2013 ACM / IEEE International Symposium on Empirical Software Engineering and Measurement

Identifying Barriers to the Systematic Literature Review Process

Jeffrey C. Carver*, Edgar Hassler†, Elis Hernandes‡, and Nicholas A. Kraft*
*Dept. of Computer Science; Univ. of Alabama; Tuscaloosa, AL, USA
†Dept. of Information Systems, Statistics, and Management Science; Univ. of Alabama; Tuscaloosa, AL, USA
‡LaPES - Software Engineering Research Lab; Federal Univ. of São Carlos; São Carlos, SP, Brazil
carver@cs.ua.edu, ehassler@cba.ua.edu, elis_hernandes@dc.ufscar.br, nkraft@cs.ua.edu

***Abstract*—Conducting a systematic literature review (SLR) is difficult and time-consuming for an experienced researcher, and even more so for a novice graduate student. With a better understanding of the most common difficulties in the SLR process, mentors will be better prepared to guide novices through the process. This understanding will help researchers have more realistic expectations of the SLR process and will help mentors guide novices through its planning, execution, and documentation phases. Consequently, the objectives of this work are to identify the most difficult and time-consuming phases of the SLR process. Using data from two sources — 52 responses to an online survey sent to all authors of SLRs published in software engineering venues and qualitative experience reports from 8 PhD students who conducted SLRs as part of a course — we identified specific difficulties related to each phase of the SLR process. Our findings highlight the importance of planning, teamwork, and mentoring by an experienced researcher throughout the process. The paper also identifies implications for the teaching of the SLR process.**

Keywords*—*systematic literature review, survey, empirical software engineering.

I. Introduction

Literature review is the basis of all research. Researchers perform literature reviews to motivate new research or to summarize the state of the art in a particular area. That is, a literature review can be used to establish the foundations of a new investigation or to summarize what is currently known or unknown about a topic.

Software engineering (SE) researchers have traditionally performed *ad hoc* literature reviews (i.e., by searching databases and following references). The primary drawback of an unsystematic review is that the lack of rigor can bias the results or cause the researcher to omit important relevant literature, thereby changing the nature of the conclusions.

Medical researchers defined the systematic literature review (SLR) process to mitigate the drawbacks of *ad hoc* reviews. An SLR is a formal, repeatable method for identifying, evaluating and interpreting the available research regarding a topic or question of interest. The primary difference between an SLR and an *ad hoc* review is the level of advanced planning. Prior to conducting the review, the researchers develop a protocol that documents: the research questions that guide the review, the search strategy (including specific databases and keywords), the criteria for selecting appropriate papers, a method for assessing the quality of the selected papers, the specific data to be extracted from each paper, and a plan for analyzing and synthesizing the extracted data to draw conclusions.

Medical researchers, practitioners, and policy makers have long relied on SLRs, because they integrate and critically evaluate current knowledge to support decisions about important issues. Seeking these same benefits, the SE community has begun conducting SLRs. Indeed, with the growing emphasis on empirical SE research, SLRs are growing in importance, because they allow researchers to bring together disparate evidence to understand the effects of various SE techniques and tools. Unfortunately, though SLRs are important to the maturation of SE and to the adoption of SE research practices by industry, SLRs are difficult and time-consuming to conduct.

SLRs are often performed by relative novices (e.g., PhD students preparing for their dissertation research), for whom the SLR process can be especially difficult. Therefore, it is important to better understand the most common difficulties researchers face when conducting SLRs so that mentors can be better prepared to guide novice researchers through the process. Based on our experiences with a graduate SE course at the University of Alabama in which 8 PhD students each conducted an SLR, we sent a survey to authors of published SE SLRs to gather information about their experiences with various aspects of the SLR process. We received 52 responses.

The goal of this paper is to use our experiences and the survey results to identify the barriers that researchers, especially novice researchers, face when performing an SLR. We organize the discussion around the primary phases of the SLR process. We anticipate that the results of our analysis will increase not only the quantity, but also (and more importantly) the quality of SLRs (and thus of SE research in general).

The remainder of this paper is organized as follows. Section II provides an overview of the SLR process. Section III describes the data sources used for the analysis. Sections IV and V analyze the study results. Section VI discusses the implications of the results both for researchers and for educators/mentors. Finally, Section VII concludes the paper.

II. Systematic Literature Review Process

For our analysis we use the SLR process described by Kitchenham [13], [14], who first ported the SLR process from the medical field to SE. In this description the process comprises three primary phases: planning, execution, and documentation.

978-0-7695-5056-5/13 $26.00 © 2013 IEEE
DOI 10.1109/ESEM.2013.28

A. Planning Phase

During the planning phase, the researchers define the protocol that guides the SLR execution. The goal of the protocol is to reduce researcher bias and provide a repeatable, transparent process for conducting the SLR. During the planning phase, the researchers should define and document in the protocol, at a minimum, the following information:

- **P1** Motivation for conducting the SLR
- **P2** Research question(s)
- **P3** Search strategy (incl. target databases and search strings)
- **P4** Strategy for identification of primary studies (i.e., inclusion and exclusion criteria)
- **P5** Quality assessment criteria
- **P6** Data extraction form

After defining the protocol, the researchers should solicit an independent expert or panel review for completeness and validity, and after any subsequent changes to the protocol, the expert or panel should review the revision(s).

B. Execution Phase

During the execution phase, the researchers conduct the SLR in five steps:

- **E1** Identify relevant research using the search strategy
- **E2** Select primary studies using the inclusion and exclusion criteria
- **E3** Evaluate selected studies using the quality assessment criteria
- **E4** Extract required data using the data extraction forms
- **E5** Analyze extracted data and synthesize resulting information to draw conclusions

Researchers use the search string to query multiple databases and to identify a set of candidate studies. Next, they use the inclusion and exclusion criteria to eliminate candidate studies that are irrelevant, using titles first, abstracts second, and the full text third. During each iteration, the researchers eliminate a candidate study only when it is clear that the study is not relevant. After selecting the primary studies, the researchers perform a quality assessment of each selected study to evaluate the reliability and importance of its results. Finally, the researchers extract important data from all remaining studies, the extracted data is analyzed, and the resulting information is synthesized. To reduce researcher bias in the process, members of the research team should perform each step independently and then meet to review the results and resolve any conflicts.

C. Documentation Phase

During the documentation phase, the researchers report the results of the planning and execution phases to document the review in a publication. This phase has two steps:

- **D1** Specify dissemination strategy
- **D2** Format SLR report

III. Methodology

We describe our two primary data sources in Section III-A then provide a brief overview of our analysis process in Section III-B.

A. Overview of Data Sources

We drew information from two primary data sources: a graduate course taught by the first author and a survey of published SLR authors. The remainder of this section describes each of the data sources in detail.

1) SLR Graduate Course: In Spring 2012, the first author taught a graduate Advanced Empirical Software Engineering course. There were eight PhD students enrolled in the course, four from Computer Science and four from Management Information Systems. The main focus of this course was for the students to learn about and conduct SLRs. As such, the course had two primary goals:

1) Perform an SLR
2) Evaluate the SLR process

Each student conducted their SLR as a semester-long project. Realizing that most SLRs cannot be completed within one semester, it was expected that work would continue beyond the semester to make these SLRs publishable. At this point, two of these papers have been accepted in conferences [11], [22], at least five other papers will be submitted to various journals and conference in the near future, and most of the papers will become parts of the students' dissertations. In addition, each student acted as a second reader for the SLR of one of their classmates. The first author of this paper oversaw all SLRs, provided input on the protocols, and helped to resolve any conflicts during the paper selection and data extraction steps.

In addition to the interaction among the primary author and the second reader throughout the semester, a large portion of the class meeting time was devoted to discussing each SLR and to evaluating the SLR process. Each student had the opportunity to present his protocol and to receive feedback from his classmates. The class also spent a substantial amount of time discussing the logistics of the SLR process and identifying common problems encountered by multiple students. Each student, utilizing notes retained throughout the process, wrote a report at the end of the semester describing their SLR process, noting their difficulties, and suggesting improvements.

Each student produced two deliverables for the course:

1) Draft SLR report, which in most cases needed further revision to become publication-ready
2) Report describing experiences with following the standard SLR process, including specific areas in which difficulties were encountered

2) Survey of SLR Authors: To obtain more detailed insight into the SLR process, we created a list of SLRs published in SE venues by searching publisher's websites (IEEE, ACM, Elsevier, Springer and Wiley) and Kitchenham et al.'s SLR on SLRs [15]. Using that list of SLRs, we generated a list of authors. We sent to the authors a survey that asked them to describe: the SLR process followed, the difficulties they encountered in the SLR process, the time they spent during

the SLR process, and the aspects of the SLR process most in need of tool support. We received 59 responses to this survey. Table I lists the survey questions, which include free-response and multiple-choice questions.

B. Overview of Analysis Process

We performed a qualitative analyses of the graduate student experience reports and the survey responses independently to discern differences between novice and experienced researchers. We then synthesized these results. We employed a grounded theory approach to examine the qualitative responses [21].

We first analyzed the responses with open coding, providing an initial level of conceptual extraction. Coding is the process of breaking down the data [21] for the purpose of discovering theoretical concepts, their properties, and the relationships among concepts. The data may be examined at many levels; the analyst may examine the data elements of words, phrases, sentences, paragraphs, or entire documents [21]. In addition, a researcher may move among the levels to examine the data from multiple perspectives. Coding entails the examination of a data element and attaching a meaningful conceptual label to it [21]. Coding often takes place in the form of margin notes. However, coding could also be performed on note cards, or in the form of database elements. A critical element is that coding should not be a mere restatement of the data. Theoretical elements must be at a higher conceptual level than the data from which they are derived.

Following the initial open coding process, we performed axial coding. The axial coding process is similar to open coding, but focuses on making connections between the elements identified in the open coding process. As axial coding concludes, a tentative core variable that explains the behavior in question is selected. Utilizing the core variable as a guide, researchers then recode data to provide a clearer conceptual model of the underlying phenomenon.

To alleviate bias in the process, two of the authors independently performed the coding process. After each round of coding the authors met to compare results of the coding process for agreement and to resolve any differences in coding. The results of the coding meetings indicated a high degree of convergence in the codes identified with only minor differences were easily resolved through the application of synonyms. The results of this coding process appear in the sections that follow.

In addition to the qualitative data, the survey produced a large amount of quantitative data. We analyzed the quantitative data using traditional descripive and inferential statistics along with graphical presentations in charts.

IV. General Results

This section provides an overview of the data via descriptive statistics.

A. Expertise

We evaluated the SLR expertise of the survey respondents using both an objective metric and a subjective metric. Objectively, the respondents had co-authored 3.19 SLRs on average (median of 2.5) with 2.5 of those SLRs having been published

TABLE I. Survey questions

1	What was your motivation for conducting your SLR? ○ Want to learn about a new area ○ Expert in the area trying to summarize the literature ○ Other: _____
2	For how many SLRs have you been a co-author?
3	How many of those SLRs have been accepted/published in a peer-reviewed conference or journal?
4	What are the top three most difficult aspects of the SLR process? ___ Defining the research question(s) ___ Identifying appropriate keywords ___ Creating search strings ___ Searching the databases ___ Selecting appropriate papers to include ___ Extracting data from the papers ___ Assessing the quality of the papers
5	What are the top three most time-consuming aspects of the SLR process? *Same options as Question 4*
6	What are the top three aspects of the SLR process most in need of tool support? *Same options as Question 4*
7	Did your protocol deviate from the standard SLR procedure as originally defined by Kitchenham? (Yes/No)
7a	If so, how did your protocol deviate from the standard SLR procedure and why?
8	Prior to beginning your review, did you have your protocol reviewed by someone outside the author team? (Yes/No)
8a	If yes, did you make any changes based on the review? Explain
8b	If no, why did you not have your protocol reviewed?
9	During the review process, did you deviate from your protocol? (Yes/No)
9a	If so, why did you deviate from the protocol?
10	How would you characterize the expertise of the author team? ○ All authors have the same level of expertise ○ Some authors have more expertise than other authors ○ Other: _____
11	What was the biggest problem you faced when searching the various databases?
12	How was the exclusion at the title level performed? ○ One researcher worked alone ○ One researcher made a first pass with another reviewing a sample ○ Each was reviewed by two or more researchers ○ Other: _____
13	How was the exclusion at the abstract level performed? *Same options as Question 12*
14	How was the exclusion at the full-paper level performed? *Same options as Question 12*
15	If there were conflicts regarding paper exclusion, how were they resolved?
16	Which citation manger was used ot support your search? ○ BibTeX ○ RefWorks ○ EndNote ○ None ○ Mendeley ○ Other: _____
17	Did you perform a formal quality assessment of each paper? (Yes/No)
17a	If so, how did you create the assessment criteria? ○ Reused an existing quality criteria ○ Created a new one ○ Some combination of the two ○ Other: _____
17b	If so, how many papers were excluded based upon the quality assessment? ○ 0% ○ 10%–25% ○ <10% ○ >25%
17c	If so, if your quality assessment was performed by more than one pseron, how were conflicts resolved?
17d	If not, why did you not perform a formal quality assessment of each paper?
18	How was the data extraction performed? ○ Two or more researchers independently extracted data from each paper and compared results ○ One researcher extracted data and another reviewed the results ○ One researcher extracted data from all papers and another extracted data from a sample
19	If there were conflicts during the data extraction step, how were they resolved?
20	How did you manage the extracted data?

Fig. 1. Most Difficult Aspects of SLR Process (Question 4)

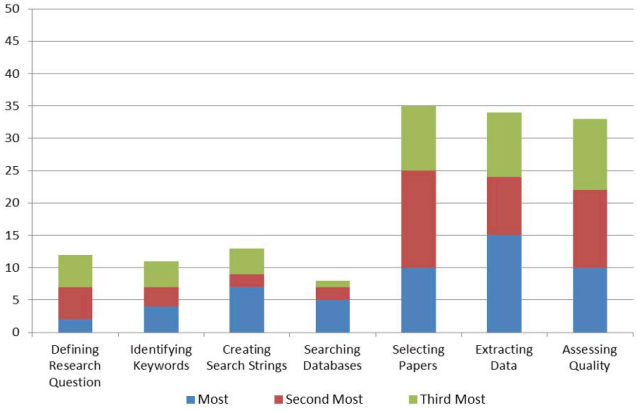
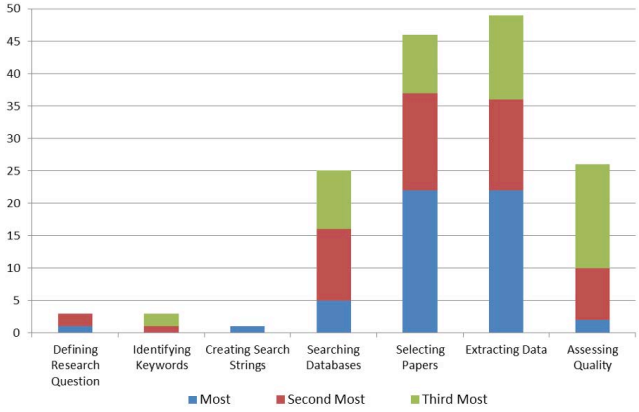


Fig. 2. Most Time-Consuming Aspects of the SLR Process (Question 5)



(median 2). Subjectively, 28 of the 59 respondents (47.5%) indicated that their motivation for conducting their SLRs was that they were experts who wanted to summarize the literature. In addition, 47 of the 59 respondents (79.7%) indicated that the author team had different levels of expertise.

B. Overall Evaluation of SLR Process

Questions 4, 5 and 6 (Table I) asked respondents to rate the following seven aspects of the SLR process on three different scales, as described below:

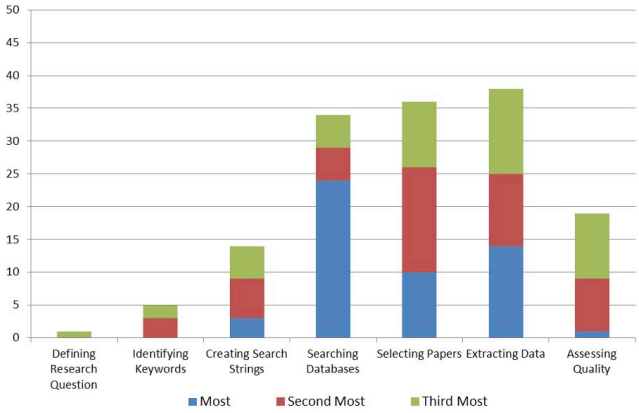
1) Defining the Research Questions
2) Identifying Appropriate Keywords
3) Creating Search Strings
4) Searching the Databases
5) Selecting Appropriate Papers to Include in the Review
6) Extracting Data from the Papers
7) Assessing the Quality of the Papers

First, regarding the **Most Difficult** aspects of the SLR process, Figure 1 illustrates the responses to Question 4 indicating which aspects of the SLR process the respondents found most difficult, second most difficult and third most difficult. As the figures shows, *Extracting Data* was the most difficult, followed by *Assessing Quality* and *Selecting Papers*.

Second, regarding the **Most Time-Consuming** aspects of the SLR process, Figure 2 illustrates the responses to Question 5 indicating which aspects of the SLR process the respondents found most time-consuming, second most time-consuming and third most time-consuming. As the figure shows, *Extracting Data* and *Selecting Papers* were the most time-consuming aspects of the SLR process, followed by *Searching Databases* and *Assessing Quality*.

Third, regarding the aspects of the SLR process **Most in Need of Tool Support**, Figure 3 illustrates the responses to Question 6 indicating which aspects of the SLR process the respondents found most, second most, and third most in need of tool support, respectively. As the figure shows, *Searching Databases* was by far the aspect most in need of tool support, followed by *Selecting Papers* and *Extracting Data*.

Fig. 3. Aspects of SLR Process Most in Need of Tool Support (Question 6)



C. Protocol

Next, the survey asked the respondents about their overall protocol design and execution. First, when asked whether their protocol deviated from the original definition by Kitchenham [13], [14], 65% said that it did not. Furthermore, when asked whether the respondents' deviated from their protocol while executing the SLR, again 65% said they did not. Interestingly, the answers to these two questions were not significantly dependent upon each other ($\chi^2 = .898$).

Surprisingly, 55% of respondents stated that they did not have their protocol reviewed by external experts. The analysis of the qualitative explanations as to why the protocols were not reviewed (Question 8b) typically fell into three categories:

- **Belief that the team's expertise was sufficient to design a valid protocol (14 cases)**. One respondent summarized this point as "It wasn't required as we authors were domain experts." Fortunately, this extreme position only occurred in 4 cases. Two cases indicated that a review was conducted by a PhD student's advisor. The other cases suggested that the large research team possessed enough expertise to conduct the review internally.
- **Lack of resources (9 cases)**. The responses here included: lack of accessible domain experts (3 cases),

lack of access to SLR experts (4 cases), time constraints (1 case) and extensive, iterative changes during the review, making re-reviews impractical (1 case).

- **Reuse of an existing protocol (5 cases)**. The belief that a previously published protocol could be utilized as a plug-and-play template dominated the responses.

Of the 45% that did have their protocol reviewed, 84% made changes to the protocol based on that review. Those changes helped the authors to improve the search process, search criteria, data extraction procedures, and other details. The review also helped authors formulate additional research questions, refine definitions, and make other changes to enhance the overall SLR process. Table II details the most common changes resulting from protocol review.

TABLE II. COMMON CHANGES RESULTING FROM EXTERNAL REVIEW

Category	Nature of Change	# Cases
Search Process	Keywords	6
	Search Strings	3
	Data Sources	3
	Procedural	2
Criteria	Inclusion/Exclusion	4
	Quality Assessment	3
Extraction	Form Design	3
	Procedural	1

In addition, during the SLR graduate course, the students' stated that the frequent discussions of their protocols with each other and with the professor helped in the definition and refinement of the protocols. The students also found it helpful for their SLR process to be guided by the professor, who is experienced in conducting SLRs. These experiences are not unique. Guidance from a more experienced researcher is crucial to the accuracy and success of the review [16].

V. SPECIFIC PROTOCOL ITEM RESULTS

This section uses data gathered from the data sources described in Section III-A to analyze the specific steps in the SLR process. The goal of the analysis is to identify aspects of the SLR process that the PhD students and SLR authors found particularly difficult, time consuming, and in need of tool support. This section is organized around the *Planning* and *Execution* phases of the SLR process as defined in Section II. In cases where the planning and execution items overlap, the results are discussed together. For the sake of clarity this section discusses data from both sources together under each heading. Note that the survey respondents did not provide any discussion regarding *P1: Motivation for conducting the SLR* or *E5: Synthesize Data to Draw Conclusions*, so they are excluded from the following discussion.

A. P2: Research Questions

The research questions are arguably the most important aspect of the protocol because they drive the remainder of the protocol. The experiences in the SLR Graduate Course indicated the types of problems that can arise if the research questions are not appropriately scoped. If the questions are too broad, searches will return too many papers to reasonably evaluate in one SLR. Conversely, if the questions are too narrow, searches will not return enough papers from which to draw useful conclusions. Scoping of the research questions is an activity in which feedback from more experienced researchers through a protocol review can be particularly beneficial. Other researchers have also noted problems with defining research questions [6].

Suggestions for novices about the Research Questions are:

- Research questions must be scoped properly
- Expert feedback can guide improvements

B. P3/E1: Search Strategy

The search strategy includes both the creation of the search string(s) from the research questions and the use of those strings in various databases This section first describes issues with search string creation in general and then discusses specific issues with executing the database searches.

Creating the search string(s) can be an iterative process as the SLR authors attempt to define the appropriate set of keywords and synonyms that cover the research space. Regarding the development of search strings in general (i.e., not including the differences among databases), the experiences in SLR Graduate Course resulted in the following observations. First, adding a '*' at the end of a key term helps to identify variant spellings. Second, when using a common term like "Open Source," restrict the search to the title, abstract and keywords to limit the number of irrelevant hits. Third, in some cases a large number of relevant papers were not returned by the initial search, causing the search string to be refined based upon the results of a secondary search (i.e., reviewing references in the identified papers) a technique called 'snowballing' [10]. The survey respondents mentioned that changes to search strings was one of the reasons they might deviate from their protocol. Table III lists the other detailed responses regarding search strings.

TABLE III. COMMON CHANGES TO SEARCH STRINGS

Category	Nature of Change	# Cases
Definition	Adapt to Different Syntax	18
	Adjust for Character Limitations	5
	Adapt to Available Search Fields	1
	Irrelevant Results Returned	1
Refinement	Initial Search String Returned No Results	4
	Refactor to Balance Precision and Recall	4
	Modify Keywords for Some Sources	1
	Adjust the Search Process	1
	Incorporate Industry Focused Material	1

Once SLR authors define the search string(s), they typically query multiple databases using the string(s). The experiences in the SLR Graduate Course suggest the following observations regarding the database searches. First, databases have different behavior for the same search string, an observation also noted by Babar and Zhang [2]. For example, in some cases, changing the order of the keywords in the search string could change the result set. In other cases, Boolean logic does not work as

expected (e.g., *AND*-ing two terms together increases the size of the result set, while *OR*-ing two terms reduces the size of the result set). These differences force SLR authors to develop distinct search strings for each database. Zhang and Babar note similar results [23]. Second, the advanced search functionality differs across databases. In some cases an advanced search returns a different result set than a basic search, even when the same search string is used. Third, there is a large overlap in the literature contained in the databases commonly used for SLRs, creating the need to identify and remove duplicate studies from the result set. A problem also observed by Chen et al. [4]. Fourth, databases differ in the content and format of the citation information provided. Finally, the databases are not consistent in their behavior regarding bulk export of references to a citation manager.

The survey respondents reported issues similar to those experienced in the graduate course. The main problems reported were inconsistencies in input format, behavior, and usability among the databases. Table IV provides more details about the specific problems identified in the survey.

Interestingly, a number of published SLRs also identified issues with the databases and the database search process [1], [3], [5], [7], [8], [16], [17], [19].

Suggestions for novices about the Search Strategy are:

- Tailor the search strategy for each database
- Plan to manipulate the search string for each database
- Databases overlap, so define a strategy for identification and elimination of duplicates results
- Do not underestimate the effort required to combine references exported from different databases

C. P4/E2: Identification of Primary Studies

After obtaining search results, SLR authors must determine which papers to include in the SLR. This step encompasses both the definition of the inclusion/exclusion criteria (P4) and the application of those criteria to select the papers (E2).

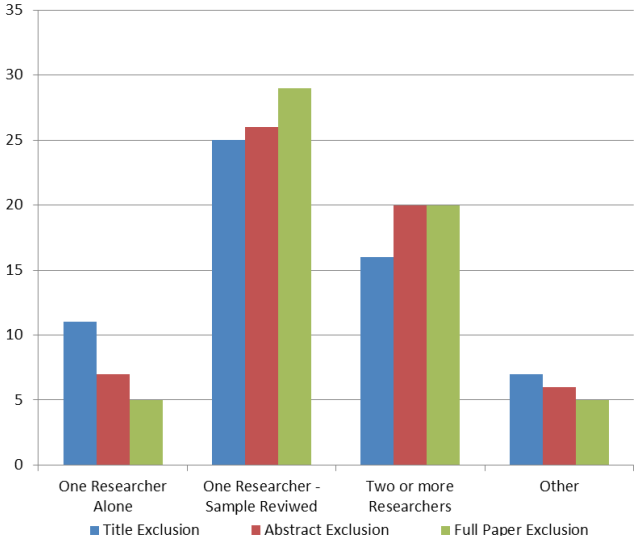
The SLR Graduate Course provided the following observations regarding the definition of inclusion/exclusion criteria:

1) They must conform to the goals of the current SLR and should not be copied verbatim from other SLRs
2) They should be reviewed by an expert
3) They should be as specific as possible
4) They should be reviewed, and may need to be updated, during the search process

Regarding the paper selection process, the SLR Graduate Course provided the following observations:

1) Review of the titles and abstracts may not be sufficient for excluding papers
2) Citation management is needed to track references
3) Scripts and manual examination are needed to handle identify and eliminate duplicate results
4) Certain terms or content may not be evident in the title/abstract/keywords, but a quick scan of a paper's full text can quickly eliminate some irrelevant papers
5) This step is very time consuming, particularly completing multiple reviews of papers

Fig. 4. Method of Selecting Papers



6) Coordinating team members' schedules is difficult

Questions 12-14 of the SLR author survey asked the respondents to indicate the process they followed for applying the inclusion/exclusion criteria at the Title, Abstract and Full-Paper levels respectively. The options were:

1) One researcher alone
2) One researcher reviews all, another reviews a sample
3) Two or more researchers review all papers
4) Other

The results in Figure 4 show that the most common method of selecting papers is to have a primary researcher review all papers with a second researcher reviewing only a sample. The second most popular approach was to have multiple researchers review each paper. These results indicate that performing an SLR is a highly-collaborative process that is difficult to carry out effectively as a single researcher. The collaborative nature of the process raises another issue. What happens when conflicts arise? The survey respondents indicated that most conflicts could be resovled through discussions or through consultation with another expert (who breaks the tie).

Suggestions for novices about Identification of Primary Studies relate to the need for:

- Expert review of inclusion/exclusion criteria
- Collaboration in reviewing and selecting the papers
- Strategies for team management and conflict resolution

D. P5/E3: Quality Assessment

After identifying a set of candidate primary studies, the SLR authors must perform a quality assessment of those studies. The results of the SLR survey indicated that 32.2% of the respondents did not perform a quality assessment. Table V shows the reasons given for the lack of quality assessment. These reasons relate to the lack of standards, resources, and time, as well as to the belief that it was not necessary due to the expertise of the authors.

For the 67.8% of the respondents who performed a quality assessment, the methods for defining the criteria were:

- Reused existing criteria — 27.5%
- Created new criteria — 17.5%
- Some combination — 52.5%
- Other — 2.5%

Survey respondents used the same methods for resolving conflicts about the quality assessment as they did for resolving conflicts about the inclusion of primary studies.

In terms of the actual exclusion of papers based upon the quality assessment, the survey respondents reported that 30% of the time no papers were excluded, 30% of the time less than 10% of the papers were excluded, 22.5% of the time between 10% and 25% of the papers were excluded, and 17.5% of the time greater than 25% of the papers were excluded.

The SLR Graduate Course also provided the following observations. First, the quality assessment checklist should be part of the data extraction form. Second, the quality criteria must be relevant to the specific topic and not simply reused from a prior SLR. Third, the researcher must ensure that the quality assessment criteria will actually differentiate between papers of different quality. Finally, the quality assessment should be performed by at least two researchers to avoid bias.

Suggestions for novices about Quality Assessment relate to the need to:

- Define appropriate quality assessment criteria
- Apply the criteria to eliminate low-quality studies
- Have multiple authors independently assess quality

E. P6/E4: Data Extraction

Once the SLR authors have arrived at a final set of papers that are to be included review, the next step is data extraction. The results of the SLR author survey showed that data extraction was the most difficult and most time consuming aspect of the SLR process. The survey respondents indicated that changes to the data extraction procedures or the data extraction form were reasons to deviate from the defined protocol.

Similar to the question about how the inclusion/exclusion criteria was applied, the survey asked the respondents how they performed data extraction. The results showed that 45.8% of the time one researcher extracted data from all papers and a second researcher reviewed all extracted data, 30.5% of the time two or more researchers independently extracted data from each paper, 10.2% of the time one researcher extracted data from all papers and a second researcher reviewed a sample, and 13.6% of the time another process was used. Survey respondents used the same methods for resolving conflicts about the data extraction as they did for resolving conflicts about the inclusion of primary studies.

The SLR Graduate Course resulted in the following observations. First, the data extraction form needs to be reviewed by an expert in SLRs and an expert in the domain of the review. Second, the data extraction form may need to be refined throughout the data extraction process as the authors better understand the papers. Third, the extracted information needs to be reviewed by collaborators to eliminate any bias. Finally, there is a need for a tool that allows the extracted data to be easily recorded and analyzed across multiple papers. In our survey, most researchers used spreadsheet software such as Microsoft Excel to support data extraction.

Suggestions for novices about Data Extraction relate to the need to:

- Have an expert review the data extraction form
- Support data extraction and review by multiple authors
- Ease recording and analysis of the extracted data

VI. Discussion and Limitations

The results indicate a number of barriers for both novice and experienced SLR researchers. While many of these barriers can be overcome with additional experience and/or guidance, some are simply a matter of time and effort — factors that are easily underestimated. Overall, our results indicate that the planning phase of a review is perhaps the most critical part of the process.

From a **planning perspective**, the preparation and review of the protocol is crucial to success. Individuals new to the SLR process benefit from working through the planning details and the feedback of the review committee. The process helps establish more realistic expectations as to the time and effort that will be required to complete the review. Experienced researchers may also benefit from the review process as indicated by our survey results, where a large percentage of the protocols that were reviewed ended up being modified. The addition of a "second set of eyes" often leads to improvements in the search process, selection and evaluation criteria, and data extraction processes. These small changes, in our experience, may lead to substantial reductions in both the time and effort required during the conducting phase of a SLR.

A critical first step in the planning phase is the **identification of appropriate research questions**. The research questions serve as the foundation of the review and drive the search strategy, inclusion/exclusion criteria, data extraction, and synthesis of the review. Research questions must be properly scoped to avoid the two problems of data explosions and insufficient data for a complete analysis. Expert feedback from both a domain perspective and a research perspective is beneficial to formulating appropriate research questions. These conclusions are similar to those drawn by Riaz et al. specifically about novice researchers [20].

Our results indicate a number of issues surrounding the **search process** for the review. First, there are numerous sources — both traditional and electronic — from which existing research may be drawn. Identifying appropriate sources appears, at this point, to be a matter of experience. For the novice researcher this means relying on the feedback of those experienced in the domain for guidance in planning.

Similarly, the **keywords** and their appropriate combinations for search purposes are a challenging aspect of SLR planning. The difficulties associated with defining search strings is further compounded by the nuances of the individual data stores. At present, these issues can only be overcome through experience. This experience may come in the form of feedback from a review committee or in the form of pilot testing with the individual data sources.

TABLE IV. COMMON DATABASES ISSUES

Category	Nature of Issue	# Cases
Consistency	Inconsistency of Data among Databases	9
	Indexing Problems	5
	Coverage of the Database Is Unclear	4
	Duplicity of References	4
	Lack of Confidence in Results	2
Interface	Retrieve and Manage Papers	5
	Download Large Number of PDF Files	3
	Format of the Results	3
	Limited Access to Some Papers	1
	Multiple Interfaces	1
	Different Filtering Capabilities	1
	Tool Synchronization Problems	1
	Not Allowed to Crawl the Results	1
Studies Content	Abstract Is Unclear	1
Volume	Volume of Results	1

TABLE V. COMMON REASONS NOT TO EVALUATE THE PAPERS BY QUALITY CRITERIA

Category	Reason Given	# Cases
Standards/Resources	Papers Do Not Fit in the Standards of Quality Assessment	4
	Have Not Found How	3
	Quality Assessment Is Out of Scope	1
	Insufficient Resources	1
Time	Large Volume of Papers	2
	Insufficient Time	1
Unnecessary	Quality Is Subjective	1
	Papers Already Evaluated by the Community	1
	Performing a Systematic Mapping	1
	Search Is Exploratory	1
	Did Not Seem Necessary	1
	Assessment in Another Way	1

TABLE VI. HOW DID YOUR PROTOCOL DEVIATE FROM THE STANDARD SLR PROCEDURE AND WHY?

Category	Reason for Deviation	# Cases
Planning	Some Researchers Have Followed Other Guidelines	2
	Introduce Iterative Process	1
	Unable to Report Every Step of the Protocol	1
	Used PICOC	1
Study Identification	Search Procedures Changed	4
	Added Manual Search	1
Selection	Relaxed Quality Assessment Compared to Guidelines	1
	Two Researchers Did Full Independent Selection Process	1
	Added Relevance Assessment Step	1
	Used Advanced Strategies for Study Selection	1
	Meetings to Resolve Differences in Selection	1
	Assessed Quality Based on Rigor and Relevance	1
	Did Not Measure the Degree of Agreement Among Raters Quantitatively	1
Extraction	Clustered Structured the Area of Research Using Bubble Plot	1
	Conducted an Expert Survey to Verify the Research Issues Extracted	1

In any case, researchers must not only plan the search itself, but must also plan for the **storage and manipulation of the search results**. We noted a number of bibliographic management tools which can aid in this process, some of which include the vital ability to eliminate duplicate search results. The two most commonly used tools were BibTex and EndNote.

The identification of clearly defined **inclusion and exclusion criteria** is another critical step in planning a review for both the experienced and the novice researcher. Clearly defined criteria simplifies the identification of studies to be included in the analysis. Filtering of the search results is one of the more time-consuming aspects of executing a review. It is made even more challenging when the titles and abstracts are not clearly reflective of the content of a paper. However, the barriers of time and effort to complete this stage can be reduced through careful planning, the incorporation of feedback from a protocol review, and the use of appropriate tools designed to aid in bibliographic management.

The **quality assessment** of the studies selected for inclusion in the review is also a very challenging aspect of the SLR process. It is both time and effort intensive in that it requires multiple researchers to review and assess each of the selected studies. The original SLR process concept as utilized in the field of medicine focused primarily on the synthesis of randomized controlled trials. While more recent guidance concerning SLRs in the field of software engineering (SE) [14] recognizes the variety of methodologies utilized in the domain of SE and that the original hierarchy of evidence may not be appropriate, there is still a lack of clear consensus as to how the quality of different methodologies should be assessed. Our results suggest researchers attempt to overcome this hurdle by utilizing previously published quality criteria with little, if any, refinement. Interestingly, Dybå and Dingsøyr described a series of quality assessment systems used in other disciplines that could be ported to SE [9]. In addition there are some quality assessment tools used by medical researchers in performing SLRs which may be of use to the SE community [12]. The need for these validated quality criteria is an open question and poses a need for further research within the SE community.

Finally, we note that the **data extraction** process may seem deceptively simple, but in actuality may represent a significant hurdle. In both our data sources we found the data extraction process to be a common source of deviation from the defined protocol. While our results indicate that better planning of the data to be extracted as well as improvements in data extraction form design may help alleviate some of the issues, further research is needed to fully understand the difficulties associated with this portion of the SLR process. Based on the results found in both the survey and course feedback, it is clear that planning for data extraction must include not only what is to be extracted and how, but also consideration for the storage, manipulation, and tracking of the extracted data. Protocol planning should therefore include a data management plan which is then followed by all researchers participating in the review.

We note several limitations for the interpretation of our results. First, our sample of novice researchers (i.e., PhD students enrolled in the SLR Graduate Course) is small and consists of individuals that are not only new to the SLR process, but also new to research in general. However, the diversity of the group and consistency of their responses alleviate this concern to some degree. Individuals ranged in age from mid-20s to late 40s, came from a variety of ethnic and cultural backgrounds, were in the third or fourth year in doctoral programs from two disciplines (CS and MIS), and had a variety of industry experience. We must also consider that the novice group is drawn from one course. This raises a concern that the results from the group may in fact be an artifact of the instruction in that course. We believe this concern is offset by the similarity of findings in the survey results and previous experiences delivering the course.

VII. Implications and Conclusions

"An ounce of prevention is worth a pound of cure."
— Anonymous

From a teaching perspective, the results of this study indicate several key success factors to be considered in the design and execution of a course of instruction over a single semester. First, an emphasis must be placed on the identification and proper scoping the research questions that will drive the study. For the novice researcher working within the time constraints of a single semester, this is a critical first hurdle with consequences that ripple throughout the remainder of the SLR process. A similar observation was noted by Oates and Capper who indicated that novices need additional guidance on the construction of the research questions [18].

Similarly, the construction and subsequent revision of the protocol for a study has far-reaching implications for both novice and expert researchers alike. Many of the barriers identified during the execution phase of an SLR study can be overcome, or significantly reduced, through extensive planning and pilot testing. For novice researchers, shorter iterations with feedback over individual protocol items may assist in driving progress towards a completed and feasible protocol. While our analysis supports an emphasis on the planning phase of the SLR process, additional research is needed to fully understand the impact which it has regarding the conservation of time and effort in the conducting phase of a review. Future research should also examine the documenting phase of the review and the impact of the planning and conducting phases on the documentation process.

We have also identified four primary areas in need of infrastructure support. Each of these areas represent opportunities for improvements in existing support infrastructure for researchers and possible areas of exploration for future research.

First, the process of systematically identifying relevant papers is largely manual and thus very labor intensive. This process is more difficult when research topics cross traditional disciplinary boundaries, as many interesting topics increasingly do. While some advancement has been made in database functionality, as a discipline, we still lack adequate tools to assist in the extraction and compilation of relevant information from existing research. Using most common search engines for SE literature (e.g., IEEExplore, ACM, or Google Scholar), a search may yield thousands of results, of which a large percentage may be irrelevant for various reasons (i.e., overloaded terminology or containing the right word combinations by chance). In addition, because databases are not mutually-

exclusive, the result set will likely contain duplicates, which requires manual intervention by a researcher for resolution.

Second, tool support is lacking for collaborative SLRs. After identifying the relevant set of articles, multiple researchers must extract important information from each paper and compare the results for consistency. Again, this step is typically performed manually. There is no tool support designed specifically to ease this step or to in the inter-rater reliability assessment necessary to evaluate the accuracy of the extracted information.

Third, there is no mechanism to store the data extracted from papers so that it can be updated and reused. It is quite likely that the same paper may be relevant to multiple SLRs. While the data extracted from a paper may differ somewhat depending on the research questions, there will likely be a lot of common data items. Because there is no central repository for storage of extracted data, researchers must fully repeat this extraction process for each new SLR. Such a repository would not only reduce effort by enabling a researcher to extract only the additional data relevant to the new research question(s), but also facilitate collaboration by allowing researchers to identify others working on similar topics.

Finally, there is no mechanism to enable SLRs to evolve over time. Ideally, an SLR should be a "living" document that could evolve as new research results become available. Because current publications are static, appropriate infrastructure is needed to support the evolution of SLR results by allowing researchers to easily create new versions or fork off related topics. Making SLRs living documents that incorporate the latest research results will allow them to be more useful both to researchers and practitioners.

An SLR may serve as the foundation for new research or as a way to summarize an existing body of literature. As many SLRs are performed by PhD students conducting background research, it is important to understand the common difficulties and barriers to conducting a successful review. This study has identified a number of such barriers and found that most are not unique to the novice, but rather are shared by experienced researchers as well. By better understanding the difficulties and barriers in the SLR process, we hope that the research community will be able to improve the instructional process, infrastructure needed to support the production of SLRs, and by extension both the quantity and quality of published SLRs.

REFERENCES

[1] M. S. Ali, M. A. Babar, L. Chen, and K.-J. Stol, "A systematic review of comparative evidence of aspect-oriented programming," *Information and Software Technology*, vol. 52, no. 9, pp. 871–887, Sep. 2010.

[2] M. A. Babar and H. Zhang, "Systematic literature reviews in software engineering: Preliminary results from interviews with researchers," in *Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 346–355.

[3] L. Chen, M. A. Babar, and C. Cawley, "A status report on the evaluation of variability management approaches," in *13th Int'l Conf. on Evaluation and Assessment in Soft. Eng.*, 2009.

[4] L. Chen, M. A. Babar, and H. Zhang, "Towards an evidence-based understanding of electronic data sources," in *Proceedings of the 14th international conference on Evaluation and Assessment in Software Engineering*. Swinton, UK, UK: British Computer Society, 2010, pp. 135–138.

[5] L. Chen and M. A. Babar, "A systematic review of evaluation of variability management approaches in software product lines," *Information and Software Technology*, vol. 53, no. 4, pp. 344–362, Apr. 2011.

[6] F. Q. B. da Silva, A. L. M. Santos, S. C. B. Soares, A. C. C. França, and C. V. F. Monteiro, "A critical appraisal of systematic reviews in software engineering from the perspective of the research questions asked in the reviews," in *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. New York, NY, USA: ACM, 2010, pp. 33:1–33:4.

[7] I. F. da Silva, P. A. da Mota Silveira Neto, P. O'Leary, E. S. de Almeida, and S. R. de Lemos Meira, "Agile software product lines: a systematic mapping study," *Software: Practice and Experience*, vol. 41, no. 8, pp. 899–920, 2011.

[8] E. Dominguez, B. Perez, A. L. Rubio, and M. A. Zapata, "A systematic review of code generation proposals from state machine specifications," *Information and Software Technology*, vol. 54, no. 10, pp. 1045–1066, Oct. 2012.

[9] T. Dybå and T. Dingsøyr, "Strength of evidence in systematic reviews in software engineering," in *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement*. New York, NY, USA: ACM, 2008, pp. 178–187.

[10] S. Jalali and C. Wohlin, "Systematic literature studies: database searches vs. backward snowballing," in *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*. New York, NY, USA: ACM, 2012, pp. 29–38.

[11] A. Kakar and J. Carver, "Best practices for managing the fuzzy front-end of software development (sd): Insights from a systematic review of new product development (npd) literature," in *Proceedings of International Research Workshop on IT Project Management*, 2012.

[12] S. Karunananthan, C. Wolfson, H. Bergman, F. Beland, and D. Hogan, "A multidisciplinary systematic literature review on frailty: Overview of the methodology used by the canadian initiative on frailty and aging," *BMC Medical Research Methodology*, vol. 9, no. 1, p. 68, 2009.

[13] B. Kitchenham, "Procedures for performing systematic reviews," Keele University, Keele, UK, Joint Technical Report TR/SE-0401 and NICTA 0400011T.1, 2004.

[14] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keel University, EBSE Technical Report EBSE-2007-01, 2007.

[15] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering âĂŞ a systematic literature review," *Information and Software Technology*, vol. 51, no. 1, pp. 7–15, Jan. 2009.

[16] E. Mendes, "A systematic review of web engineering research," in *International Symposium on Empirical Software Engineering*. Institute of Electrical and Electronics Engineers, 2005, pp. 481–490.

[17] S. Montagud, S. Abrahao, and E. Insfran, "A systematic review of quality attributes and measures for software product lines," *Software Quality Journal*, vol. 20, no. 3-4, pp. 425–486, Sep. 2012.

[18] B. J. Oates and G. Capper, "Using systematic reviews and evidence-based software engineering with masters students," in *Proceedings of the 13th international conference on Evaluation and Assessment in Software Engineering*. Swinton, UK, UK: British Computer Society, 2009, pp. 79–87.

[19] M. Riaz, E. Mendes, and E. Tempero, "A systematic review of software maintainability prediction and metrics," in *3rd International Symposium on Empirical Software Engineering and Measurement*. Institute of Electrical and Electronics Engineers, Oct. 2009, pp. 367–377.

[20] M. Riaz, M. Sulayman, N. Salleh, and E. Mendes, "Experiences conducting systematic reviews from novices' perspective," in *Proceedings of the 14th international conference on Evaluation and Assessment in Software Engineering*. Swinton, UK, UK: British Computer Society, 2010, pp. 44–53.

[21] A. Strauss and J. Corbin, *Basics of qualitative research: Grounded theory procedures and techniques*, A. Strauss and J. Corbin, Eds. Sage, 1990, vol. 2nd.

[22] S. Thompson, M. Keith, and C. Posey, "Putting privacy in its place: A taxonomy of the costs and benefits of location data disclosure," in *Proceedings of The Dewald Roode Information Security Workshop*, 2012.

[23] H. Zhang and M. Ali Babar, "On searching relevant studies in software engineering," in *Proceedings of the 14th international conference on Evaluation and Assessment in Software Engineering*. Swinton, UK, UK: British Computer Society, 2010, pp. 111–120.