematic reviewers* and *traditional reviewers*, were identified and grouped after the searches. We extracted and synthesized the data from the identified SLRs only.

Based on the *preliminary findings* from the first stage interview, we designed two questionnaires, one for *systematic reviewers* and the other for *traditional reviewers*. The questionnaires were published as web surveys after a short internal trial. The two groups of reviewers, whose contact details were extracted from the tertiary study, were invited to participate in the surveys. The gathered qualitative and quantitative data were analyzed respectively. In the meantime, the citations of the identified SLRs and TLRs were collected for the impact analysis.

Finally, the results gathered from different research methods were put together to generate the findings: value and motivation (by combining the reflection from the interviews and surveys), topic and impact (from the tertiary study and impact analysis), and practice and experience (from the surveys and tertiary study). These findings provide the evidence for answering the research questions of SLR's adoption in software engineering.

C. Research Implementation

This section describes the technical process of the implementation of our multi-method research project. We can only provide a brief description of each research method used in this research program because of the limited space.

1) Interviews: Based on the pilot literature search and our knowledge of the researchers active in SLRs, we invited a number of researchers in SE to participate in two stages of interviews. The invitees can be classified into three categories: ***advocates*** who introduced SLR methodology and evidence-based practice into SE and published many SLRs, ***followers*** who were experienced SE researchers and had

participated in at least one SLR, and ***novices*** who were research students when performing SLR(s) [1].

In the first stage, we invited 24 researchers, and 17 invitees agreed to be interviewed (cf. [1] for details of the first stage of interview process). In the second stage, we further interviewed nine more researchers (eight *followers* and one *novice*) identified in the search of SLRs ②. In total, 26 SE researchers were interviewed and all the interviews were transcribed for analysis.

2) Tertiary Study: After the first stage of interviews, we carried out a tertiary study that sought the secondary studies (SLRs and TLRs) between 2004 and the middle of 2009. This tertiary study was later updated ⑧ till the end of 2010 after the data analysis of the interviews and surveys ⑦.

Search for SLRs: We employed a Quasi-Gold Standard (QGS) based systematic literature search approach [25], integrating manual and automated search, for identifying the SLRs in SE. For manual search, we screened the publication venues related to empirical software engineering, EBSE, and premier SE venues, such as EMSE, ESEM, EASE, TSE, IST, JSS, ICSE and IEEE Software. It was followed by an automated search through five of the major digital library portals in SE [24]: IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, and Wiley InterScience.

Note that some secondary studies selected in [12] and [15] were not considered as SLRs in our tertiary study, as we followed a more strict criteria [25] to appraise a study as an SLR. For example, [23] was excluded from our study as its random sampling of ICSE papers may be not repeatable by other researchers.

Search for TLRs: Unlike SLRs, the publication of TLRs is much scattered among venues. Hence we applied only automated search to seek TLRs in SE published during the time span through the abovementioned five digital portals.

We identified a TLR by its usual structure which begins with presenting the earlier work in a particular area and proceeds with the most important past work up to the present [19], and the applied approach that does not use formalized methods of systematic review [18]. In terms of our observation, TLR is often combined with other research methods, such as survey, case study, or framework development. We excluded papers with TLR section if the literature review was not the major contribution. All included TLRs were grouped into two clusters: *full* TLR, which reports a literature review only; and *partial* TLR, in which literature review is one of the two major contributions reported.

3) Questionnaire Based Surveys: Based on the preliminary findings from ①, we carried out web-based survey using two questionnaires with comparable questions for the *systematic reviewers* and *tradition reviewers* respectively. The questionnaires[1] were developed by following Kitchenham and Pfleeger's guidelines for survey research in SE [16]. We extracted the potential participants of the survey study from the review studies (through the literature search ②) as those authored by researchers who could make up a sample set very close to our target population. After an internal trial within our research colleagues and external check by SE empiricists, the questionnaires were published online and remained open for gathering responses for a few months.

Survey for Systematic Reviewers: The SLRs identified through our literature search ② were reported by 124 authors. In order to minimize the sampling error and selection bias of the survey study, we maximized the sample size as close as possible to the target population (systematic reviewers in SE) by extracting their contact information from the publications. We found 4 authors without email address in their papers. Though both authors of this paper have also published some SLRs, we were excluded from the survey to prevent potential response bias. As a result, 118 invitations to participate in the survey were distributed to the identified *systematic reviewers*. We received 12 messages of non-delivery due to unknown email addresses, and 3 auto-replies that mentioned the invitees were unavailable during the survey period.

We received 52 responses to the survey with a response rate of 50% (52/103). Only one author explicitly rejected our invitation.

Survey for Traditional Reviewers: During the search of TLRs ②, we even found more authors of (*full* plus *partial*) TLRs than of SLRs. However, after excluding the authors who had also reported SLRs, we were left with 109 *traditional reviewers*. We extracted their contact information from the publications, and sent out 98 invitations to this survey to the authors of TLRs (11 authors without email addresses in their papers). We received 14 non-delivery messages and 2 auto-replies due to their unavailability.

We received 27 responses to this survey and the response rate was 33% (27/82). There was again one author who explicitly rejected our invitation.

As mentioned above, the survey questionnaires were designed to collect both quantitative and qualitative data. The gathered quantitative data were entered into MS Excel and MiniTAB for statistical analysis; the qualitative data from the surveys were analyzed using NVivo[2].

4) Impact Analysis: The SLRs and TLRs in SE were systematically searched ② and updated ⑧ by the end of 2010. We excluded those published in 2010 from impact analysis due to the short observation period for citation. In order to avoid the citation for the contributions other than literature review in the *partial* TLRs, we also excluded them from impact analysis.

The citations received by all SLRs and *full* TLRs published by the end of 2009 were collected through Google Scholar in February 2011, which provided at least 13 months

[1]The surveys were published online at http://systematicreviews.org.

[2]NVivo 7, QSR International (http://www.qsrinternational.com/)

for observing the impact of the reviews published in 2009.

IV. Value and Motivation

This section answers the RQ1 by synthesizing the advantages and strengths of SLRs valued by the methodology users; presents the systematic reviewers' motivators for performing SLRs and the reasons for "*why didn't traditional reviewers do SLRs?*" They are from the interviews ① ④ and the web surveys ③ in this research.

A. Value of SLRs

The SLR guidelines [13] state several advantages (strengths) of SLRs: *1*) well-defined methodology with less bias, *2*) effects across a wide range of settings and methods, and *3*) possibility of performing meta-analysis. One obvious disadvantage (weakness) of SLR is that it requires considerably more effort than TLRs. Here we compare the reflections from the categories of literature reviewers in SE with respect to the above mentioned claims.

We have analyzed the qualitative data gathered from the respondents to open-ended questions, which were used to mainly capture the explanations for responses provided using the Likert scale and different aspects of SLRs in SE. Our analysis of the responses to the question about the value of SLRs in SE revealed that a large majority of the respondents (both *systematic reviewers* and *traditional reviewers*) identified several values of SLRs; an overwhelming number of respondents were convinced that SLRs provide a systematic way of building a body of knowledge about a particular topic or research question. Other valuable aspects of SLRs reported by the survey respondents were '*more reliable findings based on synthesis of literature*', '*repeatability*', '*identification of problem areas for new research*', and '*a source for supporting practitioners' decisions about technology selection*'. Our analysis also showed that most of the claimed advantages of SLRs [13] were confirmed by the respondents of our survey.

When asked about the potential strengths of SLRs compared to TLRs, most of the respondents reported similar concerns as mentioned in response to the question about the value of SLRs such as a transparent and systematic approach, comprehensive and traceable review of the available literature, identification of more relevant sources of literature, and a basis of drawing reliable and unbiased conclusions. We found that the responses from the traditional literature reviewers also identified similar strengths of SLRs such as '*well-defined process*', '*objective selection of the papers*', and '*traceable and reliable findings*'. These identify some of the value and strengths of SLRs perceived by the participants of our surveys.

B. Motivators for Doing SLRs

Apart from the abovementioned value and advantages of SLR in SE, the systematic reviewers also share their more specific motivators. Table I enumerates the encouragements and fulfillment that researchers reported for doing SLRs [1] during the interviews ① and ④.

TABLE I
Motivators for doing SLRs

Motivator	Adv.	Fol.	Nov.
New research findings & ideas from SLR	0	5	2
Clear statement & structure of state-of-the-art	1	3	1
Learning from studies & getting knowledge	1	2	1
Recognition from community	0	3	0
Paper publication (e.g., motivated by IST Journal)	0	1	3
Working experience	0	1	0
Learning research skills (SLR methodology)	0	2	0

The most important motivators for conducting an SLR are '*getting new research findings and ideas from the results of SLR*', '*clear statement and structure of state-of-the-art*', and '*learning from studies and getting knowledge*', both of which are related to reviewer's research interests. The top motivator is further confirmed by 80% of systematic reviewers in ③ that SLRs sometimes bring them new research innovations *unexpectedly*.

Over half of systematic reviewers (53%) believe that SLRs in SE can be as effective as in other disciplines. About 22% of the respondents were more pessimistic about this questions. And the remaining respondents were '*not sure*'.

C. Reasons for Not Doing SLRs

Table II shows the responses to the surveys ③ when the traditional reviewers were asked "*why didn't you apply SLR methodology to your previous literature review?*". About 50% of the respondents '*did not know SLR methodology when they did the past (ad hoc) literature reviews*'. One quarter of them believed that a narrative review was more suitable for their previous study. Among the 5 respondents chose '*other*', three of them thought '*the study area is relatively small*' and/or '*insufficient data were available*', for which SLR might not be suitable.

TABLE II
Reasons for not doing SLRs

Reason	Respondent
Didn't know SLRs at the time of past literature review	14(52%)
Systematic reviews are time-consuming	10(37%)
Narrative review is more appropriate to my study	7(26%)
Other	5(19%)

Further, when the traditional reviewers were asked "*if you could go back and redo your literature review (survey), would you carry out an SLR instead of traditional (narrative) literature review?*", 74% of the respondents (traditional reviewers) gave positive answer, and showed their intent to carry out an SLR in their future literature reviews.

The traditional reviewers also showed their confidence in SLR's effectiveness in SE. About 50% of the respondents replied '*yes*', even they had not tried SLR methodology themselves. Only 11% of the respondents replied '*no*' to this question.

V. Topic and Impact

This section addresses the RQ2 by summarizing the SLRs reported by topic and year. This question also investigates the methodology diffusion and impact in SE. The evidence has been mainly extracted from the tertiary study (②, ⑤, ⑧ & ⑨), and also generated through the influence analysis ⑩.

A. Topics of SLRs

Table III presents an overview of the reported SLRs on a variety of topics in SE per year. Our search (② & ⑧) found 142 SLRs[3] (139 secondary studies and 3 tertiary studies) reported in 154 publications over the past 7 years. We identified over 30 SE research topics addressed by SLRs. Among them, *global development*, *cost estimation*, *requirements engineering*, *empirical research*, and *agile development* are the most investigated topics attracting systematic reviewers' research interests.

Figure 2 shows the number of *new* SLRs in SE per year has been continuously increasing over the past 7 years. There were noticeable jumps of the number in 2007 and 2009.

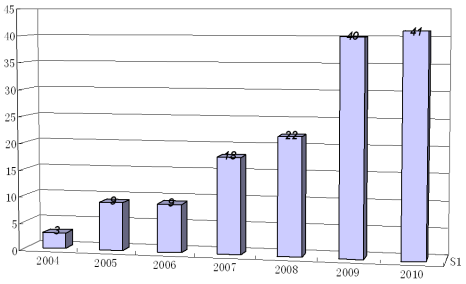

Figure 2. Numbers of *new* SLRs per year

In addition, systematic reviewers and traditional reviewers indicated 38 topics in SE (through the surveys ③) on which they would like to have further SLRs conducted. The most expected topics include *requirements engineering*, *software process improvement*, and *agile development*.

B. Impact of SLRs

To investigate SLR's diffusion and impact in SE, we collected the distribution of the *systematic reviewers* and citations of the SLRs *vs.* TLRs in the impact analysis ⑩.

1) Diffusion among Researchers: Based on the systematic reviewers extracted from the literature search ② & ⑧, we examined the diffusion of SLR methodology among the populations.

Figure 3 shows the geographic distribution of systematic reviewers from 2004 to 2009. As a newly introduced research methodology, it has convinced a large number of SE researchers in Europe. Nevertheless, the numbers of SLR users in the other regions were still quite low.

[3]The full list of SLRs in SE is available at http://mendeley.com/groups/965201/systematic-reviews-in-software-engineering/papers/.

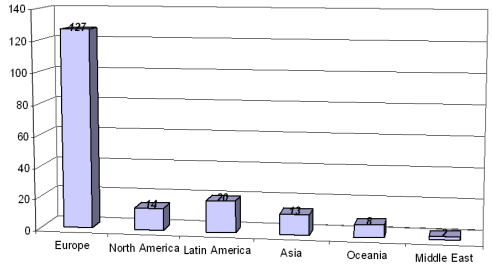

Figure 3. Geographic distribution of *systematic reviewers*

2) Impact on Researches: The citation information about both SLRs and TLRs was collected through Google Scholar in February 2011, and then summarized here by year and category. Figure 4 explicitly compares the average citations of the SLRs and the *full* TLRs year by year. It is obvious that the average citations of SLRs are higher than *full* TLRs in most years. This phenomenon to some extent reveals the higher impact of SLRs on SE research compared to TLRs.

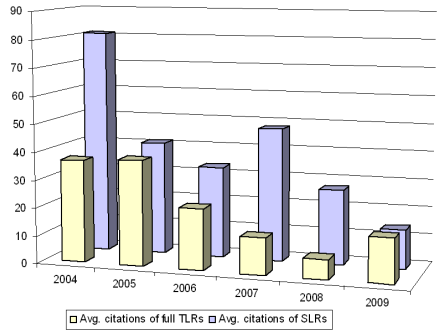

Figure 4. Citations of SLR & full TLRs over years

Figure 5 presents the statistical distribution of the citations received by the SLRs from 2004 to 2009 (collected from ⑦). It is noted that the standard deviations of the citation for SLRs varies significantly, which may imply the quality of SLRs was not so stable or some topics (or research questions) were not so attractive to other researchers.

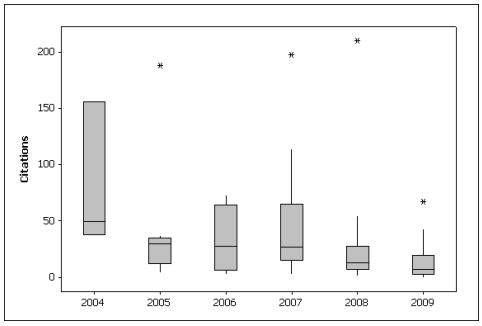

Figure 5. Citations of systematic reviews per year

TABLE III
OVERVIEW OF SLR STUDIES AND PUBLICATIONS IN SOFTWARE ENGINEERING RESEARCH BY TOPIC AND YEAR

Rank	Review topics	2004	2005	2006	2007	2008	2009	2010	Sum
1	Global development					◇◇◇	◇◇◇ ◇	♦* ◇◇◇ ◇◇◇ ▽	14/15
2	Cost estimation	♦	♦ ◇◇▽	♦ ◇	♦♦♦* ◇ ◇		▽		12/13
3	Requirement engineering			◇	◇	◇	♦ ♦ ◇◇◇* ▽ ▽	♦ ♦ ◇	12/13
4	Empirical research		♦ ◇ ▽	♦	♦ ♦ ◇	♦ ▽	♦ ▽		11/11
5	Agile development					♦	♦ ◇	♦ ◇◇◇ ◇◇▽	10/10
6	Inspection & testing	♦		♦	▽	♦* ◇	♦ ♦ ♦* ◇	♦ ♦* ◇	9/12
7	Software architecture			▽		▽	♦ ▽	♦ ♦ ◇ ▽	8/8
8	Software process improvement				♦ ◇	♦ ♦ ◇	◇	♦	7/7
9	Software process modeling			▽		◇	◇	♦ ♦ ◇ ◇*	6/7
10	Software product lines						♦ ◇ ◇ ▽	♦*	4/5
10	Open source development						▽ ◇*	♦ ◇ ◇	4/5
12	Software measurement			◇		♦ ◇		♦	4/4
12	Software tools				♦		♦	♦ ◇	4/4
14	Tertiary study						♦	♦* ◇ ▽	3/4
15	Software security					♦ ◇		♦	3/3
15	Software maintenance				♦	◇	◇		3/3
15	Software design				◇	◇	♦		3/3
15	Web engineering		◇		◇	▽			3/3
19	Motivation in software engineering					♦ ♦*	♦		2/3
20	Service-oriented & grid computing						♦ ♦		2/2
20	Unified modeling language						♦ ♦		2/2
20	Software evolution				♦			◇	2/2
20	Aspect-oriented programming						◇	♦	2/2
20	Program analysis						♦ ◇		2/2
20	Business process			◇				♦	2/2
26	Knowledge management					♦	♦*		1/2
	Other topics	♦	◇		♦ ▽		◇	♦ ♦	7/7
Total	*–new/all SLR reports–*	3/3	9/9	9/9	18/19	22/24	40/44	41/46	**142/154**

▽ workshop paper or short paper ◇ conference paper ♦ journal paper (* paper updating or extending previous review report)

From our analysis ⑩, the most 'influential' (cited) SLRs per year are Jorgensen's [10] in 2004, Sjoberg et al.'s [21] in 2005, Davis et al.'s [6] in 2006, Jorgensen and Shepperd's [11] in 2007, Dyba and Dingsoyr's [7] in 2008, and Kitchenham et al.'s [12] in 2009.

VI. PRACTICE AND EXPERIENCE

Methodological rigor is one major claimed strength of EBSE [14]. It is also known that SLR requires considerable amount of effort and expertise compared to TLR [17]. Accordingly, rigor and effort are two distinct characteristics of SLR in practice. These characteristics always need to be balanced during the course of performing an SLR. In addition, unlike TLR, teamwork is necessary for minimizing the potential bias during SLR. However, there is no guidance on forming a review team that includes people with the required diversity [17] in the SLR guidelines for SE [13].

This section concentrates the above three distinct characteristics of SLR to investigate the systematic reviewers' practice in SE. We distilled the results and findings focusing on these three aspects, which are mainly collected from activity ⑤, ⑦ & ⑨, in order to benefit the current and future systematic reviewers in SE.

A. Teamwork

To enhance its rigor and minimize the potential bias, an SLR is normally undertaken by more than one reviewers. Hence, setting up a review team is always a necessary step before starting an SLR.

Team Size: From the surveys ③, more than half of our respondents (51%) believe that an ideal SLR team should consist of three members. Equally 18% of the respondents regarded either two or four members can form a desirable team. The reason can be explained by the findings from the interviews ① & ④ that a '*too small*' team size (e.g., single reviewer) is difficult to control the potential bias; a '*too large*' size, on the other hand, may lead to a much higher communication and coordination overhead.

Team Distribution: In the interviews ① & ④, our interviewees addressed their procedure of performing SLR in either local (within organization) or distributed (cross organizations) settings. A local team may make communications '*simple*' and '*convenient*', while a distributed team could ensure the '*expertise*' and reinforce '*independence*' of the individual's reviews.

We further examined the systematic reviewers' preference in a larger pool (the surveys ③). The results ⑦ show that many people (36%) are inclined to a local team. However, a large number of our respondents (47%) would like to make decision between the two settings based on the tradeoff (of the abovementioned benefits) in the real context.

B. Rigor

It has been reported many times that rigor is one of the top strengths claimed by SLR that ensures the review quality. We were interested in revealing how the rigor is implemented in SLRs.

Activity: Many systematic reviewers (35%) believe that *protocol design* is the most important activity to ensure the

rigor of an SLR. It is followed by *study selection*, *quality assessment*, and *data extraction*, which received 18%, 14% and 14% support respectively. Only one respondent thought *data synthesis* is the most critical for SLR quality.

Bias Control: In order to minimize the potential bias during the review process, peer-review is the most common method used by 80% systematic reviewers. Many reviewers also sought help from external checkers (33%), conducted self-review (29%), or validated the agreements by statistical techniques (24%), e.g., Kappa.

C. Effort

In estimating the effort (time) for performing an SLR, Allen and Olkin [2] present a regression formula for determining the number of hours as a function of the number of references returned (x),

$$hours = 721 + 0.243x - 0.0000123x^2 \quad (1)$$

As this formula was based on empirical observations on SLRs in medical discipline, its accuracy in SE needs to be further assessed. However, there is no guidance for estimating the required effort distribution across the three phases of an SLR.

The box-plot in Figure 6 shows the distribution of the respondents' (*systematic reviewers*) estimation of effort consumed over the three phases of SLR according to their experiences. It indicates that the *conducting* review phase takes around half of the effort of undertaking an SLR. In the other half, the *planning* and *reporting* review phases roughly share an equal quarter of the overall effort. The most time-consuming activity (through ③) was *data extraction* (43% responses) followed by *study selection* (27% responses).

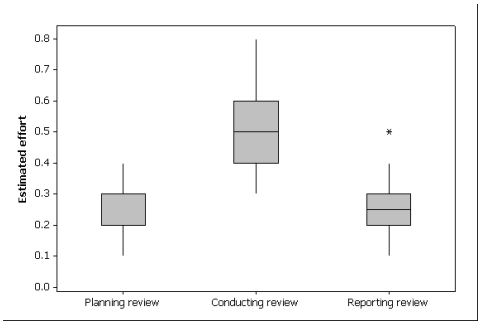

Figure 6. Effort consumed over SLR phases

Through ①, many interviewees mentioned the iterative rework between study selection and data extraction, especially for novices. This issue was further investigated with the survey ③. Only 18% systematic reviewers in SE never experienced rework between these two activities.

VII. Reflection and Discussion

We have adopted a multi-method approach to gain an in-depth understanding of different aspects of the status of adoption, perceived and real advantages, and practice of SLRs in SE since the introduction in 2004. We expected the use of multi-method approach to provide us with complementary research strategies to produce robust findings by systematically gathering and analyzing the required data [22]. We assert that a multi-perspective overview of adoption and practice of SLRs in SE over the past seven years can help SE researchers and practitioners to understand the perceived value, current or potential impact, and practices of a research methodology that has been reported to have significant influence on practitioners' and policy makers' decisions in several other disciplines.

A. Motivations

The proponents of SLRs in SE claim that SLR is a systematic approach to building a body of knowledge about a particular topic or research question(s) by its practitioners, and to identifying problems for future research and support decision making and technology selection. Our findings reveal that many *traditional reviewers* also agree with various value of SLRs claimed by the *advocates* and confirmed by *followers* of SLRs in SE, and are also willing to try SLR in their future literature reviews.

Most of the *systematic reviewers* believe that SLRs provide the mechanism of achieving reliable and traceable findings that can help practitioners to make more informed decisions. Our research has also revealed that SLR users expect to identify new issues in the studied area and innovative ideas. This can be considered the most important motivator for performing SLRs. The participants of our research realize that there are significant differences between medical and software engineering disciplines that is why it is difficult to expect of EBSE to have similar results as evidence-based medicine has produced in the near future. However, it is a good surprise that a majority of the participants were quite optimistic about the effectiveness of SLRs in SE as in other disciplines, such as sociology and education. We also found the reasons for researchers choosing TLRs over SLRs. It is an interesting finding that half of the respondents were not aware of SLRs when they carried out their literature review. However, compared to *systematic reviewers*, we received a relatively small number of responses to our survey invitation from the *traditional reviewers*. This can be considered an indication that *systematic reviewers* are very enthusiastic about this particular research methodology and are willing to share their perspectives and experiences in order to help improve the methodology.

B. Practices

Over the past seven years, a large number of SLRs (142) have been reported on a variety of research topics (30+) in

SE. However, their distribution among the topics was not yet balanced. We found SLRs had been intensively conducted and reported on a few topics, such as *cost estimation* and *requirements engineering*, particularly *global development* in the recent years. Our research shows SLRs, compared to the TLRs reported in the same period, have made a wider research influence in terms of citations. Whereas, systematic review, as a new research methodology, has not been widely adopted by the researchers outside Europe.

An SLR requires more rigor and effort. The rigor is based on its teamwork, bias control and systematically defined process. With respect to their experiences, the users of SLR suggested an ideal team size of three, and *designing protocol* as the most important activity to ensure rigor and repeatability of SLR. Based on the statistics, *conducting review* is the most effort-consuming phase in SLR, especially the *data extraction* activity. The effort distribution reflected by the experienced *systematic reviewers* may help the future reviewers estimate and plan their own SLRs.

C. Limitations

This research has some limitations that we consider worth mentioning. We have used a multi-method approach which is expected to help produce more robust results which are based on the combination of complementary empirical research methods. However, we need to be aware of the limitations of the individual research methods, e.g., interview and questionnaire-based survey. Our study has explored the perceptions and views of SE researchers about their experiences of applying SLRs in SE research through semi-structured interviews. That means our results are based on the recollection of the interviewees. We tried to minimize this risk by audio-taping all the interviews with the interviewees' permission. The transcriptions of the interviews were verified with the notes taken. Moreover, we tried to have both authors present in most of the interviews.

A shortcoming of the questionnaire-based survey sessions is that respondents are provided with a list of potential reasons and expected benefits of a particular technology (i.e. SLR methodology in our case) and are asked to select from that list. This approach may limit the respondents to consider only those options and questions provided to them. However, we tried to address this issue and provided the respondents an opportunity to share their perceptions and experiences by seeking explanation for most of the responses with open-ended space. We gathered significant amount of qualitative data. A relatively small number of responses to our survey invitation from traditional reviewers can be another limitation of this study.

Considering the wide distribution of traditional literature reviews in SE publication venues, we only performed an automated search to capturing TLRs. This might result in an incomplete set of the identified TLRs compared to the outputs of our '*exhaustive*' search of SLRs based on the QGS approach [25]. However, the purpose of searching TLRs in this research was for impact analysis (comparison with SLRs) in terms of *average* citations, rather than to produce a complete list of TLRs. Hence, we have confidence on the validity of our findings about the impact that was based on a large sample set of TLRs retrieved from automated search.

Generalizability can be another risk. However, we tried to manage this by involving the participants from different organizations and located in different parts of the world. It should also be noted that a large majority of the participants (i.e. interviewees and survey respondents) reported similar experiences and lessons. It increases our confidence in the findings of this study. One of the main limitations of the multi-method approach is the huge amount of time and effort required for planning, executing, and analyzing each phase of our research. However, we believe that such investment is necessary in order to achieve reliable and robust results.

VIII. Conclusions and Future Work

Our long-term research goal is to build an empirically supported body of knowledge to improve the adoption and use of SLRs in SE with the objective of supporting practitioners' decision making for selecting software technologies. Furthermore, we also intend to contribute to the development of scientific and technological support to exploit the full potential of SLRs. We are approaching these goals by firstly concentrating on gaining an in-depth understanding of researchers' perspectives about and motivations for conducting (or not conducting) SLRs, capturing the current status of the adoption and impact of SLRs in SE research, and also studying the practices being followed for conducting SLRs.

This research has gathered empirical evidence to advance the knowledge about different methodological aspects and logistics involved in *planning*, *conducting*, and *reporting* SLRs in SE. The findings provide support to several advantages and strengths of SLRs claimed by the advocates of this methodology based on the perceptions of the users of systematic review as well as of the users of traditional (ad hoc) review. The results also provide SE researchers and practitioners an evidence-based understanding of different aspects and potential value of SLRs for decision making.

Systematic review is a relatively new research methodology for SE researchers. The appropriateness and application of SLR in SE have yet to be fully explored and assessed. However, there is clear evidence that more and more researchers support the move to perform secondary studies in a systematic and rigorous manner as even *traditional reviewers* are of the view that systematic reviews are valuable to SE research and practice. Researchers can also gain motivation for continuously improving the methodology and reporting rigor of their studies in order to deliver high quality SLRs. This study has also identified a few best practices that are expected to be useful for researchers intending to undertake systematic reviews. Additionally, the findings

that published SLRs appear to be more influential than TLRs should provide satisfaction to those who have invested significant amount of time and effort in conducting SLRs and motivate those who have been contemplating to apply this methodology. We also expect that our initiative to study different aspects of research methodology will stimulate researchers to carry out similar studies for increasing the understanding of the value and the adoption of SLRs in SE.

ACKNOWLEDGMENTS

We would like to record our thanks to all those who have participated in this research and shared their perspectives and experiences.

REFERENCES

[1] M. Ali Babar and H. Zhang. Systematic literature reviews in software engineering: Preliminary results from interviews with researchers. In *3rd International Symposium on Empirical Software Engineering and Measurement (ESEM'09)*, pages 346–355, Lake Buena Vista, FL, Oct. 2009. IEEE.

[2] I. E. Allen and I. Olkin. Estimating time to conduct a meta-analysis from number of citations received. *Journal of the American Medical Association*, 282(7):634–635, 1999.

[3] J. Biolchini, P. G. Mian, A. C. C. Natali, and G. H. Travassos. Systematic review in software engineering. Technical report, Universidade Federal do Rio de Janeiro, 2005.

[4] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil. Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, 80(1):571–583, 2007.

[5] D. Cruzes and T. Dybå. Synthesizing evidence in software engineering research. In *4th International Symposium on Empirical Software Engineering and Measurement (ESEM'10)*, pages 1–10, Bolzano/Bozen, Italy, Sept. 2010. ACM.

[6] A. Davis, O. Dieste, A. Hickey, N. Juristo, and A. M. Moreno. Effectiveness of requirements elicitation techniques: Empirical results derived from a systematic review. In *14th IEEE International Requirements Engineering Conference (RE'06)*, pages 179–188, Minneapolis/St. Paul, MN, Sept. 2006. IEEE.

[7] T. Dyba and T. Dingsoyr. Empirical studies of agile software development: A systematic review. *Information and Software Technology*, 50(9-10):833–859, 2008.

[8] T. Dyba, T. Dingsoyr, and G. K. Hanssen. Applying systematic reviews to diverse study types: An experience report. In *1st International Symposium on Empirical Software Engineering and Measurement (ESEM'07)*, pages 225–234, Madrid, Spain, Sept. 2007. IEEE.

[9] T. Dyba, B. Kitchenham, and M. Jorgensen. Evidence-based software engineering for practitioners. *IEEE Software*, 22(1):158–165, 2005.

[10] M. Jorgensen. A review of studies on expert estimation of software development effort. *Journal of Systems and Software*, 70(1-2):37–60, 2004.

[11] M. Jorgensen and M. Shepperd. A systematic review of software development cost estimation studies. *IEEE Transactions on Software Engineering*, 33(1):33–53, 2007.

[12] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. Systematic literature reviews in software engineering: A systematic literature review. *Information and Software Technology*, 51(1):7–15, 2009.

[13] B. Kitchenham and S. Charters. Guidelines for performing systematic literature reviews in software engineering (version 2.3). Technical report, Keele University and University of Durham, 2007.

[14] B. Kitchenham, T. Dyba, and M. Jorgensen. Evidence-based software engineering. In *26th International Conference on Software Engineering (ICSE'04)*, pages 273–284, Edinburgh, Scotland, UK, May 2004. IEEE.

[15] B. Kitchenham, R. Pretorius, D. Budgen, O. P. Brereton, M. Turner, M. Niazi, and S. Linkman. Systematic literature reviews in software engineering – a tertiary study. *Information and Software Technology*, 52(8):792–805, Aug. 2010.

[16] B. A. Kitchenham and S. L. Pfleeger. Principles of survey research: Part 2 6. *ACM SIGSOFT Software Engineering Notes*, 27-28, 2002-2003.

[17] J. H. Littell, J. Corcoran, and V. Pillai. *Systematic Reviews and Meta-Analysis*. Pocket Guides to Social Work Research Methods. Oxford University Press, 2008.

[18] M. Petticrew and H. Roberts. *Systematic Reviews in the Social Sciences: A Practical Guide*. Wiley Blackwell, 2006.

[19] G. Rugg and M. Petre. *The Unwritten Rules of PhD Research*. Open University Press, 2004.

[20] D. I. Sjoberg, T. Dyba, and M. Jorgensen. The future of empirical methods in software engineering research. In *International Conference on Software Engineering, Future of Software Engineering Track*. IEEE, 2007.

[21] D. I. Sjoberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanovic, N.-K. Liborg, and A. C. Rekdal. A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering*, 31(9):733–753, 2005.

[22] M. Wood, J. Daly, J. Miller, and M. Roper. Multi-method research: An empirical investigation of object-oriented technology. *Journal of Systems and Software*, 48(1):13–26, 1999.

[23] C. Zannier, G. Melnik, and F. Maurer. On the success of empirical studies in the international conference on software engineering. In *28th International Conference on Software Engineering (ICSE'06)*, pages 341–350, Shanghai, China, May 2006. ACM.

[24] H. Zhang and M. Ali Babar. On searching relevant studies in software engineering. In *14th International Conference on Evaluation and Assessment in Software Engineering (EASE'10)*, Keele, England, Apr. 2010. BCS.

[25] H. Zhang, M. Ali Babar, and P. Tell. Identifying relevant studies in software engineering. *Information and Software Technology*, 53(6):625–637, 2011.