Performing Systematic Literature Reviews With Novices: An Iterative Approach

Mathieu Lavallée, Pierre-N. Robillard, and Reza Mirsalari

***Abstract*—Reviewers performing systematic literature reviews require understanding of the review process and of the knowledge domain. This paper presents an iterative approach for conducting systematic literature reviews that addresses the problems faced by reviewers who are novices in one or both levels of understanding. This approach is derived from traditional systematic literature reviews and based on observations from four systematic reviews performed in an academic setting. These reviews demonstrated the importance of defining iterations for the eight tasks of the review process. The iterative approach enables experiential learning from the two levels of understanding: the process level and the domain level.**

***Index Terms*—Computer science education, engineering education, engineering students, reviews.**

I. Introduction

PERFORMING a systematic literature review (SLR) requires an expert to find the relevant studies, compile the important conclusions, analyze the key data, and synthesize the state of knowledge. Even with expert support, recent evidence shows that reviews performed by novices are not repeatable [1], unlike reviews made entirely by experts [2]. This therefore raises the following research questions.

RQ1) What can be done to ensure that the two main qualities of systematic reviews, completeness and repeatability [3], are optimal when novices are involved?

RQ2) Can novices successfully perform a systematic literature review, where success is defined as a review that is both complete and repeatable?

This paper presents a new iterative systematic review (iSR) approach designed for users who are novices in the domain and/or in the review process itself. This approach is based on the current state of the literature and the authors' own experience guiding students in performing SLR. It was built and tested on four systematic reviews: The first two were used to build the approach, and the final two were used to test it. Although the method presented here was developed within a software engineering program, it is not specific to this domain and could be applied to other areas.

II. Related Work

The two main qualities demanded of an SLR are completeness and repeatability [3]. Completeness implies that it must cover the whole field under study and not be limited to subjectively selected papers. Repeatability implies that an independent team following the same process would reach the same conclusions.

Repeatability of the systematic literature review process has been confirmed, but only with teams comprising domain and systematic review experts [2]. The results of novices are much less repeatable, casting doubt on the capability of novices to perform publication-grade quality literature reviews [1]. Repeatability ensures that independent teams following the same systematic literature review process in the same field would reach the same conclusion. Poor repeatability can throw serious doubt on the scientific value of the review.

Completeness has not been clearly confirmed: The lack of reporting standards in the domain makes it difficult to define an adequate search strategy [4]. Different authors use different definitions for the same concept, making the creation of an appropriate search string nearly impossible for many topics. Completeness ensures that the systematic literature review covers the entire area and is not a reflection of some handpicked literature. Poor completeness may indicate that the selection of the material under study was biased to draw specific conclusions.

To achieve both completeness and repeatability, many authors recommend reiterations of certain tasks [5]. However, in all the reported cases, the iterations remain limited to a set of specific tasks, mostly related to question definition and search strategy. As will be shown in this paper, iterations are beneficial for other tasks, such as data extraction, analysis, and synthesis, especially when novices are involved. This is because repeatability and completeness are related to the entire review process, from the identification of the problem to the writing of the conclusions.

The iSR process is based on the theory of experiential learning, where procedural knowledge is acquired as the tasks are performed, or, as Kolb writes:

> *"[Learners] must be able to involve themselves fully, openly, and without bias in new experiences; reflect on and observe these experiences from many perspectives; create concepts that integrate their observations into logically sound theories; and use these theories to make decisions and solve problems"* [6].

The iSR approach defines these experiences in eight tasks:

1) Review planning: Plan the review effort and training activities.
2) Question formulation: Define the research questions.
3) Search strategy: Define the review scope and search strings.

Manuscript received June 20, 2013; revised September 06, 2013; accepted November 02, 2013. Date of publication December 05, 2013; date of current version July 31, 2014.

The authors are with Génie informatique et génie logiciel, Polytechnique Montreal, Montreal, QC H3T 1J4, Canada (e-mail: mathieu.lavallee@polymtl.ca).

Digital Object Identifier 10.1109/TE.2013.2292570

4) Selection process: Define inclusion and exclusion criteria.
5) Strength of the evidence: Define what makes a high quality paper.
6) Analysis: Extract the evidence from the selected papers.
7) Synthesis: Structure the evidence in order to draw conclusions.
8) Process monitoring: Ensure the process is repeatable and complete.

A. Review Planning in the Literature

The literature recommends that the following topics be presented to novice students at some point during the review process: SLR, along with an explanation of its use, depending on the focus of the research [7]–[12]; literature review planning [7], including a planning tutorial [7]–[12]; an introduction to the scientific method [10], to help students understand what makes a high quality paper; approaches to empirical studies [10], [12]; a brief overview of common statistical calculations [10], [13]; and an introduction to common research biases [10].

On the teaching of statistics, Brereton *et al.* [13] say that "*some statistical knowledge is needed and we found that training in basic statistical techniques, such as proportions, confidence intervals, and significance levels, was required*."

B. Question Formulation in the Literature

The formulation of the research question can be initiated with a very generic question, such as "*What has already been written on subject X?*" [5]. More targeted questions can be introduced as the domain becomes better understood [5], [14]. MacDonell *et al.* propose that question writing should follow the PICO approach of Population, Intervention, Comparison intervention, and Outcome [2]. Question formulation has been found by many authors to be a very difficult a task [5], [7], [9], [12]. They suggest a four-step procedure for writing good research questions [15]: identify the problem; write down the relevant definitions; write down the assumptions made; identify your own preconceived ideas (hypotheses).

C. Search Strategy in the Literature

The construction of a search string often uses the "population AND intervention AND outcome" structure [13], [16]. Building a good search string involves the same paradox as building a good research question. In Brereton's words [7]:

> *"It is quite a challenge, especially when you are not an expert in the topic of study. Of course the data should come from the studies, but it is hard to establish the best search strings until you are familiar with the topic."*

The challenge of building a good search string is compounded by the fact that database search engines do not use a standardized language. To alleviate these problems, Rainer and Beecham present a short iterative procedure for the construction of a search string [15], an initiative also supported by Oates and Capper, who state that "*the search strategy is likely to be adjusted as the results are inspected and the research question evolves*" [5].

A good search string should retrieve all the papers found in an SLR of the same domain (per the recall metric), as well as retrieve as few irrelevant papers as possible (per the precision metric). This "quasi-gold standard" string evaluation approach, however, is only possible in domains where a quality SLR exists [17].

An alternative to search strings, the snowballing approach, starts with a body of high-quality relevant papers. It then searches papers that are in the reference list of these starting papers (backward snowballing) and papers that cite these starting papers (forward snowballing) [18]. However, the snowballing approach can be very sensitive to the starting papers chosen [19]. Therefore, a mix of database searches and snowballing seems to be preferable.

D. Selection Process in the Literature

The selection process typically uses the following steps, each with their own inclusion and exclusion criteria: 1) selection based on the paper title [13], [16]; 2) selection based on the paper keywords [13]; 3) selection based on the paper abstract [8], [13], [16].

Brereton *et al.* [13] give an example of a case where an exclusion criterion became evident only after the review was well under way. Inclusion and exclusion criteria should therefore be periodically revised. Another problem is that one study can be presented in multiple papers, and similarly, one paper can present multiple studies [13]. This can cause a study to be overrepresented as it is used for multiple analyses across different papers.

The literature reports that novices can either be too restrictive [9] or not restrictive enough [20] in their choice of inclusion and exclusion criteria. Instructors should therefore keep a close eye on the progress of the selection process in order to ensure that the criteria are clearly and fully defined. Inter-rater agreement metrics like the Cronbach Alpha [21] and the Fleiss' Kappa [22] can be used to assess the quality of a selection process.

E. Strength of the Evidence in the Literature

Many authors [5], [9], [13] consider this task to be the most difficult for novices to perform. According to Kitchenham [8], "*Students found evaluating the quality of studies found in a systematic review particularly problematic*." As a result, Brereton *et al.* decided to skip this activity during their 2011 review [7]. As they describe it:

> *"Students were not expected to assess study quality, an activity considered especially challenging, even for experienced researchers."*

They affirm that novice reviewers are not accustomed to the way in which empirical papers are written. Even the use of a simple checklist does not improve the situation. Rainer *et al.* report that some "*students simply did not use the checklist, a number of students used it poorly, whilst some students used it well. The varied use of the checklist is surprising, as it had already been used in some tutorials and lectures*" [12].

This means that research quality evaluation is often based on the student's perception, rather than on the evidence in the

TABLE I
CONTEXT OF THE SLRS PERFORMED TO BUILD AND TEST THE iSR PROCESS

Review	Domain	Participants	Process expertise	Domain expertise	Process used	Results
Fall 2010	Software process improvement	5	Novices	Novices	Biolchini et al. [28]	Partial success (repeatable but incomplete): Analysis and synthesis had to be completely redone.
Fall 2011	Web tool use	14	Novices	Novices	Biolchini et al. [28]	Failure (non-repeatable and incomplete): Bad research question was not corrected, resulting in poor inclusion/exclusion criteria.
Summer 2012	Pair programming	2	Proficient and expert	Novices	iSR	Success (repeatable and complete).
Fall 2012	Web information lookup	3	Novices	Novices	iSR	Success (repeatable and complete).

paper. The quality evaluation thus becomes a matter of personal opinion.

F. Analysis in the Literature

Strauss [23, p. 19] writes that qualitative data collection and codification should follow an iterative and incremental approach. The data collected, the code used to qualify the data, and the memos used to organize the codes must be constantly refined as the selected articles become better understood.

Petticrew and Roberts [14, p. 165] recommend focusing on the studies having the highest research quality. The main synthesis work (and conclusions) should be based on the data extracted from these top-quality papers, with the lower-quality ones used for either confirmation or to support minor conclusions. They also recommend not transforming qualitative data into quantitative data [14, p. 191]. Practices like tally counting can be useful, but they result in a large amount of important contextual information being neglected, and so should be used with care. The works of Hannay *et al.* show how one can adequately pool data from contextually different studies into a single meta-analysis [24].

Brereton *et al.* [7] maintain that a "key problem" in data extraction is that students have to know which data are relevant, even though they are domain novices. The extraction must therefore be adjusted as the need arises and as the students' comprehension of the field improves.

G. Synthesis in the Literature

The activity of producing a synthesis has been reported as difficult by many authors [8], [16]. One reason for this stems from the use of data extracted by other reviewers. Baldassarre *et al.* describe filling out the data aggregation forms as "*a demanding task, as it requires combining the individual work of a number of students"* [16].

The poor quality of syntheses in literature reviews reported by Cruzes and Dyba [25] can be traced back to the existing review processes, which provide very few details on how to perform a synthesis. Consequently, there is a real need for a defined synthesis procedure.

Janzen and Ryoo [26] present a successful synthesis process when working with students. They asked each of their students to read 17 articles and summarize them. The synthesis approach was to write one global summary based on the 17 individual article summaries. The quality of the summaries produced proved to be equal to or better than that of the summaries produced by experts. This successive summarization approach also proved useful to initiate systematic review syntheses.

Another synthesis approach seen in the literature is to use a validated model, such as a recognized domain taxonomy, as a basis for the categorization of the evidence [6].

H. Process Monitoring in the Literature

This task is not defined in the traditional SLR processes. The "big upfront" approach to traditional literature reviews implies that all resources are assigned at the beginning and cannot be easily moved during process execution. This task evaluates the repeatability and completeness of the review results and produces recommendations for the next iterations.

III. CONTEXT OF THE REVIEWS PERFORMED

To build and validate the iSR approach, four systematic reviews were performed, each in a different domain, as described in Table I. All reviews were performed with students, over one semester of a software engineering graduate course. The goal was to introduce the students to the state of the art in a specific domain: The whole class therefore worked as a single team.

Students attended a weekly 3-h lecture, and then carried out a home assignment either individually or in pairs. The lecture was split into three parts: During the first hour, the instructor provided feedback on the previous week's assignment; during the second hour, the instructor introduced new concepts; and for the third hour, students and instructor discussed issues seen in the results and planned the next assignment.

The first review, performed during the Fall 2010 semester, involved domain and process novices. The domain targeted was the impact of software process improvement on developers. The review was successful and the results were published [27]. Nevertheless, a number of problems emerged, mostly at the analysis and synthesis stages, necessitating some rework of the analysis and synthesis tasks.

The second review was performed during the Fall 2011 semester and involved domain and process novices. The do-

main targeted was the use of Web-based tools during software development. The review failed, mostly because the size of the body of research on distributed (or global) software development had been underestimated. The results did not provide an accurate synthesis of the domain and were found to be too biased to be pursued.

The problems encountered in the Fall 2010 and Fall 2011 reviews motivated the development of the iSR approach. To validate this new approach, a third review was conducted during the summer of 2012, whose objective was to make a synthesis of the various definitions of Pair Programming practices reported in software engineering studies. The literature review was conducted by two graduate students familiar with SLR processes, but not with the iSR approach. They were domain novices. The review was successful and, through multiple iterations, they managed to produce a good summary of the targeted domain. The review results led to a paper submission.

The fourth review was performed during the Fall 2012 semester by three graduate students, all of whom were domain novices and process novices. The domain targeted was software developers' search for information on the Web. The review demonstrated the usefulness of an iterative approach, as the research questions had to be adjusted multiple times during the review to account for the state of the domain. The review results have since been integrated into a paper currently being written.

A. Evaluation of the Students

Students were graded on their completing the week's assignments, with the grade itself being based on the process activities rather than the content. Disagreements on the content were discussed in class during plenary sessions, which typically resulted in changes in the process and in the task assignments for the following week. Relevant interventions during plenary sessions require that students master self-evaluation capabilities. Self-evaluation is most efficient when the students execute the process tasks over and over again, hence the importance of iterations.

B. Roles Within the Process

This process defines two main roles: instructor and student. Both instructors and students can be domain novices. Ideally, a domain expert should validate the work performed, but the iterative approach of the iSR process can produce good results without expert feedback. The students can also be process novices, but the instructor should at least be familiar with the process. The role of the instructor is to present the necessary training to the students and to gather the results of their subsequent work.

The role of the student is to perform the actual reviewing work, build the search strings, execute the selection process, read the selected material, extract the data, synthesize the results, and so on.

IV. Lessons Learned From the iSR Approach

This section presents the lessons learned from the eight tasks comprising the iSR approach. Fig. 1 presents a graphical view of the effort expanded on the eight tasks of an iSR process composed of ten iterations; the curves for each task are normalized.

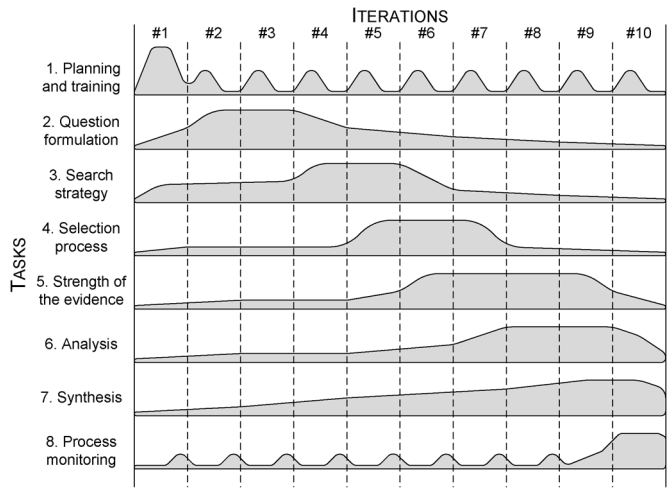

Fig. 1. Example of the iSR process in action. The shaded area represents an artistic view of the effort expanded on each task for a given iteration.

A. Planning and Training

This task has two objectives: first, to plan the review work, and second, to plan training sessions and tutorials to provide novice students with the technical know-how required to perform a literature review.

All documentation provided to the students should be carefully written, as the repeatability of the process requires clarity and understandability. Any ambiguous statement can result in divergent results, which degrade repeatability. The use of an iterative approach resolves part of this problem by offering opportunities to correct any misunderstanding.

The seminar presentations, tutorial, and exercises should follow the first steps of Bloom's taxonomy of educational objectives [29], in order to introduce the concepts progressively. The following three levels should be considered:

- *Remember*: Ask students to follow instructions by rote.
- *Understand*: Ask students to provide a feedback on the appropriateness of the instructions given.
- *Apply*: Ask students to tailor the instructions to the context at hand.

During the systematic review process, the students will need to differentiate strong search strings from weak ones, high-quality studies from biased ones, and so on. It is important to teach students how to perform self-evaluations of their work. These quality evaluations require a high level of understanding, something that cannot be achieved with the rote application of a method.

All the knowledge required should not be transmitted to the novice students right at the start of the process. For example, the importance of distinguishing high-quality from low-quality papers does not emerge until a number of relevant papers have been selected. Tutorials on the scientific method can therefore be delayed until those concepts are needed.

The reviews showed the importance of adapting the work plan. Some domain idiosyncrasies only become known during

the execution of the review, which requires reiterating the work performed. The review planning therefore needs to be regularly revised.

Another problem is the pressure to meet a deadline. It is tempting to cut corners in order to complete the review process. It is, however, better to have an incomplete review containing good quality data that can be completed later on by another team than a botched review that is useless since it needs to be completely redone.

B. Question Formulation

The aim of this task is to define research questions and review objectives. Calibrating the research question is essential in order to obtain a manageable, yet significant, body of literature. It is usually not practical to carry out an SLR based on hundreds of papers since these could take years to analyze and synthesize.

The question formulation task suffers from a paradox: The domain must be well understood in order to produce good research questions. However, to understand the domain, the research must first be performed.

The main cause of the failed review in the Fall 2011 semester was the fact that the question posed was not suited to the current state of the domain. Redefining the research questions is useful when the current state of the domain does not match initial expectations, which is a frequent issue when domain novices are involved. Consequently, research questions must be iteratively readjusted.

C. Search Strategy

This task's objective is to define the journals and conferences to be targeted, the databases to be searched, to decide whether snowballing will be applied or not, and whether "grey" literature[1] will be searched or not, and the like.

The "term impact analysis" string validation method consists of testing the search string with and without each term, to evaluate how it affects the results. A term with no impact on the results can be safely discarded. A term responsible for a large part of the results might be too generic and could be refined. This methodology could resolve the problem that "*students provided poor explanations in their reports of how their searches were conducted*" [11].

Another problem with the search strategy is that reviewers tend to perform a typical Google-type search. They are looking for the one exact answer to the research question instead of looking for all the relevant answers.

There were also some technical problems inherent in building search strings. The most common problems found were the following: inappropriate use of wildcard characters (?,*); orthographical errors; subsumed terms ("software" OR "software engineering"); and unbalanced strings ("software" OR "test-driven development"). Most of these problems can be detected by a validation technique like term impact analysis.

Iterations based on a pilot approach also work well for building the search string. A small sample of the papers found with a search string can be transferred to the selection, analysis, and synthesis tasks. This can help the reviewer spot relevant and irrelevant keywords that were not included in the initial search string.

Finally, some of the important keywords are not immediately obvious, and only become clear as they are seen in the papers. The search string must therefore be adjusted each time a new keyword emerges as either relevant or irrelevant. The major risk is therefore to find new, relevant keywords late in the review process.

D. Selection Process

Existing literature review processes often mix the quality in terms of the relevance of the paper for the purposes of the review and quality in terms of the paper's research methodology. "Relevance quality" is measured during the Selection Process task. "Research quality" is measured during the Strength of Evidence task.

Relevance quality is determined by the inclusion and exclusion criteria that define which papers are relevant and should be kept, and which papers are off-topic and should be rejected.

Two guidelines are proposed for this task.

- The selection-by-title stage rejects only duplicates titles and proceedings introductions since there is often little information in a title.
- The selection-by-abstract is based on the relevance of the content.

An "overview step" is needed that consists of looking at the tables, figures, and conclusions of the paper. The main purpose is to perform a final relevance quality evaluation, along with a preliminary research quality evaluation [7].

Most review processes suggest the use of a spreadsheet application like Excel to briefly document the inclusion or exclusion rationale for each paper. These justifications will help in the revision of the inclusion and exclusion criteria.

Misinterpretation of inclusion and exclusion criteria explains some of the differences found in reviews performed by novices [1]. When misinterpretations occur in the selection process, these must be discussed freely and openly with the students, to establish the reasons for the poor results. These discussions enable the instructors and students to reach agreement on the value of the papers. Students were rarely to blame for poor inter-rater agreement; rather, the culprit was generally poorly worded assignments and ambiguous tasks.

However, Excel is not entirely appropriate for collaborative work, and synchronization proved problematical when two students worked in parallel.

The third literature review, on Pair Programming (PP), generated some discussions on what constitutes a PP paper and what does not. Since the definition of the practice proved to be very fluid, many papers initially thought to be irrelevant were reconsidered. The importance of this selection criterion did not emerge until synthesis began. Thus, one benefit of the iterative approach is that not all the resources were expended at the start of the synthesis task. A common issue of this task is the definition of overly restrictive inclusion and exclusion criteria, typical of a Google-type search.

[1]Grey literature comprises unpublished papers or papers published in non-peer-reviewed journals. These items mostly consist of technical reports, the quality of which can vary from excellent to mediocre.

E. Strength of the Evidence

This task is aimed at evaluating the research quality of the selected papers. The process uses a two-step approach for the research quality evaluations. The first step is to produce the overview, which consists of the verification of the presence of three critical context details: a description of the sample and its population, a description of the study context, and clearly written study conclusions.

The second step is to carry out a thorough quality evaluation, using a checklist based on the works of Dyba and Dingosyr [30]. Filling out this form, however, requires a thorough reading of the full text, a task typically performed in parallel with the Analysis task. Each element of the checklist uses a three-level evaluation: (0) the element is absent or very poor; (1) the element is present but could be more detailed; (2) the element is present and sufficiently detailed.

While previous studies reported problems with the use of a checklist [12], it was found that multiple iterations on research quality evaluation resolved these issues.

F. Analysis

In this task, the papers are analyzed to extract the relevant data, typically using a provided extraction form. The data extracted is not always the data needed for the synthesis, because it is not known beforehand which is the data required for the synthesis. In many cases, the extraction had to be redone once the synthesis approach had been clarified.

Providing novice reviewers with a sample completed form helped them understand what is required in each field. The form needs to be sufficiently detailed to ensure that the results are comparable from one reviewer to another. Students should also focus on the conclusions of the studies, as they often provide the "meat" of the synthesis.

According to qualitative feedback from the reviewers, this step is the most time-consuming activity of the process, although not the most mentally demanding. In light of this, extraction work should start as soon as possible, in order to ensure that reviewers are familiar with the extraction procedures and the form and understand what is required of them. Early extraction can also avoid a sudden drop in extraction quality toward the end of the review, when deadline pressures loom. In addition, some extractions tend to be too succinct, and others too verbose. Constant evaluation and feedback on the results is therefore required to ensure that reviewers understand what is expected of them.

A learning iteration is recommended for novice reviewers, where they are assigned a short synthesis exercise based on good and bad extractions performed by other reviewers. The goal of such an exercise is to clarify what differentiates a good extraction from a bad one.

Another approach used was to perform the analysis in stages. During the first analysis iteration, a small batch of papers is analyzed, and an attempt is made to synthesize the extracted data. The extraction form is then modified based on the synthesis results. During subsequent iterations, the data extraction process is corrected, and the analysis is then performed on a larger batch of papers. This approach proved fruitful, as in most cases the final extraction form was very different from the initial one.

G. Synthesis

The reviews show that synthesis benefits greatly from an iterative approach. To find some structure in a large amount of qualitative data requires many attempts. Each attempt might also call for a different view of the data, which might in turn require a revision of the extraction form. In some cases the data extracted did not support some interesting synthesis that had emerged. This required the reviewers to revisit the source papers to extract the missing information.

Students are typically expecting Google-type search responses. As a result, they discard many articles as irrelevant, even though they provide interesting answer fragments.

The synthesis approach that provided the best results in the reviews involves successive summarization, much like Janzen and Ryoo's approach [26]. The following are the four steps in the approach.

1) Each selected paper is summarized in a single paragraph, based on its extraction form.
2) Each summary paragraph is further summarized into a single phrase.
3) The summary phrases from all the selected papers are put together to see how they support or contradict one another.
4) A single paragraph is built based on the phrases summarizing all the selected papers.

At every stage of the approach, the cohesiveness of the results is monitored. The phrases and paragraphs built must make a coherent whole. The resulting paragraph does not consider the context of the studies and should not be used as-is in the review synthesis. The objective of this exercise is limited to finding a potential structure for the evidence found in the selected papers. Feedback from the students underlines the fact that the synthesis activity is not especially time-consuming, but it is mentally demanding. The successive summarization approach helps the students find the relationships between the pieces of evidence, but building a conceptual map of the evidence still involves a significant mental strain.

H. Process Monitoring

This task is related to the collection of data during the systematic review. Its goal is to improve the process for future reviews and to plan the resources needed for the forthcoming iterations.

The results obtained are not always good enough to warrant a transition to the next task. Mistakes made in the previous tasks must be corrected before more effort can be invested in the next ones, to ensure that results reflects the state of the literature (completeness), and can be redone with similar results (repeatability).

This paper is the result of multiple improvements to the systematic review processes currently available. These improvements were made possible by gathering data during every task of the process. Reviewers were required to report the time spent on the tasks and to provide feedback on the difficulties encountered. These reports enabled the authors to detect problems and propose solutions to common issues arising in systematic review processes.

Solutions like the term impact analysis for search string validation and the successive summarization approach for the synthesis of the evidence were introduced because of reported issues with these tasks.

V. Conclusion

Lessons learned from multiple systematic reviews demonstrate that an iterative approach can be beneficial when working with domain and process novices. As the review progresses, the perception of the domain by novices changes, and the design of the review should evolve accordingly. The research questions may have to be rewritten, the selection procedure may need some adjustment, the extraction forms and analysis tables may require revision, and the synthesis conclusions may need to be redesigned. This approach should produce better and more accurate results with each iteration, reflecting the progressive gain in expertise and understanding of the novices. It should also enable instructors to calibrate the effort output required through the addition or removal of iterations.

References

[1] B. Kitchenham, P. Brereton, Z. Li, D. Budgen, and A. Burn, "Repeatability of systematic literature reviews," in *Proc. 15th EASE*, Durham, U.K., 2011, pp. 46–55.

[2] S. MacDonell, M. Shepperd, B. Kitchenham, and E. Mendes, "How reliable are systematic reviews in Empirical Software Engineering?," *IEEE Trans. Softw. Eng.*, vol. 36, no. 5, pp. 676–687, Sep.–Oct. 2010.

[3] B. Kitchenham, P. Brereton, and D. Budgen, "Mapping study completeness and reliability—A case study," in *IET Seminar Dig.*, 2012, vol. 2012, pp. 126–135.

[4] O. Dieste, A. Griman, and N. Juristo, "Developing search strategies for detecting relevant experiments," *Empir. Softw. Eng.*, vol. 14, no. 5, pp. 513–539, Oct. 2009.

[5] B. J. Oates and G. Capper, "Using systematic reviews and evidence-based software engineering with masters students," in *Proc. 13th EASE*, Durham, U.K., 2009, pp. 79–87.

[6] F. O. Bjørnson and T. Dingsøyr, "Knowledge management in software engineering: A systematic review of studied concepts, findings and research methods used," *Inf. Softw. Technol.*, vol. 50, no. 11, pp. 1055–1068, Oct. 2008.

[7] P. Brereton, "a study of computing undergraduates undertaking a systematic literature review," *IEEE Trans. Educ.*, vol. 54, no. 4, pp. 558–563, Nov. 2011.

[8] B. Kitchenham, P. Brereton, and D. Budgen, "The educational value of mapping studies of software engineering literature," in *Proc. 32nd ICSE*, Cape Town, South Africa, 2010, pp. 589–598.

[9] A. Rainer and S. Beecham, "A follow-up empirical evaluation of evidence based software engineering by undergraduate students," in *Proc. 12th EASE*, Bari, Italy, 2008, pp. 78–87.

[10] M. Jorgensen, T. Dybå, and B. Kitchenham, "Teaching evidence-based software engineering to University Students," in *Proc. 11th METRICS*, Como, Italy, 2005, pp. 213–220.

[11] A. Rainer, S. Beecham, and C. Sanderson, "An assessment of published evaluations of requirements management tools," in *Proc. 13th EASE*, Durham, U.K., 2009, pp. 98–107.

[12] A. Rainer, T. Hall, and N. Baddoo, "A preliminary empirical investigation of the use of evidence based software engineering by under-graduate students," in *Proc. 10th EASE*, UK, 2006, pp. 91–100, Keele University.

[13] P. Brereton, M. Turner, and R. Kaur, "Pair programming as a teaching tool: A student review of empirical studies," in *Proc. 22nd CSEET*, Hyderabad, India, 2009, pp. 240–247.

[14] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences*. Malden, MA, USA: Blackwell, 2008.

[15] A. Rainer and S. Beecham, "Supplementary guidelines and assessment scheme for the use of evidence based software engineering," Univ. Hertfordshire, Hertfordshire, U.K., Tech. Rep. CS-TR-469, 2008.

[16] M. T. Baldassarre, N. Boffoli, D. Caivano, and G. Visaggio, "A hands-on approach for teaching systematic review," in *Proc. 9th PROFES*, Rome, Italy, 2008, pp. 415–426.

[17] H. Zhang, M. A. Babar, and P. Tell, "Identifying relevant studies in software engineering," *Inf. Softw. Technol.*, vol. 53, no. 6, pp. 625–637, Jun. 2011.

[18] S. Jalali and C. Wohlin, "Systematic literature studies: Database searches vs. backward snowballing," in *Proc. 6th ACM-IEEE ESEM*, Sweden, 2012, pp. 29–38, Lund University.

[19] M. Skoglund and P. Runeson, "Reference-based search strategies in systematic reviews," in *Proc. 13th EASE*, Durham, U.K., 2009, pp. 31–40.

[20] P. Brereton, M. Turner, and R. Kaur, "Pair programming as a teaching tool: A student review of empirical studies," in *Proc. 22nd CSEET*, Hyderabad, India, 2009, pp. 240–247.

[21] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951.

[22] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, Nov. 1971.

[23] A. L. Strauss, *Qualitative Analysis for Social Scientists*. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[24] J. E. Hannay, T. Dyba, E. Arisholm, and D. I. K. Sjoberg, "The effectiveness of pair programming: A meta-analysis," *Inf. Softw. Technol.*, vol. 51, no. 7, pp. 1110–1122, Jul. 2009.

[25] D. S. Cruzes and T. Dybå, "Research synthesis in software engineering: A tertiary study," *Inf. Softw. Technol.*, vol. 53, no. 5, pp. 440–455, May 2011.

[26] D. S. Janzen and J. Ryoo, "Seeds of evidence: Integrating evidence-based software engineering," in *Proc. 21st CSEET*, Charleston, SC, USA, 2008, pp. 223–230.

[27] M. Lavallée and P. N. Robillard, "The impacts of software process improvement on developers: A systematic review," in *Proc. ICSE*, Zurich, Switzerland, 2012, pp. 113–122.

[28] J. Biolchini, P. Gomes Mian, A. Candida Cruz Natali, and G. Horta Travassos, "Systematic review in software engineering," PESC, Rio de Janeiro, Brazil, Tech. Rep. RT-ES 679/05, 2005.

[29] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwolh, *Taxonomy of Educational Objectives: The Classification of Educational Goals*. London, U.K.: Longman, 1956.

[30] T. Dyba and T. Dingsoyr, "Empirical studies of agile software development: A systematic review," *Inf. Softw. Technol.*, vol. 50, no. 9–10, pp. 833–859, Aug. 2008.

Mathieu Lavallée received the B.S.E. degree in software engineering from Polytechnique Montreal, Montreal, QC, Canada, in 2010, where he is currently pursuing the Ph.D. degree in the impact of information flows within software development teams on product quality.

He has been working with the Laboratoire de Recherche en Génie Logiciel under the supervision of Pierre N. Robillard since 2009 and is the author of eight papers related to knowledge management and team dynamics within software engineering. His current research works focus on information acquisition within software development teams.

Pierre-N. Robillard received the B.Sc degree in physics from the University of Montréal, Montréal, QC, Canada, in 1969, the M.Sc. degree in high-energy particles and M.A.Sc. degree in computer communications from the University of Toronto, Toronto, ON, Canada, in 1973, and the Ph.D. degree electrical engineering from the University Laval, Quebec City, QC, Canada, in 1977.

He is a Professor with Polytechnique Montréal, University of Montréal, and was the founding Chairman of the Department of Computer and Software Engineering in 2000. He has authored or coauthored about 120 conference and journal papers. His research interests are in software process engineering, software quality, and team dynamics.

Dr. Robillard is a professional engineer and a member of various professional associations. He serves as a reviewer for several journal publications.

Reza Mirsalari received the B.S. degree in software engineering from the Najafabad University, Najafabad, Iran, in 1996, and the M.Sc. degree in computer science and software engineering from Polytechnique Montreal, Montreal, QC, Canada, in 2013, and is currently pursuing the Ph.D. degree at Polytechnique Montreal, focusing on software quality and software development teams.

From 2000 to 2010, he worked as an industrial expert on business process management and systematic process management as it pertains to information systems. He has also worked recently on business process automation within a software development company in Montreal.