Experiences conducting systematic reviews from novices' perspective

Article · April 2010

CITATIONS
35

READS
146

4 authors:

Mehwish Riaz
University of Auckland
14 PUBLICATIONS **215** CITATIONS
SEE PROFILE

Muhammad Sulayman
University of Auckland
16 PUBLICATIONS **182** CITATIONS
SEE PROFILE

Norsaremah Salleh
International Islamic University Malaysia
57 PUBLICATIONS **617** CITATIONS
SEE PROFILE

Emilia Mendes
Blekinge Institute of Technology
223 PUBLICATIONS **4,458** CITATIONS
SEE PROFILE

Experiences Conducting Systematic Reviews from Novices' Perspective

Mehwish Riaz[1], Muhammad Sulayman[2], Norsaremah Salleh[3], Emilia Mendes[4]
Department of Computer Science, The University of Auckland, Level 5, 38 Princes Street, Auckland 1142, New Zealand
{[1]mria007, [2]msul028, [3]nsal017}@aucklanduni.ac.nz, [4]emilia@cs.auckland.ac.nz

Background: A systematic review (SR) is a sound methodology for collecting evidence on a research topic of interest and establishing the context of future research. Unlike ordinary or even expert literature reviews, SRs are systematic thus increasing the confidence in the findings from the previous published literature. SRs can be carried out by both experienced and novice researchers; however, while expert researchers' experiences with conducting SRs are important for improving the SR body of knowledge, we believe that novice researchers' experiences are equally important to establish what distinct problems they face while carrying out SRs. With a prior knowledge of these issues, novice researchers can better plan their SRs and seek guidance from expert researchers.
Aim: The aim of this paper is therefore to report on experiences conducting SRs from the perspective of novice researchers. The paper reports first hand experiences of novices conducting SRs and compares them with the experiences of an expert as well as with the experiences reported in the previous literature.
Method: An instrument was created and used to gather the experiences conducting SRs from three PhD students and their supervisor. The instrument covered all the SR steps; it was individually filled out by each of the participating subjects and its data was later on aggregated.
Results: The results show that the problems faced by novices in terms of time taken to conduct the review; defining the research questions, inclusion/exclusion criteria, data extraction and data synthesis forms are not faced by expert researchers. Moreover, problems faced by novices related to defining quality criteria are different in nature than those faced by expert researchers.
Conclusions: It has been observed that while numerous problems are faced by both novices and experts, many others are specific to novices, where several of these can be solved with the help of domain and SR experts.

Keywords: Systematic review, empirical software engineering, experience report.

1. INTRODUCTION

Systematic Reviews (SRs), also known as Systematic Literature Reviews, have been recently introduced in the area of Software Engineering (SE) [9] [20] as a structured and systematic methodology for performing literature reviews. Since then, SRs have gained significant importance in SE as a means of identifying, evaluating and interpreting all available evidence relevant to a particular research question, topic area, and phenomenon of interest [7]. While to date numerous SRs have been published on diverse SE topics such as cost estimation [8], Web engineering [12], organizational motivators for adopting CMMI-based software process improvement [18], agile software development [6], software maintainability prediction and metrics [15], and software process improvement for small and medium Web companies [19], there have also been studies reporting their lessons learnt and experiences conducting SRs (e.g., [3], [5], [17]). Almost all of these SRs have followed Kitchenham's guidelines [7] [20] for performing SRs in SE.

As the use of SRs in SE is new, various researchers have reported problems relating to different SR activities such as formulation of research questions [17], [3], conducting of literature searches [5] [17], inter-rater reliability checks [17], selection bias [6], quality assessment of primary studies [6] [5] [17]; and also with the overall SR process as being very time consuming [1] [13]. We believe that reporting such problems as well as the experiences while dealing with these problems are not only important for other researchers who conduct SRs, but are also a supplement to the guidelines for performing SRs in SE.

While most of the studies reporting SRs or experiences conducting SRs have been carried out from the perspective of experienced researchers [1], we believe that the problems faced by new researchers such as PhD students may be considerably different in nature as well as in magnitude from that of experienced researchers. This paper, therefore, reports first hand experiences of three PhD students (novices, the term adopted from [1]) while carrying out different steps of SRs. The main aims of this research are therefore to:

- Collect and report experiences conducting SRs in SE from novices' perspective.
- Report the issues the novices faced while conducting SRs and the actions that they took to overcome those issues.
- Compare the experiences of an expert – their supervisor – with the experiences of novices and identify common issues.
- Compare the experiences reported by experts – both in previous literature and by the novices' supervisor - and compare with the experiences of novices to identify common issues as well as differentiate between the issues faced by experts and novices.

The main contribution of this paper is therefore, a first hand experience report of the novices conducting SRs and how these compare with experts' experiences. We believe that the findings reported herein can be beneficial for future PhD students, their supervisors and other researchers conducting SRs in SE.

The remainder of this paper is organized as follows: Section 2 briefly presents an overview of SR activities, followed by an account of related work in Section 3; Section 4 reports the experiences of the three PhD students conducting SRs in SE as well as presents their background and the instrument used for recording experiences; Section 5 compares the experiences of the expert and the three novices; a discussion based on the comparison of the collective experiences of the novices and expert with the previously published experiences is given in Section 6 along with some suggestions for novices. Finally conclusions are presented in Section 7.

2. SYSTEMATIC REVIEWS

An SR is a sound methodology to understand current state of the art and to identify future opportunities for research in a given topic area [7]. SRs are very helpful in [7] [20]:

- Identifying the existing evidence regarding a treatment or technology.
- Providing a context for properly placing new research activities.
- Examining the extent to which empirical evidence supports or contradicts theoretical hypotheses.
- Identifying gaps in the existing research.
- Assisting the generation of new hypotheses.

The three main phases of a SR process are to plan, conduct and report the review [7] [20], which comprise the following general steps [7]:

- Identification of the need for conducting a SR.
- Formulation of purposeful research question(s).
- Development of a research protocol.
- Exhaustive search of primary studies.
- Selection of primary studies based on the inclusion criteria.
- Quality assessment of selected studies.
- Extraction of data from selected studies.
- Synthesis of extracted data against the research questions.
- Interpretation of results.
- Report writing.

It is our view that an SR is a very important methodology for use by novice researchers because it enables them to have confidence on the results of their literature review, provides them with a sound ground to base their future research on and teaches them various key aspects of conducting research. Therefore, the conducting of SRs by novices always proves beneficial even if the outcomes of an SR in terms of answering the research questions are not as expected (e.g. small number of primary studies).

3. RELATED WORK

A summary of the problems with conducting SRs reported in previous studies, and whenever applicable their proposed solutions or suggestions, is given in Table 1. In addition, some studies also suggest best practices [1] [3] for conducting SRs, challenges faced while conducting SRs [1], and point out a need for improvement in the SR guidelines related to the aspects of quality assessment, experience/examples, simplified version, quantitative analysis, qualitative analysis, and protocol templates [1]. Moreover, Kitchenham et al. [11] also suggest that broader automated searches can find more studies than restricted manual searches; however these are more effort and time consuming in comparison to manual searches.

Note that all the related work in this section has been conducted by expert researchers and is mostly based on their experiences conducting SRs. The two exceptions are [1] [11], which are based on interviews with SR practitioners and an SR's replication, respectively. In addition, Oates and Capper [14] discuss their experience teaching SRs to Masters' students and Baldassarre et al. [2] discuss their experience involving graduate students in the data extraction process. To the best of our knowledge only one study to date discusses the issue of conducting SRs from the perspective of novice researchers - Babar and Zhang [1], who synthesized experiences from various researchers including Advocates, Followers and Novices by conducting an interview-based survey. They report best practices, challenges, and suggestions for improving SRs. Although their work considered novices, the experiences reported are not first hand, and are generic in nature. Therefore, we believe that this paper is the first to explicitly detail experiences conducting SRs from novices' perspective.

Table 1: *Problems and solutions/suggestions conducting SRs reported in previous studies*

Problems Reported	**Solutions/Suggestions Provided**
<u>Overall Process</u>: *The overall process of conducting SRs is time and effort consuming [1], in particular processes of planning [13], execution [13] and data extraction [3].* *Difficulty in initial learning [4], lack of domain knowledge [1] [3] and guidance [1], education of the supervisors [1], not clear when to stop the piloting process [17], not finding many studies [5], distributed nature of team complicates the process of protocol development [3], limited quantity and quality of record keeping [3], and lack of consistency among researchers for defining conceptual structure during SR [4].*	*A review protocol template for optimizing the process of planning and execution [13] and for data extraction - one person extracting data and other acting as a checker [3] is more efficient than two people extracting data and comparing inconsistencies [3]. The problem of lack of consistency among researchers for defining a conceptual structure can be improved by having the support of a formalized common terminology of involved concepts represented by an ontology [4].*
<u>Defining Research Questions (RQs)</u>: *Problem scoping RQs [3], RQs structure not fitting comfortably with Population, Intervention, Comparison, Outcome, Context (PICOC) [17] and guidelines not considering impact of RQ type on review procedures [4].*	*For scoping RQs, perform a systematic preview mapping study [3] and define complementary RQs and unit of analysis [17].*
<u>Search Strategy</u>: *Non-standardized keywords [3] and difficulty in identifying synonyms [4].*	*Ask field/area specialists for help with synonyms [4].*
<u>Search Process</u>: *Databases (DBs) have limitations on Boolean expressions [4] [5] and number of characters [4], have different search syntaxes [3], give different results for basic or advance search [17], and have different order of evaluation of Boolean expressions [3]; in some DBs e.g. ACM searches cannot be restricted to abstracts and titles [5]; most publisher-specific DBs do not return any unique results[5]; and search strings are different for different DBs due to their different underlying models[3].*	*Use a broader search string [17] and use of a common tool that can aggregate results from DB searches [17].*
<u>Study Selection</u>: *Unstructured abstracts [3] on the basis of which judgment on inclusion/exclusion of studies cannot be made [5].*	*Promote the use of structured abstracts in SE [3] [5].*
<u>Quality Assessment</u>: *Not knowing how the threats to the validity were controlled [17] and if including a large number of primary studies is better than conducting high-quality review with more selective quality assessment criteria [5]. Unavailability of scoring methods for diverse study types [5]. Problems assessing quality of primary studies and that of SRs [6].*	*Use of a quality criterion only for assessing publication bias, internal validity, and external validity [3]. For diverse study types, a framework was developed to assess quality of combined qualitative and quantitative empirical research was developed [5]. For assessing quality of SRs Critical Appraisal Skills Programme (CASP), Grades of Recommendation Assessment, Development and Evaluation (GRADE), and Meta-Analysis of Observational Studies in Epidemiology (MOOSE) were used [6].*
<u>Selection bias</u>: *Bias in selection [6] and no inter-rater reliability checks [17].*	*Pilot every part of review process and use documentation and multistage process for selection [6].*
<u>Data Extraction and Synthesis</u>: *Not clear how much categorization is done during extraction and how much during data synthesis [17]; no standard method for synthesizing data from qualitative or mixed methods studies [5]; mechanism to undertake meta-analysis are not clear [4]; primary studies report multiple tests for one study, report subset of required data, or use different means of reporting results [3]; and data extractors not as familiar with the statistical terms as members who defined data extraction form [3].*	*Meta-ethnography was used for synthesizing qualitative, mixed methods studies [5]. For studies reporting multiple tests, refine data extraction process to ensure results from each study were not double counted [3]; for subset of data, define a process for managing missing values [3]; and for different means of reporting, amend data extraction and aggregation process [3].*
<u>Reporting the Review</u>: *Limited space while publishing [3] [5] resulting in assessment of reporting quality rather than research quality [5]. Publication problems due to lack of reviewer's knowledge of SRs [1]. Not clear if the final protocol should be shown or its evolution process [1].*	*Show final protocol with footnotes where changes took place to indicate its evolution [1].*

4. EXPERIENCES WITH CONDUCTING SRs FROM THE NOVICES' PERSPECTIVE

In this Section, the experiences of three of the four authors of this paper - Riaz, Sulayman and Salleh (SR novices), with conducting SRs during their PhD studies are presented, organized according to the SR steps (see Section 2). They are all supervised by the same person (author Mendes) - who is experienced in conducting SRs. The time they spent in both developing the protocol and executing the SR was less than a year; Riaz took approximately 8.5 months, whereas Sulayman and Salleh spent about 10 months each. For the remainder of this paper the SR novices' are identified as N1 (Riaz), N2 (Salleh) and N3 (Sulayman). Their SR reports and papers are listed below.

- N1 → Mehwish Riaz, Emilia Mendes and Ewan Tempero. A Systematic Review of Software Maintainability Prediction and Metrics [15].
- N2 →Norsaremah Salleh, Emilia Mendes and John Grundy. Empirical Studies of Pair Programming for CS/SE Teaching in Higher Education: A Systematic Literature Review [16].
- N3 →Muhammad Sulayman and Emilia Mendes. Systematic Review of Software Process Improvement in Small and Medium Web Companies [19].

The instrument used for recording their experiences is given in the Appendix. This instrument was created by the first author and was then validated until all the four authors were satisfied that it provided coverage to all the steps of the SR on which they wanted to report their experiences. Note that the same instrument was also used by the expert. The instrument was filled out individually by all four authors by recalling their experiences as well as by consulting the notes made during the review meetings, minutes of meetings and by searching through notes and footnotes in the various documents produced while conducting the SRs. This exercise was done in year 2009 i.e., within a year and a half of conducting the SRs by all three novices and after 4 years of conducting SRs by the expert. Their individual experiences were synthesized into tables by the first author, which were later reviewed in a meeting attended by all four authors. Given that a large part of the data used to report the authors' experiences were based on existing documentation from conducting SRs, we believe that the experiences reported herein are trustworthy.

4.1. Developing the SR protocol

Some similarities as well as differences were observed in the SR novices' experiences related to the development of the SR protocol (see Table 2). In particular, similar problems were faced while formulating the research questions (RQs) and determining the inclusion/exclusion (I/E) and quality criteria. Since the formulation of RQs was guided by the structure of PICOC, they found that identifying the PICOC really depended on the SR's context. For example, in all the three studies, the "comparison" component was not applicable because the SRs did not intend to perform any comparisons of the intervention. The formulation of RQs was also driven by the SR novices' domains of interest, and the challenge was to ensure that the RQs were relevant and appropriate. The RQs were revised iteratively until consensus was achieved among the review team - a novice (PhD student) and the supervisor (Mendes).

The main challenge in determining the I/E criteria was based on the difficulties in specifying the conditions of a study that would make it appropriate for inclusion in the SR. The ability to define the scope of the review helped in devising the inclusion criteria. The SR novices also found that discussing this issue with their supervisor was very helpful. Regarding the quality criteria, they faced similar problems in developing the quality checklist especially in finding suitable questions for assessing the studies' quality. This led one of the SR novices to use a separate checklist for quantitative and qualitative studies. In most situations, the challenges were overcome by consulting the supervisor and referring to the existing SR guidelines, books and published articles.

In table 2, only the problem faced by a particular novice is followed by the novice identifier, N1, N2 or N3. The common problems are given without this distinction.

4.2. Conducting the SR

In the initial stage, the main problem was related to the outcomes of executing the search string as defined in the protocol. The SR novices retrieved either few or no results when searching using the comprehensive search string, however obtained more results when they broadened the search string. In addition, they also observed that there were different types of facilities and search limitations posed by different databases, adding additional complexities to the literature searching process (e.g. SpringerLink, LNCS and ACM allow only 10 terms per search string, thus the search strings had to be adapted accordingly).

The difficulty in determining whether a study was relevant to be included in the SR was also one of the major challenges. Apart from the titles, both the abstracts and conclusions were also referred to during the initial study selection. The SR novices found that in many studies there was a lack of clarity in the abstract, thus they could not rely their selection solely upon abstracts. This led them to having to read the full texts of several papers, and in some cases to also contact the paper's authors (whenever the author(s) could still be traced).

The SR novices faced similar problems while extracting data from the studies as certain details were not clearly described or explicitly documented in the primary studies. To ensure the consistency of data extracted from the studies, each of the review teams had a joint meeting to compare the extracted data for a sample of the primary studies. With regards to applying the quality criteria, a few questions in the quality checklist could not be answered due to the lack of precise information reported in some of the primary studies. The SR novices' experiences suggest that overall each of them encountered almost similar problems while conducting the SRs (see Table 3). Note that as in Table 2, Table 3 also identifies a problem specifically faced by a novice by the novice identifier. The common problems faced are given without an identifier.

Table 2: *Summary of problems encountered while developing the SR protocol*

Challenges/problems faced	**Strategy to overcome the problems**
Formulation of research questions	
- The challenge was to ensure that the RQs were relevant, appropriate, made sense with regards to conducting SR, and covered all the aspects of the area to be investigated.	*-Read more about the subject area. Consult the supervisor. The RQs were revised many times until the SR covered all the aspects of the investigation.*
Identification of PICOC	
-Unable to find the 'Comparison' context in PICOC.	*-Remove the comparison part from PICOC: became PIOC.*
Devising Search Strategy	
-The related papers found did not meet the minimum inclusion criteria but had relevant keywords. Some papers that seemed relevant did not have a Keywords section (N1).	*- Keywords were included from related papers and some terms were picked up from papers that seemed relevant but had no Keywords section. It was documented in the protocol that the keywords were coming from studies that were not part of the SR (N1).*
-Too rigid to restrict the search strategy based on the synonyms or by applying Boolean operators. This created a very complex search string, which returned fewer results (N2).	*- Refer to the SR guidelines and consult a subject librarian (N2).*
Determining Inclusion/Exclusion Criteria	
-Terms used in the previous literature were often found to be confused with interrelated but different terms (N1).	*-The distinction was made between the interrelated terms and only the term relevant to the SR's context was considered (N1).*
-Should studies that discussed metrics but that did not validate them be included? The answer to this question also determined whether a quality check to include only empirical studies was enforced while study selection. (N1)	*-Devise the inclusion criteria to include only studies that performed empirical validation of new metrics (N1).*
-Problem in defining the scope of the review. Initially it was difficult to define the exclusion criteria.	*-Identify the focus of the study and refer to the PICOC. Also, discuss with the supervisor.*
Determining Search Phases	
-What years should be included in the search? (N1)	*-Run a preliminary search to find out in which year the literature relevant to the review topic was first published (N1).*
- The challenge was in knowing whether the resources were sufficient or comprehensive enough to cover the research topic.	*-Refer to the list of databases subscribed by the University, existing studies to check their publication venue, and to the SR guidelines for the list of possible venues.*
Determining the quality criteria	
-Whether to develop one quality checklist for both qualitative and quantitative studies?(N1)	*-Two separate checklists were initially created; however, they were later on merged into only one (N1).*
-The challenge was to assess the quality of quantitative and qualitative studies using a single checklist (N3).	*-Two separate checklists were prepared for both types of studies (N3).*
-To identify which quality indicators were suitable or important in determining the studies' quality.	*-The quality checklists were determined by the SR topic and whether studies were quantitative or qualitative. Refer to the SR guidelines and books, and discuss with the supervisor.*
Determining study selection process	
-Whether titles and abstracts are sufficient to base the initial selection on?(N1)	*-In addition to titles and abstracts, conclusions should also be considered (N1).*
Creating data extraction forms	
-Difficulties in finalizing the data extraction form in terms of what data items to include so that they effectively answer the RQs.	*-Revise RQs and refer to SR guidelines and existing SRs.*
Creating data synthesis forms	
-Difficulties in identifying a strategy to synthesize the evidence – whether feasible or not to aggregate results using meta-analysis.	*-Refer to SR guidelines and existing SR and learn by example.*
-Was not sure what was to be included in the data synthesis forms?	*- Discuss with supervisor and refer to existing literature.*

Table 3: *Summary of problems encountered while conducting the SRs*

Challenges/problems faced	**Strategy to overcome the problems**
Conducting the searches	
-Few or no results returned from the complete search string.	*- Consult the subject librarian and use a smaller search string.*
-ACM Digital Library returned millions of results and did not let searches be restricted to only abstracts and titles.	*-Contacted ACM. They indicated their limitations as well as the fact that searches could be restricted using other criteria that could be applied once the results were displayed.*
-ACM & Springer databases returned many irrelevant studies.	*-Performed manual searches to identify relevant studies.*
-Large result sets and duplicate results across different databases (N1).	*-Upon contacting IEEE Digital Library and SpringerLink, both indicated that IEEE Xplore and LNCS could be ignored respectively as IEEE Xplore gave same results as IEEE Digital Library and LNCS gave same results as SpringerLink (N1).*
Documentation of search process	
-Some databases did not provide the ability to download the search results (e.g. ACM Digital Library) making it very time consuming to manually document the results.	*-Search results were manually stored for each of the databases that did not allow downloading in an appropriate format.*
Selection of studies	
-The searches returned many totally irrelevant results.	*-Results from totally different areas were removed.*
-Quality of the abstracts was very poor.	*-Read full texts and contacted Study authors for clarification.*
-Limiting SR to studies specific to database-driven applications lead to only fewer results (N1).	*-The scope of the SR was broadened to include all types of applications (N1).*
Applying quality criteria	
-Some questions could not be answered.	*-A conservative answer was recorded. Authors were contacted too.*
-It was difficult to assess a study's quality using a subjective measure.	*-Discussed the quality issue in a review meeting.*
Extracting data	
-Some of the data related to the research questions was not found in the studies (N3).	*-The fields for unanswered bits were left blank (N3).*
-Sometimes studies did not discuss/report findings/methodology clearly or explicitly (N1).	*-Only what was mentioned in the paper was used, instead of making any assumptions. Also, a review meeting was conducted to validate the data extracted (N1).*
Synthesizing & analyzing data	
-Finding patterns in the data was not easy because no universal measures were used in the included studies.	*- Models described in the included studies were not compared.*
-Insufficient evidence to answer the RQ (i.e. more detailed data needed to be extracted from the paper)	*-Revised the data extraction form.*
-Tabulating important data was a challenge. Some research questions were unanswered and the tables were not filled out for synthesized data (N3).	*-Revised the tables for data synthesis. The empty tables were included in the review report identifying gaps for further investigation (N3).*
Reporting the results of SR	
-Space constraints due to number of pages allowed for publication.	*-Web links were added to present detailed reports.*
-Lesser number of included studies appeared as threat to the importance of investigated research area (N3).	*-Further depth was added to the discussion section of the review to highlight large gaps in the body of knowledge and it was also supported by the literature that was relevant but did not pass SR's specific I/E criteria (N3).*
Other issues	
-Some of the research questions were not addressed in the studies (N3).	*-The unanswered questions became research gaps.*
-The studies which met the inclusion criteria were only 4 (N3).	*-The retrieval of a very small number of primary studies, which may be interpreted by some as an indication of the research area not being worthy of investigation; however can also indicate that the research area is new and therefore needs to be investigated (N3).*

5. COMPARISONS OF NOVICES' EXPERIENCE WITH EXPERT's EXPERIENCE

In this section the comparison of novices' experiences with that of the expert (Mendes) is presented. The experiences reported by Mendes are based on the following two systematic reviews:

- Barbara A. Kitchenham, Emilia Mendes, Guilherme H. Travassos. Cross versus Within-Company Cost Estimation Studies: A Systematic Review [8].
- Emilia Mendes. A Systematic Review of Web Engineering Research [12].

The first SR was conducted by the expert with two other colleagues. However, the expert only narrates the problems faced by her that fit the context of this study. Other non-comparable problems such as those related to being geographically apart have not been reported as they did not apply to the novices.

Some commonalities were observed when comparing novices' experiences to the expert's experience both in developing the protocol and in conducting the SR. The common problems are summarized in Table 4.

Table 4: *Common problems faced by the expert and novices while conducting SRs*

Steps of the SR	Summary of challenges faced by the expert and novices
Formulation of RQ(s)	*Took a while to finalize the RQ(s)*
Devising search strategy	*Search string too long/complex*
Determining Inclusion/ Exclusion criteria	*To identify precisely the conditions under which a study would be of importance for the SR*
Conducting the searches	*Troublesome to run searches; overall process was extremely time consuming*
Extracting data	*Very time consuming*
Reporting the SR	*Page limit of conferences and journals*

Defining the PICOC was a challenge to the expert as it required some re-thinking of the meaning of population within the context of a SR. As for the SR novices, they found that the comparison part was not applicable for any of the three SRs. In terms of determining the quality criteria, the problem reported by the expert was related to the use of quality criteria that mixed the quality of reporting and quality of the methodology/analysis. In contrast, the SR novices had difficulties to identify suitable questions sufficient enough to assess the quality of studies included in the SR. On other issues, the expert mentioned that being geographically apart did add to the amount of effort needed to carry out the SR as the communication was done mostly by email. As for the novices, all SRs were conducted locally.

Overall, the time taken on average by the expert to conduct the SRs was of 2.5 weeks per person working full time. This in its own is a large difference when compared to the SR novices' experiences, as they each spent at least 8.5 months. None that one of the SRs conducted by Mendes contained more than 100 primary studies, thus in the case of the expert, the volume of primary studies was not related with the time taken to carry out a SR.

6. DISCUSSION

An SR is in our view a very useful methodology for collecting evidence on a topic of interest, thus playing an important role in establishing the context of research on a given topic. We also believe that the presence of concrete guidelines for conducting SRs and experiences of people performing SRs can greatly strengthen the SR body of knowledge. In this Section, we present a discussion of our experiences compared to that of other researchers who have also published papers aimed at reporting their experiences conducting SRs. A brief comparison is presented in Table 5, followed by a detailed discussion.

Table 5: *Comparison of findings of the previous studies with the findings of this research*

Findings common in this research and previous studies	Findings of this research
Difficulty in defining the scope and relevance of RQs [3] (N1, N2, N3).	*-It is more difficult for novices to define the scope and relevance of RQs than it is for experts.*
Difficulty in finding synonyms and keywords for compiling the search string [3] [4] (N1, N2).	*-It is not always the case that the RQs are based on each component of the PICOC template. Some components may not be applicable depending on the context of the SR.*
Difficulty in dealing with different databases [3] [4] [5] [17] (N1, N2, N3).	*-Keywords from relevant papers, if found relevant, may be included for compiling search strings, even if the probability of their inclusion in the SR is low.* *-It is useful to consult a domain expert for identifying search terms and thereby compiling the search string.*
Unstructured and badly written abstracts pose difficulty in study selection process [3] [5] (N1, N2, N3).	*-The search string may be broadened with the help of the subject librarian if the complete search string yields very few or no results.*
Difficulty in primary study quality assessment due to the reason that the quality related aspects are not mentioned in the primary studies [5] [6] [17] (N1, N2, N3).	*-Simpler search strings lead to manual searching within the search results, requiring extra time and effort.*
Missing values in the primary studies while extracting data [3] (N3).	*-I/E criteria can be difficult to determine. It should be based on the domain as well as on the importance of study in terms of its relevance to the SR.*
Data extraction forms are revised in the data extraction step [3] (N1, N2, N3).	*-The effectiveness of using two separate quality checklists for both qualitative and quantitative study types over one quality checklist, or vice versa, is not clear.*
	-Creation of data extraction and data synthesis forms is iterative.
	-An unanswered RQ is likely to imply that there is a gap in the research under review by that RQ.

We observed that the formulation of the research questions is equally challenging for both novices and experienced researchers. Previous studies [3] [4] and our experiences for conducting SRs narrate the difficulty in defining the scope and relevance of research questions. However, the definition for the research questions' coverage criteria is more of a challenge for novices than it is for expert researchers. For novices, it is helpful to consult supervisors and domain experts, whereas for experienced researchers a systematic mapping study can be very effective [3]. In our case i.e., for novices, we revised the research questions of all three SRs we carried out until we were satisfied that the RQs covered all the aspects of the topic area on which we wanted to collect the evidence.

In our case i.e., for novices, none of the SRs had the 'comparison' context of PICOC. None, except for one of the previous studies, have discussed PICOC before, saying that it was difficult to fit the research questions to the PICOC template [17]. We believe that PICOC depends totally on the focus and domain of the SR and hence it may not always be feasible to use all of its components.

Identifying keywords and synonyms was found difficult in previous studies [3] [4]. N1 also encountered the problem of identifying keywords from relevant studies (see Table 2). In such scenarios, we believe that the search terms can be identified effectively with the help of domain experts – for novices, the supervisors can provide the help.

A common problem the novices of this study faced and which is not explicitly mentioned in previous studies is finding a very small number of studies when applying a complex search string. The solution was to use broader search strings. It can also be useful for novices and expert researchers to consult subject librarians and seek their help to define search strings that can find most suitable results and to also deal with issues related to using numerous databases. This is what the SR novices did when conducting their SRs.

We, the novices and the expert, observed similar issues with the behaviour of various databases that in the past also caused problems to experienced researchers during the search process (see Tables 1, 2 and 3). The use of simpler search strings, as suggested by both the supervisor and the subject librarian, proved very helpful for the novices. However, simpler search strings further led to more manual searches within search results so increasing substantially the amount of work. Also for search processes, the two major challenges faced by the novices were determining which years to include in the searches and not knowing which sources were more appropriate for running searches. Both problems were perhaps due to the SR novices' lack of experience with the search process. To tackle the first problem, the novices considered the searches to be run from the year in which the first study of the relevant topic was published (as suggested by the supervisor); and as solution to the second problem the novices relied upon the suggestions made by their supervisor and also in the guidelines for conducting SRs in SE [7], as well as the databases subscribed for by the University of Auckland's library.

While comparing our experience with that reported in previous studies, we observed that none of those studies reported any problem relating to determining the inclusion/exclusion criteria. However, this problem was also faced by the expert i.e., the fourth author. Our combined experiences suggest that the definition of the inclusion/exclusion criteria depends on the domain under study as well as on how important the quality of the studies is with regards to their relevance and importance to the SR.

While developing the protocol and determining the quality criteria, the novices also faced certain problems. It was not clear whether to have one quality checklist or two separate checklists for qualitative and quantitative studies. Therefore, two separate checklists were developed for both qualitative and quantitative studies to be used for study quality assessment (N1 and N3). In our opinion, there should be some discussion in the guidelines on the effectiveness of using a single or two separate quality checklists for different study types under different situations. A single set of questions to assess both qualitative and quantitative studies enables for the quality of all primary studies to be compared throughout; however the use of a single checklist may compromise the level of quality assessment of the evidence used in a SR. Finally, we also believe that the items in the quality checklists are greatly determined by the SR topic.

It is also our experience that creating the data extraction and data synthesis forms while creating the review protocol is an iterative activity, and is largely determined and should be aligned with the RQs.

With regards to study selection, we observed that the problem with unstructured and badly written abstracts posed issues for novices as well as experts (see tables 1 and 3). This resulted in reading more full texts and posed an extra challenge in terms of the time spent on this phase. We also had to consult primary studies' authors for clarification on various aspects of the studies. However, we observed a low response rate which we believe is a possible threat to the validity of our SRs as this may have affected the number of primary studies selected.

In relation to quality assessment, novices as well as expert researchers face the problem of determining the quality of the evidence provided by the selected studies. In addition, many quality-related aspects are not reported, thus resulting in either the exclusion of the study or the granting of a low quality score.

While extracting data both novices and experts faced similar issues of missing values and tabulating the results without altering the data extraction forms (e.g., see [3] and table 3). The novices mainly narrated the results in simple forms whereas experts faced more specific problems of synthesizing data according to studies' types (e.g., see [5]). The latter devised their own methods not explicitly mentioned in the SR guidelines. We believe that data synthesis is always in accordance with the RQs and should attempt to answer the RQs. However, if a RQ remains unanswered due to lack of evidence, this may indicate a research gap.

While reporting an SR, the problems observed by novices as well as experts are similar, i.e., limited space allowed by conferences and journals.

One of the limitations of this research is that it only reports the experiences of three PhD students. We believe that a larger sample size can be a better indicative of problems that are more general to SRs and specific to novices. We also believe that little domain knowledge and lack of prior experience in conducting SRs were two main reasons for many of the novice-related problems reported in this paper. The degree of such issues can be substantially reduced with the help of the supervisor and subject librarian.

7. CONCLUSIONS

In this paper, we have presented first hand experiences of novices – PhD students – conducting SRs and have compared their experiences with that of the expert researchers. We have observed that novices take much more time to conduct SRs than experts. We have also observed that while some experiences are common to both novices and experts, certain problems are only faced by novices e.g., formulating RQs, selecting studies, creating data extraction and synthesis forms, etc. The problems while determining quality criteria are also different in nature than the problems with quality criteria faced by expert researchers. Almost all of the issues encountered by novices are a result of little domain and/or SR knowledge and can be overcome by consulting the supervisor and/or domain expert and by referring to SR guidelines and published literature on SRs in SE. However, we believe that for novices to conduct quality SRs, following suggestions may be the best to follow:

- Seek help from an expert in the area of SRs so that important decisions especially the protocol development can be reviewed and the judgement bias may be removed.
- Gain a good knowledge of the SRs by reading and understanding the SR guidelines before attempting to conduct SRs.
- Expect the RQs as well as other parts of the protocol to evolve while conducting SRs. This will lead to extra work as well as rework in some cases.
- Consult a subject librarian and the supervisor as much as possible.
- Document each and every step thoroughly.
- In case of ambiguities, try to contact authors of primary studies instead of considering their exclusion from the SR.

The common issues observed that are faced by both novices and expert researchers are those related to databases and resource to be searched, quality assessment, quality of the primary studies, and space for reporting reviews. These problems need special attention and can benefit SRs greatly, if solved. We, therefore, believe that there is an urgent need for a single integrated software tool that in addition to providing a common interface for searching databases and for dealing with issues related to different models on which these databases are created, can also provide assistance with other steps of conducting SRs. We also believe that more work is required to strengthen the field of SRs, especially to deal with the above mentioned issues.

8. REFERENCES

[1] Babar, M.A. and Zhang, H. (2009) Systematic literature reviews in software engineering: Preliminary results from interviews with researchers. *Empirical Software Engineering and Measurement (ESEM 2009)*, Lake Buena Vista, FL, USA,15-16 Oct. 2009.

[2] Baldassarre, M. T., Caivano, D., Kitchenham, B., and Visaggio, G. (2007) Systematic review of statistical process control: An experience report. *Evaluation and Assessment in Software Engineering (EASE'07)*, UK, April 2007. British Computer Society.

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

[4] Biolchini, J. C., Mian, G.M., Natali, A.C.C., Conte, T. U., and Travassos, H.H. (2007) Scientific research ontology to support systematic review in software engineering. *Advanced Engineering Informatics* 21 (2007), pp. 133-151.

[5] Dyba, T., Dingsoyr, T., and Hanssen, G.K. (2007) Applying systematic reviews to diverse study types: An experience report. *Empirical Software Engineering and Measurement (ESEM 2007)*, Madrid, Spain, 20-21 Sept. 2007, pp. 225-234.

[6] Dybå, T. and Dingsøyr, T. (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*, Kaiserslautern,

Germany, 9-10 Oct. 2008. ACM, New York, NY, pp. 178-187.

[7] Kitchenham, B. and Charters, S. (2007). *Guidelines for Performing Systematic Literature Review in Software Engineering*. EBSE Technical Report, 2.3. Keele University.

[8] Kitchenham, B.A., Mendes, E., and Travassos, G.H. (2007) Cross versus Within-Company Cost Estimation Studies: A Systematic Review. Software Engineering, *IEEE Transactions on Software Engineering*, vol.33, no.5, pp. 316-329. May 2007.

[9] Kitchenham, B.A., Dyba, T., and Jorgensen, M. Evidence-based software engineering. *26th International Conference on Software Engineering (ICSE '08),* 23-28 May 2004, 273 – 281.

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

[11] Kitchenham, B., Bereton, P., Turner, M., Niazi, M., Linkman, S., Pretorius, R., and Budgen, D. (2009) The impact of limited search procedures for systematic literature reviews – A participant-observer case study. *3rd International Symposium on Empiricl Software Engineering and Measurement (ESEM 2009)*, Lake Buena Vista, FL, USA,15-16 Oct. 2009.

[12] Mendes, E. A Systematic Review of Web Engineering Research (2005) *ACM/IEEE International Symposium on Empirical Software Engineering*, Noosa heads, Australia, pp. 408-418.

[13] Mian, P., Conte, T., Natali, A., Biolchini, J., and Travassos, G. (2005) A Systematic review process for software engineering. *Software Engineering Latin American Workshop (ESELAW'05)*, Brazil.

[14] Oates, B. J. and Capper, G. (2009) Using systematic reviews and evidence-based software engineering with masters students. *13th International Conference on Evaluation and Assessment in Software Engineering (EASE 2009)*, 20 - 21 April 2009.

[15] Riaz, M., Mendes, M., and Tempero, E.D. (2009). A Systematic Review of Software Maintainability Prediction and Metrics. *3rd International Symposium on Empirical Software Engineering and Measurement (ESEM 2009)*, Lake Buena Vista, FL, USA,15-16 Oct. 2009, pp. 367-377.

[16] Salleh, N, Mendes, E. and Grundy, J. (2009) Empirical Studies of Pair Programming for CS/SE Teaching in Higher Education: A Systematic Literature Review. *IEEE Transactions on Software Engineering* (paper submitted for review), 2009.

[17] Staples, M., and Niazi, M. (2006) Experiences using systematic review guidelines. *10th International Conference on Evaluation and Assessment in Software Engineering (EASE 2006)*, Keele University, UK, 10 - 11 April 2006.

[18] Staples, M. and Niazi, M. (2008) Systematic review of organizational motivation for adopting cmm-based software process improvement. *Information and Software Technology*, 50(7-8):605–620, 2008.

[19] Sulayman, M. and Mendes, E. (2009) Systematic Review of Software Process Improvement in Small and Medium Web Companies. *International Conference on Advanced Software Engineering & Its Applications (ASEA 2009)*, 10-12 Dec. 2009, Jeju Island, Korea, pp. 1-8. Springer-Verlag Berlin Heidelberg.

[20] TR/SE0401. (2004) Kitchenham, B. *Procedures for Performing Systematic Reviews*. (Keele University) and Technical Report 0400011T.1 (National ICT Australia).

Appendix

Sr. No.	**Steps of the SR**	**Challenges/problems faced**	**How were the challenges/ problems overcome?**
1	*Formulation of research questions*		
1a	*Identification of PICOC*		
2	*Devising Search Strategy (from identifying synonyms to applying Boolean AND)*		
3	*Determining Inclusion/Exclusion Criteria*		
4	*Determining Search Phases (both primary & secondary)*		
5	*Determining the quality criteria*		
6	*Determining study selection process (both initial & final)*		
7	*Creating data extraction forms*		
8	*Creating data synthesis forms*		
9	*Conducting the searches (Search phases primary & secondary)*		
10	*Documentation of search process*		
11	*Initial Study selection*		
12	*Final Study Selection*		
13	*Applying quality criteria*		
14	*Extracting data*		
15	*Synthesizing & analyzing data*		
16	*Reporting the results of SR*		
17	*Any other issues (such as working with other people, being geographically apart, resource availability, not being very expert either on the domain or on SRs, access to resources, primary studies' authors not replying, etc.)*		

***Note**: Steps 1-8 relate to protocol development*