Systematic Literature Studies: Database Searches vs. Backward Snowballing

Samireh Jalali
Blekinge Institute of Technology
SE 37179 Karlskrona, Sweden
samireh.jalali@bth.se

Claes Wohlin
Blekinge Institute of Technology
SE 37179 Karlskrona, Sweden
claes.wohlin@bth.se

ABSTRACT

Systematic studies of the literature can be done in different ways. In particular, different guidelines propose different first steps in their recommendations, e.g. start with search strings in different databases or start with the reference lists of a starting set of papers.

In software engineering, the main recommended first step is using search strings in a number of databases, while in information systems, snowballing has been recommended as the first step. This paper compares the two different search approaches for conducting literature review studies.

The comparison is conducted by searching for articles addressing "Agile practices in global software engineering". The focus of the paper is on evaluating the two different search approaches.

Despite the differences in the included papers, the conclusions and the patterns found in both studies are quite similar. The strengths and weaknesses of each first step are discussed separately and in comparison with each other.

It is concluded that none of the first steps is outperforming the other, and the choice of guideline to follow, and hence the first step, may be context-specific, i.e. depending on the area of study.

Categories and Subject Descriptors

D.2 [**Software Engineering**]: Management—*Software process models*; K.6 [**Management of Computing and Information Systems**]: Software Management—*Software process*

General Terms

Experimentation, Measurement

Keywords

Systematic Literature Review, Snowballing, Agile Practices, Global Software Engineering

ESEM'12, September 19–20, 2012, Lund, Sweden.

1. INTRODUCTION

Research literature may be divided into primary studies (new studies on a specific topic) or secondary studies (summarizing or synthesizing the current state of research on a specific topic). The secondary studies may be used to pinpoint gaps or to highlight areas that require more attention from researchers or practitioners.

Secondary studies require comprehensive searches in the published research literature. Kitchenham and Charters [11] proposed a systematic literature review (SLR) approach inspired by evidence-based medicine, which recommend starting with systematic searches in databases using well-defined search strings to find relevant literature. In the guidelines [11], it is recommended that snowballing from reference lists of the identified articles should be used in addition to the searches in the databases, i.e. to identify additional relevant articles through the reference lists of the articles found using the search strings.

However, the guidelines do not explicitly recommend forward snowballing, i.e. identifying articles that have cited the articles found in the search and backward snowballing (from the reference lists). In our experience, most systematic literature reviews (including our own) do not use snowballing as a complement to searching the databases. It is fully understandable given the amount of work needed to conduct a systematic literature review. The implication being that a review provides a limited set of all papers on the topic, i.e. a sample of the population.

Webster and Watson [4] proposed a slightly different approach to systematic literature studies in the field of information systems. They propose to use snowballing as the main method to find relevant literature. In their recommendation, they highlight both backward snowballing (from the reference lists) and forward snowballing (finding citations to the papers). The snowballing approach requires a starting set of papers, which they suggest should be based on identifying a set of papers from leading journals in the area.

Given that there exist different guidelines of how to conduct systematic literature review studies, we pose the following research questions:

1. To what extent do we find the same research papers using two different review approaches?
2. To what extent do we come to the same conclusions using two different review approaches?

The outcome of a systematic literature study is either a systematic literature review [11] or a systematic mapping study [20]. Database searches and snowballing are by no

means the only options. The use of personal knowledge or contacts [6], or mixed methods [3] has also been discussed in the literature. The focus here is, however, on the first step of two recommended methods to identify relevant literature. Thus, we have limited our study to using either database searches or backward snowballing as the first step, in particular given that, in our experience, researchers are, all to often, forced to limit their search procedures given the time it takes to conduct a systematic literature study. However, we believe that if similar patterns are identified through applying partial methods (i.e. backward snowballing), the similarity is expected to increase if forward snowballing is also performed since the overlap in the included papers would be greater. The papers found are evaluated for relevance and quality, which gives a set of primary studies for each search approach (database search or backward snowballing), and these papers are the basis for further comparisons conducted in this paper.

Given that the number of published secondary studies increases [18], it is perceived as important to understand whether or not the first step in the searches impacts the actual outcomes of the systematic literature study, in particular since many published papers do not use all steps recommended in the guidelines. This is closely related to the need to ensure reliability of secondary studies, which means whether two independent studies on the same topic would find the same set of papers and draw the same conclusions [12].

Based on the need identified, we conducted two different literature reviews on Agile practices in Global Software Engineering (GSE) using different guidelines for the literature search, and in particular we only used the first step in the recommendation, i.e. database searches [11] or backward snowballing [4]. It should be noted that we included distributed development within a country in GSE too. The main reason being that many of the challenges experienced in a global setting also occurs in distributed development within a country, although some of the challenges are amplified when going global. Both studies have the same research questions. The first study is an SLR [9], and the second study applied a snowballing approach [4]. The differences between the search methods are discussed in more detail in Section 3.

The remainder of the paper is structured as follows. Section 2 summarizes related work, and Section 3 discusses the research method and introduces the two studies forming the input to the analysis. The results are presented in Section 4, and the discussions of the findings are given in Section 5. Finally, Section 6 presents the conclusions and the future research directions.

2. RELATED WORK

Inspired from medicine, in which systematic literature reviews is an approach for synthesizing evidence, Kitchenham et al. [7] introduced the concept of evidence-based software engineering (EBSE). A couple of years earlier Webster and Watson [4] suggested a structured approach in information systems to conduct systematic literature studies.

It should be noted that the research type in medicine and software engineering (SE) are not necessarily the same (e.g. controlled experiments vs. case studies). It implies different types of data, and different types of analyses of the data. Hence, the synthesis of the data collected from an SLR in SE may not be as straightforward as in at least some parts of medicine.

However, a practitioner-oriented view was formulated based on the EBSE ideas [8], and researchers also suggested guidelines for conducting systematic literature reviews [11]. Furthermore, Brereton et al. [14] reviewed a number of existing literature reviews to examine the applicability of SLR practices to SE. They found out that although the basic steps in the SLR process are as relevant in SE as in medicine, some modifications are necessary for example in reporting of empirical studies in SE.

Although the number of literature review studies in SE has increased in the past five years [18], few studies exist which evaluate the reliability of literature search approaches for example to evaluate the repeatability of protocol-driven methods or to compare the results of literature searches conducted through different methods such as SLR and snowballing. In the following, we summarize the relevant research.

Greenhalgh and Peacock [6] conducted a study in order to describe where papers come from in a systematic review of complex evidence. They applied three different methods and found 495 primary sources related to "therapeutic interventions". Their conclusion was that protocol-driven search strategies by themselves are not the most efficient method regardless of the number of traversed databases, because some sources may be found through personal knowledge / contacts (e.g. browsing library shelves, asking colleagues), and snowballing is the best approach for identifying sources published in obscure journals.

In 2009, Skoglund and Runeson [15] investigated a reference-based search approach with the primary purpose of reducing the number of initial articles found in SLRs. Although the proposed method increased the precision without missing too many relevant papers for the technically focused reviews, its results were not satisfactory when the search area was wide or the searches included general terms. This implies that the choice of approach to searching is context-dependent.

Zhang et al. [13] conducted two participant-observer case studies to propose an effective way of identifying relevant papers in SLRs. The approach was based on the concept of quasi-gold standard for retrieving and identifying relevant studies, and it was concluded to serve the purpose and hence it can be used as a supplement to the guidelines for SLRs in EBSE. In a follow up validation study [3], a dual-case study was performed, and the proposed approach seemed to be more efficient than the EBSE process in capturing relevant studies and in saving reviewers' time. Further, the authors recommended an integrated search strategy to avoid limitations of applying a manual strategy or an automated search strategy.

MacDonell et al. [12] evaluated the reliability of systematic reviews through comparing the results of two studies with a common research question performed by two independent groups of researchers. In their case, the SLR seemed to be robust to differences in process and people, and it produced stable outcomes.

Kitchenham et al. [16] conducted a participant-observer multi-case study to investigate the repeatability of SLRs performed independently by two novice researchers. However, they did not find any indication of repeatability of such studies that are run by novice researchers.

In summary, too few studies have addressed the reliability of secondary studies. As discussed here, they have either compared different SLRs or mapping studies to check whether the same results are achieved [12] and [16], or investigated more efficient approaches of searching [6], [15] and [13]. As a complement to previous studies, we investigate the reliability of secondary studies using different search strategies. This is done by comparing the outcome of two studies on the same topic using different guidelines for finding the relevant literature. The research method used is discussed next.

3. RESEARCH METHOD

The main objective of this study is to examine whether two systematic review studies would provide the same result when the applied first step in the search strategy is different. Therefore, we planned two separate literature reviews.

The first study was conducted within 2009-2010 to capture relevant research about the most common Agile practices applied in different settings of global software engineering [9]. The second study was performed during 2010-2011 with exactly the same purpose and the same research questions [10]. The difference between the two studies was the way that the relevant papers and articles were extracted from the published research, i.e. the search strategy. The time between the two searches was a couple of months and the time between the syntheses was around eight months, and hence the details about specific papers found in the first search were not fresh in the mind of the researchers. Thus making the searches reasonably independent. An alternative would have been to have different researchers conducting the two studies. However, this would have introduced threats in relation to judgments of inclusion and exclusion of papers. The threat of having the same researchers involved was mitigated by time, i.e. by leaving several months between the two studies the researchers did not remember all the details of individual papers.

The first study **(S1)** follows the guidelines provided by Kitchenham and Charters [11] as far as it comes to conducting searches in the databases. S1 did not use snowballing from reference lists as recommended in the guidelines. In the second study **(S2)**, a backward snowballing [4] approach was used. The starting set of papers for the snowballing approach was generated through a search in Google Scholar[1] on peer-reviewed papers published in 2009 rather than using our knowledge of relevant papers gained during S1. However, the purpose was to avoid the bias at this stage. This is further elaborated below.

The snowballing search method [4] can be summarized in three steps: 1) Start the searches in the leading journals and / or the conference proceedings to get a starting set of papers, 2) Go backward by reviewing the reference lists of the relevant articles found in step 1 and step 2 (iterate until no new papers are identified), and 3) Go forward by identifying articles citing the articles identified in the previous steps. Based on that S1 was focused on a specific time period, i.e. 1999-2009, it was decided to identify a starting set of papers from 2009 and then use backward snowballing based on the papers found. Given that researchers seem to focus on the database search, despite the guidelines [11], it was decided to only compare the first step for the searches, i.e. the database search vs. the backward snowballing approach. It was done for two reasons:

1. It would make the systematic literature review using the guidelines more representative of the state of studies actually published. As a consequence we saw a need to not follow the guidelines [4] for snowballing perfectly either. Thus, trying to be as fair as possible in the comparison. It would have been unfair to follow the guidelines very closely in one case and then not in the other case.

2. It was realized that a more comprehensive use of the guidelines, i.e. following all steps recommended, would result in the outcomes getting closer to each other. With the first step, we refer to only doing the database search. However, for the papers found in the database search, we check relevance and quality to finally have a set of primary studies. The same procedure is done based on the other guideline [4], i.e. we only perform backward snowballing and identify a primary set of papers. Having done all steps recommended in the guidelines would undoubtedly mean that more papers would be included and if the searches being perfect they would end up with exactly the same set of papers. Thus, we wanted to compare the first steps in the different guidelines, since it is reasonable to believe that if these produce similar enough results, then a larger sample of papers would just increase the similarity. Thus, we are concerned with comparing the samples of papers obtained when conducting the first steps in the two different guidelines [11] and [4] respectively.

Furthermore, to make the studies as comparable as possible, we kept the search terms and keywords as similar as possible in both studies and also applied the same constraints on searches. This means that the same search terms were used in the database searches in S1 as in the Google Scholar search in S2. In addition, the same researchers were responsible for finding, evaluating, and analyzing the relevant papers in both studies in order to minimize the diversity in data collection and data analysis. Hence, the only (intended) difference between S1 and S2 is the search approach (the way we identified the relevant papers).

The assessment is performed through comparing the results of the two studies based on their primary papers and their conclusions. In summary, the research questions are:

- RQ1. To what extent do we find the same research papers using two different review approaches?
- RQ2. To what extent do we come to the same conclusions using two different review approaches?

In order to answer the research questions, we conducted an in-depth comparison of the two studies.

3.1 Details of Studies

S1: It was designed to be a systematic literature review following the guidelines by Kitchenham and Charters [11], although only doing the database searches and not snowballing. The study was conducted during 2009-2010 with the purpose of capturing the status of combining Agility with GSE [9]. The results were limited to peer-reviewed conference papers and journal articles published in 1999-2009.

[1] http://scholar.google.com/intl/en/scholar/about.html

The final set of papers (81 distinctive papers) was synthesized by classifying them into different categories (e.g. publication year, contribution type, research method and Agile practices used in GSE). More details of the S1 can be found in [9] and [2].

S2: It had the same purpose and research questions as the first study, and was conducted after we were finished with S1 (2010-2011) [10]. In this study, we followed the guidelines provided by Webster and Watson [4] regarding identification of a starting set of papers followed by backward snowballing. We searched in Google Scholar (only once) using similar search terms as in S1, and then limiting the search to 2009 to identify a starting set of papers for the backward snowballing. The main purpose with the search in Google Scholar was to minimize the researcher's bias in relation to S1 since an alternative was to begin with a set of relevant papers identified through S1. First, we evaluated the relevancy of the papers and then went through the reference list of the relevant papers in order to find additional sources. The process was stopped when we could not add any further relevant papers published in the time period 1999-2009. The analysis of the data was kept as similar as possible to S1. Some further details of S2 can be found in [10], since our objective is not to present the individual studies as such; the focus is on comparing the outcome of the two different first steps for the searches based on guidelines by Kitchenham and Charters [11] and Webster and Watson [4] respectively.

It has to be noted that the same criteria for inclusion process were applied in both studies e.g. gray literature was excluded from further analysis.

3.2 Comparison Approaches

The comparison is done in two different ways. First, we examined all papers identified in S1 and S2 regarding the papers included and the findings. However, due to the fact that the majority of the articles were identical, it was not surprising that the conclusions and the findings would be also similar in both studies. Therefore, we conducted a second comparison in which we excluded the papers, which were in common for both studies, and performed the analyses solely on the unique papers for S1 and S2 respectively. Then, we compared the findings from the two analyses. The major purpose of comparisons was to investigate similarities and differences between the extracted data for the same variables (e.g. research type) in database searches (S1) and backward snowballing (S2).

4. RESULTS

In the following, we present the differences and similarities between the findings of S1 and S2 given the different first steps for the searches.

4.1 Number of Papers

The first comparison relates to the number of papers found in the two studies. S1 resulted in 534 papers being identified from the databases. 81 papers were initially judged to be relevant. Thus, the data analysis began with 81 papers, but some articles were excluded. Papers were excluded if the report was incomplete (e.g. the results were missing), or if it was exactly the same study as another one in the list (e.g. if an empirical study formed the basis for both a conference paper and an extension published in a journal).

Finally, 53 papers were included in data analysis. In S2, we found 109 papers initially. After an analysis of the relevance, we were left with 74 papers. At the end, 42 papers were included in the data analysis. Papers were removed based on the same criteria in both studies, and hence the main difference is the initial way of finding papers, i.e. the search strategy.

There is a huge difference between the numbers of papers we initially found (109 vs. 534), but it should be noted that we have checked the title of a lot of sources in snowballing too, i.e. when browsing the reference lists of the papers identified. The latter makes it hard to compare the numbers in the first step exactly. Nevertheless, 45 paper were the same in the initial set of papers identified. This overlap was surprisingly low (8% in S1 and 41% in S2). However, the situation changes when we look at the papers in the next step, i.e. those initially judged as relevant. In this step, 41 papers were identical, which should be compared with having 81 papers in S1 and 74 papers in S2 that indicates 51% and 55% overlap respectively. The final set of papers used for data extraction include 53 papers in S1 and 42 in S2 with 27 identical papers (51% overlap for S1 and 64% for S2) which is a slight majority of the identified papers between the two studies. Figure 1 visualizes the overlapping papers at the last stage and all stages are summarized in Table 1, where the unique papers in each study are shown separately as well as the papers in common. Discussions around the differences between the unique set of papers found in S1 and in S2 are provided in the Appendix.

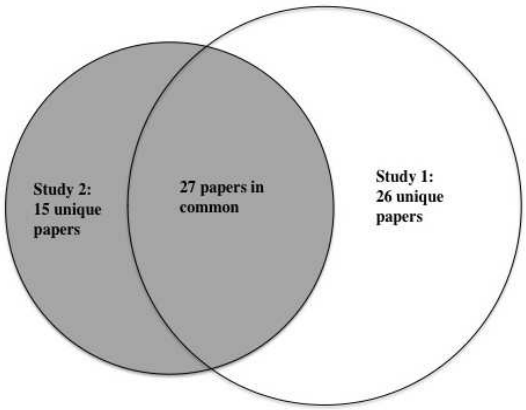

Figure 1: Venn Diagram for the Overlapping Papers

Table 1: Number of Papers in Two Different Studies

Study No.	Initial Papers	Relevant	Analyzed	Unique
1	489+45	40+41	26+27	26
2	64+45	33+41	15+27	15

The list of papers in common in S1 and S2 as well as list of unique papers for each separate study can be accessed in [19], in which the identical papers are denoted with M and are listed in Appendix A; the unique papers for S1 are denoted with D and are listed in Appendix B; and the unique papers for S2 are denoted with B and listed in Appendix C.

Table 4 in the Appendix summarizes the differences between unique set of papers found through database searches (S1) and backward snowballing (S2). For each paper, it represents: 1) the publication year, 2) the database that the

paper has been found in, 3) the terms in the title that belong to the first set of keywords, 4) the terms in the title that are in the second set of keywords, 5) if the paper is expected to be found in the other study, 6) and notes or comments. Symbol "?" in the table indicates that the information is not available.

As Table 4 shows, 10 papers in S1 do not include terms from both set of keywords in their *title* and hence cannot be found in S2 (represented by "N" in the "expected" column). Two papers (denoted by "M") could be possibly found in S2 since they include names of Agile practices in their title (i.e. "TDD" in [D19], "user stories and acceptance tests" in [D3]) and it is also evident from the title that they are within GSE. Thus, 14 papers (out of 26) should have been possible to find in S2. The reason for missing them in S2 might be either due to researchers' error or because they were not cited by other relevant papers found in the backward snowballing.

In the list of unique papers found through backward snowballing (S2), six papers (out of 15) should have been possible to find, although they were not (see Table 4 in the Appendix). Therefore, we checked the complete list of papers found initially in S1 (including 534 papers). Two out of six papers were excluded in S1 in the later data analyses because the research was already published in another paper found in S1. One of the original papers is missing in S2 while one is included. This indicates that the researchers have evaluated papers in S1 and S2 in a slightly different way. Six other studies are published in different databases than databases of S1 and hence cannot be found in S1.

Three papers ([B4], [B1] and [B7]) use a slightly different terminology and hence they are not within the set of defined keywords in S1 (see Table 4 in the Appendix). Therefore, if the same terms are used in the abstract, they cannot be found in S1. For example, a study such as [B4] uses "extremely" in the title and we could guess that it refers to "Extreme Programming". However, it is surprisingly not found in S1 although by reading the abstract in S2, we judged it as relevant. This might indicate the different ways used by the researchers and the search engines to interpret the terms in the text. In the case of [B1], "remotely located" is used in the title, which could not be found in the searches in databases since the equivalent term of "remote team" was formulated in S1. It was found through backward snowballing because only by seeing these words in the title of the article, we could immediately recognize that "remotely located" means global or distributed. The examples illustrate the challenges in formulating search strings while in snowballing it may become evident to the researcher to include a paper when reading the title of a paper.

4.2 Distribution of Papers

The next step in the comparison is to compare the distribution of papers across the years.

4.2.1 First Comparison

As mentioned above, the first comparison includes all papers found in both studies, while the second comparison (see below) compare the unique papers in each of the two studies. As shown in Table 2, the number of papers found in S1 and S2 in each year (1999-2009) is not the same.

However, the pattern of distribution does not seem to be completely different and both indicate that the number of papers has grown in the past decade.

Table 2: Number of Papers over Years

	Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
No. of	S1	0	0	0	1	2	10	6	12	14	20	17
Papers	S2	0	1	3	2	7	8	3	7	7	10	13

4.2.2 Second Comparison

The number of unique papers found by each study in each year is presented in Table 3. We have found no unique papers in S1 before 2004. Considering the number of papers in 2009, it is hard to conclude that the number of published papers is increasing in the past decade. In S2, the number of published papers seems to be constant over the years.

Table 3: Number of Papers over Years 2

	Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
No. of	S1	0	0	0	0	0	4	3	0	9	7	4
Papers	S2	0	1	2	0	2	2	2	1	1	1	3

However, we should mention that this comparison is not fully fair because the total number of papers shall be compared against each other instead of considering only the unique ones. The question is whether the differences are due to the different search strategies.

4.3 Distribution of Research Types

Next, we wanted to compare the research types using the classification from Wieringa et al. [17]. In summary, the types are defined as follows:

Evaluation Research: Techniques, methods, tools or other solutions are implemented and evaluated in practice, and the outcomes are investigated.

Validation Research: A novel solution is developed and evaluated in a laboratory setting.

Solution Proposal: A solution for a research problem is proposed, and the benefits are discussed, but not evaluated.

Conceptual Proposal or Philosophical Paper: It structures an area in the form of a taxonomy or conceptual framework, hence provides a new way of looking at existing things.

Experience Paper: It includes the experience of the author on what and how something happened in practice.

Opinion Paper: The personal opinion on a special matter is discussed in an opinion paper without relying on related work and research methodologies.

4.3.1 First Comparison

Both studies have found a majority of papers to be reported as experience reports in which practitioners have reported their own experiences on a specific issue and the method applied to alleviate it [17]. It should be noted that the number of papers in S1 and S2 for each research type is different. This difference is, however, expected due to the fact that the number of papers found in S1 and S2 are different. In addition, the order of research types according to their frequency is different. The order in S1 is: 1) experience report, 2) evaluation, 3) opinion, 4) solution, 5) validation and 6) philosophical, and in S2 it is 1) experience, 2) validation, 3) evaluation, 4) solution, 5) philosophical and 6) opinion. It is surprising that we found no opinion paper in S2 while it was the third most frequent research type in S1.

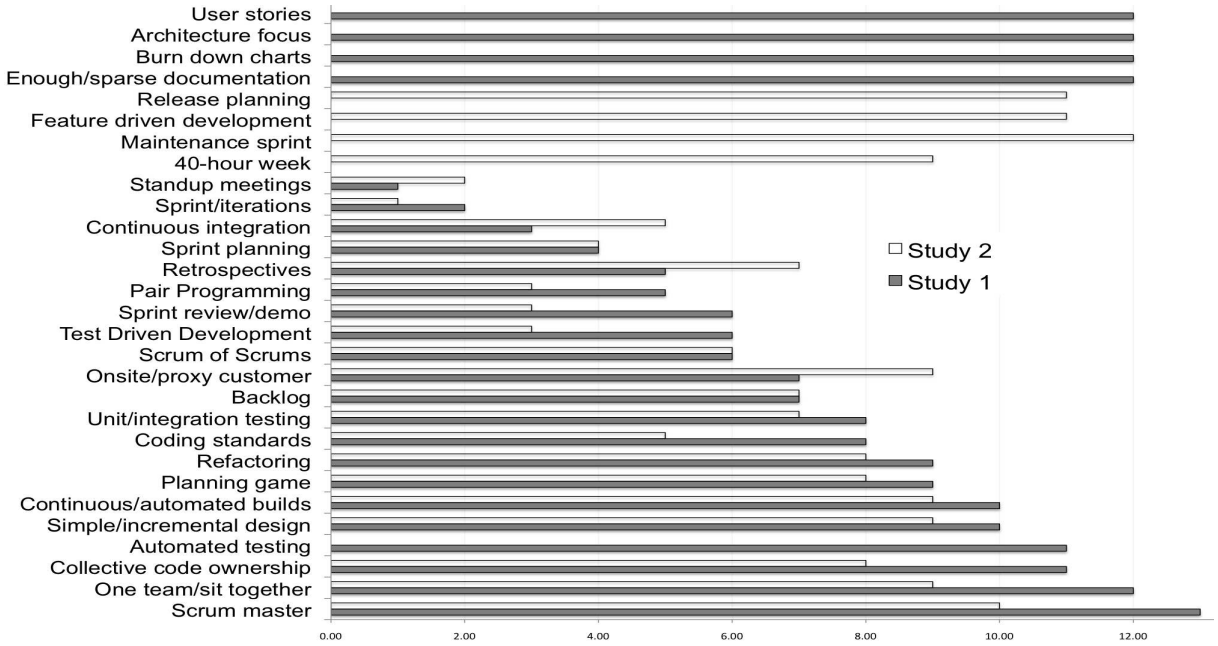

Figure 2: Agile Practices in S1 and S2 - Comparison 1

4.3.2 *Second Comparison*

When it comes to a comparison of the unique papers, the majority of the current research was found to be in form of experience reports in both studies. And in addition to the identified research types in S1, a solution paper is found in S2.

4.4 Countries Involved in GSE

It is also possible to compare the most common combinations of collaboration. The combinations include both global collaboration and distributed development.

4.4.1 *First Comparison*

In both studies, the collaboration between USA-India is found to be the most popular, and then distributed development within USA although the exact numbers are different.

4.4.2 *Second Comparison*

The same pattern as in the first comparison is found through the second comparison.

4.5 Most Efficient Practices

To identify the most efficient Agile practices used in a GSE was one of the main objectives of the literature review, and hence this is an important aspect to compare. If the studies identify completely different Agile practices, then the search strategy has indeed influenced the outcome of the literature review.

We would like to emphasize that it is not our intention to discuss the actual outcome in terms of which Agile practices are most efficient in a GSE setting. Our objective is to compare the outcomes from a search strategy point of view. Hence, the actual outcomes regarding Agility and GSE can be accessed in [9] and [10].

4.5.1 *First Comparison*

Considering the frequency of Agile practices in literature, we sorted the list of reported practices in both studies, where frequencies were counted based on the number of publications referring to a practice as being successful. We classified the practices based on their rank in a descending list. For example, if the highest frequency was found to be 18 for practice A, then practice A was assigned to class 1 together with all other practices having a frequency of 18. It means that the rank of each practice, in the sorted list, was considered as the class the practice belongs to. The practices with the same frequency have been assigned to the same class. The purpose of the classification was to be able to make a fair comparison, since the number of analyzed papers was different in S1 and S2. The result of the comparison is summarized in Figure 2 (the x-axis represents the classes for the practices).

If we take the 3 top classes of practices, S1 reported ”standup meetings” (class 1), ”sprint / iterations” (class 2), and ”continuous integration” (class 3). S2 found ”sprint / iterations” (class 1), ”standup meetings” (class 2), and three practices of ”pair programming”, ”sprint review / demo” and ”test driven development” in class 3. Thus, top three efficient Agile practices in S1 and S2 are overlapping (2 out of 3 = 66%).

According to Figure 2, general agreement of the most efficient practices is high (see classes 1 to 8). However, the higher classes (9 and higher) are infrequent, which means that these practices are mentioned by few studies and hence the difference is simply due to random, i.e. whether specific papers mentioning these practices are included or not.

4.5.2 *Second Comparison*

In this comparison, both studies found 18 practices, in which 13 of them were identical and 5 unique practices were found in each study (summarized in Figure 3).

Similar to Figure 2, the most frequent Agile practices ac-

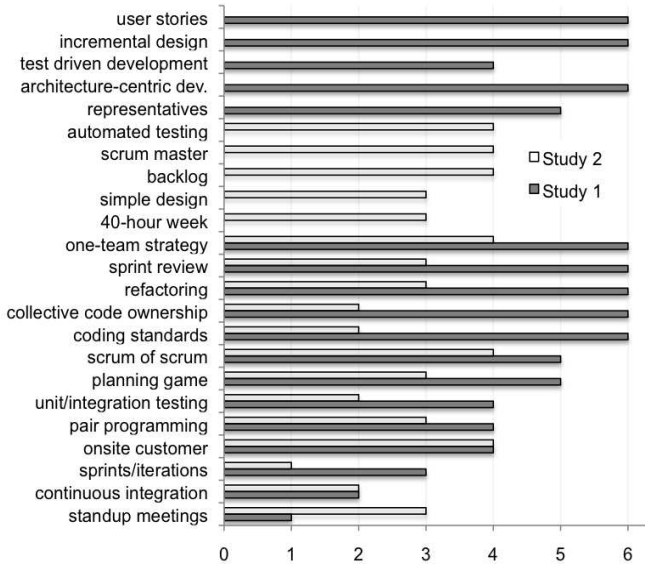

Figure 3: Practices in S1 and S2 - Comparison 2

cording to the literature (low numbers in Figure 3) show a very strong agreement.

4.6 Details for Agile - GSE Combinations

4.6.1 First Comparison

So far, we have presented and discussed which Agile practices were found most efficient in S1 and S2 as well as the countries involved in each combination of Agile method and distribution setting. For the purpose of comparison, we have assigned different scores to the combinations. If the practices and the countries found are the same for the Agile GSE combination in S1 and S2, we have assigned the score 4 to the combination; score 3 if only the practices are the same; score 2 if only the countries are identical; score 1 if neither practices nor countries are the same; and finally score 0 if the combination does not exist in the other study.

In this comparison, a majority of combinations are completely different. However, if we exclude the combinations, which were found in only one study (i.e. XP-open source in S2 and Agile-open source in S1), similar findings were identified for a majority of the combinations in both studies. Thus, it is clear that the comparison is very sensitive to individual studies.

4.6.2 Second Comparison

Unlike the previous comparison, most combinations seem to be completely different. However, it may be due to that the comparison is very sensitive to individual studies, and given that we have fewer studies here than in the first comparison, it may make it even more sensitive. It should be noted that the combinations of Agile with offshore, open source, and virtual team were found only in S1 whereas XP-open source, and XP-unclear were found only in S2. The latter combination means that the setting was unclear, although XP was used.

The higher the number shown in Figure 4 is, the higher is the similarity between the patterns in each comparison. Comparison 1 shows completely different pattern for 4 combinations out of 12, which indicates in 67% of the cases, the country involved, the Agile practices, or both have been the same for a specific combination of Agile and GSE.

In the second comparison, 7 patterns out of 11 are found to be completely different which is considerably different from results of comparison 1.

In the other words, we found exactly the same pattern (i.e. both Agile practices and the involved countries are the same in S1 and S2) for "three" distinctive combinations in comparison 1 and "one" combination in comparison 2. However, the number of combinations, which have the same score in both comparisons, is 6 out of 14, which implies 42% overlap. In addition, if we exclude the unique combinations, 6 out of 9 combinations have the same rank, which is 66% overlap.

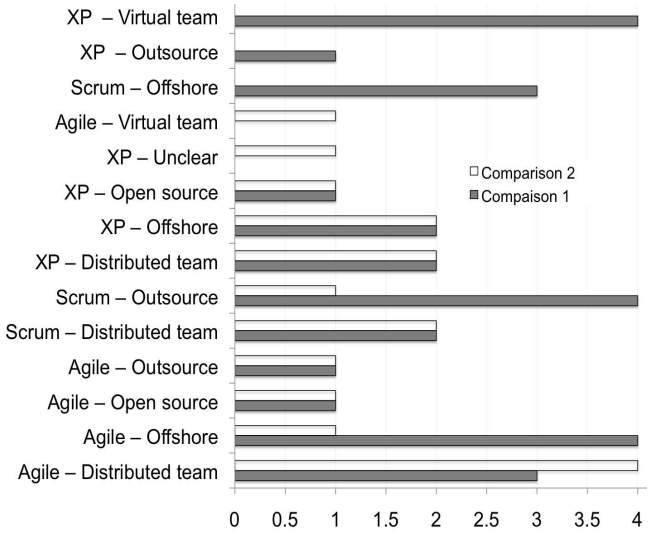

Figure 4: Patterns of Agile GSE Combinations

4.7 Limitations

In order to assure the reliability in this study, we tried to improve the reliability of the two systematic review studies in the first place (more details on this for each individual study can be found in [9] and [10]). Then considering the purpose of this study, which was to conduct a comparison, we tried to perform the comparison as fairly as possible.

Therefore, the researchers were the same in S1 and S2, and the analyses on the data were performed as similar as possible. Although we had more experiences of doing systematic reviews when we started S2, we tried to keep the gap between conducting searches as small as possible. So the data was collected with a few months difference for S1 and S2, although we synthesized them later for S2. The latter was done to ensure that the researchers did not remember all details when deciding on, for example, inclusion and exclusion of paper in S2. In addition, two researchers were involved in reviewing the comparisons and drawing the conclusions of this study.

In addition, the order of conducting the studies might have affected the results of each study and as a consequence the result of the comparisons.

5. DISCUSSION

5.1 Time and Effort Required

We cannot certainly claim that SLRs are more time consuming in formulating the search strings because snowballing requires some formulations too. However, SLRs require separate formulation for each database whilst snowballing does not explicitly require searching in more than one database. The number of initial papers in S1 was 534 and 109 in S2 which indicates greater time and effort spent to refine the searchers as well as identifying the relevant papers and discarding the irrelevant ones in the database searches approach (4.9 times more papers in S1 than S2).

5.2 Noise vs. Included Papers

Considering the term "Agile" which is a very general word and used in many papers in many disciplines, we found a lot of noise in S1. Due to this, we had to limit the search to abstract, title, and keywords. But still the number of irrelevant papers was much higher than the number of included and analyzed papers (85% noise). In the snowballing search, the balance seemed to be more reasonable (32% noise).

5.3 Judgments of Papers

In snowballing, most of the judgments were done based on the title of the paper when going backward through the reference lists (or forward in the citations if being applied). In some cases, we judged papers once more based on their abstract, i.e. it resulted in performing a stepwise judgment. In the SLR, the judgments were done on title and abstract at the same time. It should be noted that the papers with no relevant keywords in the title might be missed in snowballing. On the other hand, the papers that use different wordings (as in the example with cross-continent) might not be caught in an SLR.

5.4 Prior Experience

Prior experience of the researcher in the area of the studies as well as in performing the secondary studies may affect the results. The differences can be seen, for example, when judging the relevancy of the papers. An experienced person already knows several papers and knows several active researchers in the area, which may affect the reliability of the secondary study regarding the relevancy of included papers.

5.5 Ease of Use

We found the snowballing approach to be more understandable and easy to follow in particular it is believed to be easier for novice researchers. The SLR provides a lot of guidelines, which is good on one hand, but on the other hand novice researchers might find it confusing rather than helpful.

5.6 Identical Authors Risk

A potential threat in snowballing is that we might find several papers from the same authors since their previous research is usually relevant and is cited. Thus, the results of snowballing approach, might be biased by over presenting specific authors' research. On the other hand, database searches method performs searches on all papers in the database which eliminates this risk.

5.7 General Remark on Literature

It should be noted that a general problem with systematic literature reviews in software engineering is that in many cases existing papers are hard to classify and analyze since in many of the published studies, the contextual information is not well documented or the studies are not conducted in a realistic setting [1]. We have observed that insufficient contextual information hinders synthesizing the evidence from some studies (in particular industrial experience reports). Thus, we recommend practitioners and researchers working with industry to follow guidelines provided by Petersen and Wohlin [5] for documenting the contextual information.

6. CONCLUSIONS

In this paper, we evaluated two different first steps for conducting systematic literature studies. This was done by comparing two secondary studies on Agile practices in GSE, which were performed by the same researchers but using different search methods. First, we compared the studies against each other whether the same set of papers was found and if the included papers had resulted in the same conclusions. Secondly, we excluded the common papers from both studies and performed new analyses with the remaining unique papers for each study. Considering the fact that these comparisons did not indicate any remarkable differences between the two different studies, we compared the actual results found using the two different search methods applied. A summary of the findings is provided below. After comparing the two secondary studies in two different ways (with the common papers and with only the unique papers in each study), we did not find any major differences between the findings of the analyses. The figures and numbers were not the same, but the general interpretation of them is quite similar. We can summarize our findings as follows for the two research questions.

RQ1. To what extent do we find the same research papers using two different review approaches?
To answer the RQ1, we may observe that the papers found are different both in the number and the actual papers. In addition, the final set of papers used in data analyses, was also found to be different, although 27 papers were common. This is not really surprising given that we only used the first step in the two search methods, i.e. according to the different guidelines used. It is highly likely that the overlap would increase if we conducted snowballing from the papers found in the database searches, and also if we did forward snowballing when starting with backward snowballing. However, it is should be noted that a majority of the papers are the same despite only comparing the staring point for the comparison, i.e. database search vs. backward snowballing.

RQ2. To what extent do we come to the same conclusions using two different review approaches?
The answer to the RQ2 is more important, since it concerns the actual findings. Regardless of the differences in the actual numbers and figures, similar pattern were identified in both studies and hence similar conclusions were drawn. However, when excluding the same papers from both studies and analyzing only the remaining unique papers of each study, the identified patterns seem to be slightly different, which may be due to having fewer papers (a smaller sample). Therefore, it is not easy to draw any general conclusions with respect to the RQ2.

However, given the overlap, despite only conducting the first part in the guidelines, it indicates that the actual conclusions are at least not highly dependent on whether using database searches or snowballing. It is also quite obvious that the overlap will become larger if combining the two search strategies, although the downside being that it generates more work. Systematic literature studies are quite time consuming. Snowballing might be more efficient when the keywords for searching include general terms (e.g. Agile), because it dramatically reduces the amount of noise in database searches. Our personal experience confirms this. However, we recommend applying both backward and forward snowballing.

Although these conclusions, recommendations, and findings are based on our experiences with this comparison study as well as previous secondary studies, they seem to be in alignment with some previous studies [6] and [15], but contradictory with some others [14]. In anyway, more such comparison studies are required to be able to compare the methods fairly.

7. ACKNOWLEDGMENTS

This research was funded by the Industrial Excellence Center EASE - Embedded Applications Software Engineering, (http://ease.cs.lth.se).

8. REFERENCES

[1] M. Ivarsson, T. Gorschek (2011): A method for Evaluating Rigor and Industrial Relevance of Technology Evaluations. Empirical Software Engineering 16(3): 365-395.

[2] S. Jalali, C. Wohlin (2010): Agile Practices in Global Software Engineering - a Systematic Map. In 5th IEEE International Conference on Global Software Engineering, Princeton, USA, pp. 45-54.

[3] H. Zhang, M. A. Babar, X. Bai, J. Li, L. Huang (2011): An Empirical Assessment of a Systematic Search Process for Systematic Reviews. In the Proceedings of the 15th International Conference on Evaluation and Assessment in Software Engineering, pp. 56-65.

[4] J. Webster, R. T. Watson (2002.): Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly 26(2): xiii-xxiii.

[5] K. Petersen, C. Wohlin (2009): Context in Industrial Software Engineering Research. 3rd International Symposium on Empirical Software Engineering and Measurement, pp. 401-404.

[6] T. Greenhalgh, R. Peacock (2005): Effectiveness and Efficiency of Search Methods in Systematic Reviews of Complex Evidence: Audit of Primary Sources. BMJ 331(7524): 1064-1065.

[7] B. Kitchenham, T. Dybå, M. Jørgensen (2004): Evidence-based Software Engineering. In Proceeding of the 27th IEEE International Conference on Software Engineering, pp. 273-281, IEEE Computer Society.

[8] T. Dybå, B. Kitchenham, M. Jørgensen (2005): Evidence-based Software Engineering for Practitioners. IEEE Software 22(1): 58-65.

[9] S. Jalali, C. Wohlin (2011): Global Software Engineering and Agile Practices: A Systematic Review. Journal of Software: Evolution and Process, published online: DOI: 10.1002/smr.561.

[10] S. Jalali, C. Wohlin (2011): Global Software Engineering and Agile Practices: A Systematic Review through Snowballing, Technical Report, http://www.wohlin.eu/GS_Search_Agile_and_Global.pdf.

[11] B. Kitchenham, S. Charters (2007): Guidelines for Performing Systematic Literature Reviews in Software Engineering. Version 2.3, Technical Report, Software Engineering Group, Keele University and Department of Computer Science, University of Durham.

[12] S. MacDonell, M. Shepperd, B. Kitchenham, E. Mendes (2010): How Reliable are Systematic Reviews in Empirical Software Engineering?. IEEE Transactions on Software Engineering 36(5): 676-687.

[13] H. Zhang, M. A. Babar, P. Tell (2011): Identifying Relevant Studies in Software Engineering. Information and Software Technology 53(6): 625-637.

[14] P. Brereton, B. Kitchenham, D. Budgen, M. Turner, M. Khalil (2007): Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain. Journal of Systems and Software 80(4): 571-583.

[15] M. Skoglund, P. Runeson (2009): Reference-based Search Strategies in Systematic Reviews. In the Proceedings of the 13th International Conference on Evaluation and Assessment in Software Engineering, Durham, England.

[16] B. Kitchenham, P. Brereton, Z. Li, D. Budgen, A. Burn (2001): Repeatability of Systematic Literature Reviews. In the Proceedings of the 15th International Conference on Evaluation and Assessment in Software Engineering, pp. 46-55.

[17] R. Wieringa, N. A. M. Maiden, N. R. Mead, C. Rolland (2006): Requirements Engineering Paper Classification and Evaluation Criteria: a Proposal and a Discussion. Journal of Requirements Engineering 11(1): 102-107.

[18] M. A. Babar, H. Zhang (2009): Systematic Literature Reviews in Software Engineering: Preliminary Results from Interviews with Researchers. In the Proceedings of the 3rd International Symposium On Empirical Software Engineering And Measurement, pp. 346-355, IEEE Computer Society.

[19] S. Jalali, C. Wohlin (2011): Systematic Literature Studies: Database Searches vs. Backward Snowballing. Technical Report, http://www.wohlin.eu/Database_Snowballing.pdf.

[20] K. Petersen, R. Feldt, S. Mujtaba, M. Mattsson (2008): Systematic Mapping Studies in Software Engineering. In the Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering.

APPENDIX

Table 4 visualizes the differences between unique set of papers found through S1 and S2.

Table 4: Differences of Papers in S1 and S2

Study	Year	Ref.	Database	Keyword 1	Keyword 2	Expected	Comments
S1	2004	[D19]	ACM	?	Open Source	M	"TDD" in the title
S1	2004	[D15]	Inspec	XP	Distributed	Y	
S1	2004	[D22]	IEEE	XP	Outsourcing	Y	
S1	2004	[D7]	ACM	XP, Agile	?	N	
S1	2005	[D10]	ACM	?	?	N	
S1	2005	[D11]	IEEE	Agile	?	N	
S1	2005	[D13]	Compendex	Agile, XP	?	N	
S1	2007	[D18]	IEEE	Agile	?	N	
S1	2007	[D23]	AIS	Agile	Global, Distributed	Y	
S1	2007	[D1]	IEEE	XP	Offshore	Y	
S1	2007	[D2]	ACM	Agile	Offshore	Y	
S1	2007	[D3]	Compendex	?	Offshore	M	"User Stories", "Acceptance Tests" in the title
S1	2007	[D4]	Inspec	Agile	Global	Y	
S1	2007	[D8]	Inspec	Agile	Distributed Teams	Y	
S1	2007	[D9]	Compendex	Agile	Offshoring	Y	
S1	2008	[D12]	IEEE	?	?	N	"Continuous Integration" in the title
S1	2008	[D14]	AIS	Agile	Distributed	Y	
S1	2008	[D5]	IEEE	Scrum	?	N	
S1	2008	[D16]	IEEE	Agile	Distributed	Y	
S1	2008	[D20]	IEEE	Scrum	Offshore	Y	
S1	2008	[D21]	IEEE	?	Distributed	N	
S1	2008	[D24]	IEEE	Agile	Distributed	Y	
S1	2009	[D25]	Scopus	Agile	Distributed	Y	
S1	2009	[D26]	Scopus	Agile	Distributed	Y	
S1	2009	[D17]	Compendex	Scrum	?	N	
S1	2009	[D6]	IEEE	Agile	?	N	
S2	2000	[B14]	ACM	XP	Open-source	Y	
S2	2000	[B7]	IEEE	Daily Build	Distributed	M	"Daily Build" is not in the set of keywords of S1
S2	2001	[B8]	XP Proc.	XP	Distributed	N	
S2	2003	[B2]	Scopus	XP	Global Software Development	Y	
S2	2003	[B6]	Springer	Scrum, XP	Cross-continent	N	
S2	2004	[B12]	IEEE	XP	Global Software Development	Y	It was excluded in S1
S2	2004	[B15]	GSD Proc.	Iterative	Global Software Development	N	"Iterative" is not in the set of keywords of S1
S2	2005	[B9]	DSD Proc.	XP	Distributed	N	
S2	2005	[B13]	IEEE	XP	Distributed	Y	
S2	2006	[B4]	IEEE	XP	Distributed	M	"Extremely" refers to XP
S2	2007	[B3]	AIS	Scrum	Distributed	Y	
S2	2008	[B10]	IEEE	Scrum	Distributed	Y	It was excluded in S1
S2	2009	[B5]	MIPRO Proc.	Agile	Globally Distributed	N	
S2	2009	[B11]	Springer	XP	Offshore	N	
S2	2009	[B1]	IEEE	Agile	Remote	M	"Remote" is not within the keywords of S1