Empir Software Eng (2018) 23:1594–1663
DOI 10.1007/s10664-017-9556-7

Challenges and pitfalls on surveying evidence in the software engineering technical literature: an exploratory study with novices

Talita Vieira Ribeiro[1] · **Jobson Massollar**[1] · **Guilherme Horta Travassos**[1]

Published online: 28 October 2017

Abstract The evidence-based software engineering approach advocates the use of evidence from empirical studies to support the decisions on the adoption of software technologies by practitioners in the software industry. To this end, many guidelines have been proposed to contribute to the execution and repeatability of literature reviews, and to the confidence of their results, especially regarding systematic literature reviews (SLR). To investigate similarities and differences, and to characterize the challenges and pitfalls of the planning and generated results of SLR research protocols dealing with the same research question and performed by similar teams of novice researchers in the context of the software engineering field. We qualitatively compared (using Jaccard and Kappa coefficients) and evaluated (using DARE) same goal SLR research protocols and outcomes undertaken by similar research teams. Seven similar SLR protocols regarding quality attributes for use cases executed in 2010 and 2012 enabled us to observe unexpected differences in their planning and execution. Even when the participants reached some agreement in the planning, the outcomes were different. The research protocols and reports allowed us to observe six challenges contributing to the divergences in the results: researchers' inexperience in the topic, researchers' inexperience in the method, lack of clearness and completeness of the papers, lack of a common terminology regarding the problem domain, lack of research verification procedures, and lack of commitment to the

Communicated by: Emerson Murphy-Hill

✉ Talita Vieira Ribeiro
tvribeiro@cos.ufrj.br

Jobson Massollar
jobson@cos.ufrj.br

Guilherme Horta Travassos
ght@cos.ufrj.br

[1] Systems Engineering and Computer Science Program (PESC/COPPE), Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

SLR. According to our findings, it is not possible to rely on results of SLRs performed by novices. Also, similarities at a starting or intermediate step during different SLR executions may not directly translate to the next steps, since non-explicit information might entail differences in the outcomes, hampering the repeatability and confidence of the SLR process and results. Although we do have expectations that the presence and follow-up of a senior researcher can contribute to increasing SLRs' repeatability, this conclusion can only be drawn upon the existence of additional studies on this topic. Yet, systematic planning, transparency of decisions and verification procedures are key factors to guarantee the reliability of SLRs.

Keywords Novice researchers · Systematic literature review · Evidence-based software engineering · Exploratory study

1 Introduction

Literature reviews serve as common starting points for most scientific research, including research in the Software Engineering (SE) field. Finding and reviewing previous studies or software technologies provides benefits for researchers regarding the identification of i) knowledge and new ideas about a topic; ii) research gaps and opportunities; and iii) related work. In industrial software scenarios, practitioners can take advantage of literature reviews to support the searching for software methods, processes, techniques, and tools, among other instruments suitable for their development contexts, which lower the risk of incorrect adoption decisions in their software development settings. However, ad-hoc literature reviews can threaten their own replication, coverage, and fairness, among other features.

Systematic literature reviews (SLRs) represent a more procedural and rigorous strategy to perform literature reviews. They define a set of steps to guide the scientific literature search, producing a repeatable research protocol, allowing critical judgment about the quality of the obtained knowledge and reducing bias related to outcomes (Biolchini et al. 2005; Kitchenham and Charters 2007). *Quasi*-systematic literature reviews (Travassos et al. 2008) and systematic mapping studies (Petersen et al. 2008) are also types of SLRs. The former does not support meta-analysis due to the lack of a baseline (comparison) for evidence aggregation. The latter focus on providing an overview of an area of interest, rather than aggregating evidence for a specific purpose.

As an investigation tool, the SLR strategy play a major role in the context of evidence-based software engineering (EBSE), which aims at providing an efficient way to integrate current scientific evidence with practical experience to support the decision making in SE (Dybå et al. 2005). SLR's methodical processes for gathering, extraction, evaluation, and aggregation of evidence from various studies can assist the researchers in organizing a relevant and reliable body of knowledge regarding a specific research topic in academia. They can also assist the practitioners in finding software technologies suitable for their particular scenarios of software development in industry. As an example of the latter, Siemens Corporate Research supported the execution of an SLR aiming at investigating model-based software testing approaches (Dias Neto et al. 2007). Other examples of SLRs involving the industry can be seen in (Kasoju et al. 2013; López et al. 2015; Ulziit et al. 2015) and (Garousi et al. 2016) among others.

The importance and expected benefits of SLRs justify the concerns regarding their quality, since the topic under investigation; the experience of researchers or practitioners in both the

research method and the topic; and the means and knowledge supporting the research questions answering can compromise the results (MacDonell et al. 2010). Therefore, some guidelines to undertake SLRs have been proposed over the years, such as (Biolchini et al. 2005; Kitchenham and Charters 2007; Petersen et al. 2015) and (Kuhrmann et al. 2017). Such guidelines provide recommendations aiming at reducing threats to the validity of SLRs by advising researchers and practitioners to explain the need for the SLR and to detail the research objectives and the plan that will support the study execution. Also, many investigations concerning the planning and execution of SLRs have been published in the technical literature. In this regard, some authors claim that SLRs are robust enough to resist execution deviations, producing stable outcomes for different processes (MacDonell et al. 2010). Notwithstanding, various researchers observed incompatibilities in results in SLRs with similar goals but executed by independent investigators (Kitchenham et al. 2011, 2012; Wohlin et al. 2013) and (Munir et al. 2014) – more details in Section 2.

In this context, at an International Software Engineering Research Network (ISERN) held in 2009, a group of ISERN members raised concerns regarding the possibility of conflicting results in SLRs. At that time, they were discussing the first SLR results in the SE field. They assumed that since an SLR protocol is supposed to be explicit, precise and unbiased, its outcomes should be either equal or alike to other results obtained by other researchers or practitioners executing (replicating) it or working with SLR protocols with similar purposes. It was stressed that knowledge and experience in the method play a major role in SLR planning and execution and they could lead to differences in the outcomes, indicating that an SLR might not be suitable for those players inexperienced in the method.

Out of these discussion arise a question: if the technical literature reports inconsistencies regarding SLRs executed by novice and even expert researchers, and the EBSE relies on research-based evidence through SLRs, how can SLRs conducted by practitioners – which usually are not much acquainted with this research method – be considered reliable? This way, our aim is to discuss SLRs reliability based on the following statement: "Similar SLR protocols, executed by similar teams of novice researchers, lead to equivalent answers (outcomes) to the same research question." It is important to note that to some extent, graduate students (novice researchers) can present similar skills to practitioners, especially the less experienced ones, concerning planning and executing SLRs. About domain knowledge, practitioners may even be considered more experienced, but eventual differences in SE terminology adopted in the industry and academia can bring some difficulties to practitioners regarding SLR planning. That is, the domain knowledge may be insufficient to figure out adequate terms associated with a specific research question. A set of investigation questions were posed aiming at observing the statement above: What will happen if balanced groups of novice researchers (regarding their knowledge and experience in SLR planning and execution, and also in the research topic) plan and execute an SLR for the same research question? Should the research protocols be similar to each other, given that they address the same research question? Once similar SLR protocols are planned, should the selection of studies and reported outcomes be equal to each other, given the repeatable characteristic of SLRs? What do the differences between the planned SLRs and their results tell us about reliability (process repeatability and outcomes consistency)? How do players' (lack of) knowledge and experience affect the SLRs reliability?

To investigate these questions, we planned and accomplished an exploratory study (detailed in Section 3) to analyze the planning, execution and outcomes of seven *quasi*-SLRs carried out by novice researchers (master and doctoral students) in the context of an Experimental

Software Engineering (ESE) course in two distinct years – 2010 and 2012. The results presented in Section 4 indicate that i) when the same research question is addressed, different *quasi*-SLR protocols are planned; ii) when a similar point of view for the studies' selection strategies is reported, divergent studies are selected; and iii) when the selected studies are the same, independent teams report different results. These discrepancies reinforce the perception that the difficulties faced by novice researchers in the planning and execution of SLRs impact the approach reliability and repeatability. Based on that, we can question whether the proposed and used guidelines by the academics to carry out SLRs are feasible to support novices in the context of academia and practitioners in the industry as well.

The remainder of this paper is organized as follows. In Section 5 we present the *quasi*-SLRs scores concerning the research protocols and reports as a way to identify the main issues faced by the participants while performing the assignment. Next, in Section 6 we discuss the challenges on surveying SE evidence with novices and the strategies they can adopt to make the SLRs suitable for those inexperienced in the method and in the topic under investigation, such as practitioners (especially concerning the former). The threats to the validity of this study are in Section 7 and the Conclusions in Section 8.

2 Related Works

Several studies report on the use of novice researchers performing SLRs in SE, and even though a couple of studies mention novice researchers can undertake SLRs, they represent one of the causes for results instability in SLRs. Definition of research questions; inclusion and exclusion criteria; and data extraction and synthesis are among the main difficulties faced by novice researchers while surveying evidence in the technical literature. However, difficulties in conducting systematic reviews can also be found when expert researchers conduct them. The next subsections provide an overview of different related works that i) used students to evaluate the applicability or reliability of SLRs; ii) compared independently published literature reviews and used feedback from experts to assess the research method quality and also the barriers encountered during its execution. A summary of their results is highlighted since we used some of them to support the planning of the exploratory study presented in this paper.

2.1 SLRs and Novices

In 2006, Rainer, Hall, and Badoo presented a preliminary investigation on undergraduate students' experiences of using the EBSE approach while evaluating software technologies (Rainer et al. 2006). Overall, students had problems constructing EBSE questions, and they mainly based their questions on topics they had some experience with, for instance, programming languages to be used in their undergraduate assignments. One of their main difficulties was to formulate a question comparing software technologies. For example, the students formulated exploratory questions to identify all programming languages they could choose for their assignments, rather than developing a question to compare programming languages they were in doubt of choosing. The sources selected for collecting information to support their answers were not as expected since little scientific production was used to support their searching for information. Also, the students provided poor explanations concerning their search process, and they made different use of the available guidelines.

Oates and Capper tried to overcome some of the issues observed by Rainer, Hall, and Badoo. They carried out what they called a case study trying to answer questions related to the EBSE approach concerning its use by students (Oates and Capper 2009). They asked students to conduct an SLR on a topic of their interest and write a short essay on their experiences with the EBSE approach. The authors made some restrictions, though: they had given a question for the students to start working with; they had advised the students to search in scientific databases and to refine their search until they reached a set of articles in the range of 10 to 30. The analysis of the students' marks supported the authors' assumption that students could perform SLRs – at least upon the restrictions and guidance previously stated. The authors noticed that students need a more iterative surveying process in which they can refine their search strategy until they find works relevant to answering their research questions.

Even though Oates and Capper stated that students could perform SLRs, according to Riaz et al. their experience in conducting a complete systematic search for evidence can be quite different from experts (Riaz et al. 2010). In a study to gather the main difficulties faced by students while conducting SLRs, the authors identified issues related to building a search string that would retrieve a considerable number of papers without returning much noise; selecting appropriate works based solely on title and abstract; extracting the right amount of information from the selected works; synthesizing data that was not easily comparable; among others. While defining the research question can be challenging to both novices and experts, overall the former group faced more difficulties than the latter one.

Brereton could identify positive results in a study involving students conducting SLRs (Brereton 2011). In her case study, she observed that students were successful in undertaking most of the steps of the SLR process. The students' performance was based on marks to their activities, and, in general, students with lower marks had problems with separating the planning information from the execution information. In summary, students succeed more in planning activities, even though they mentioned that the planning phase was the most difficult part of the SLR process.

Although all these studies concluded that students could be used to perform SLRs in ES, even though they have more difficulties than experts in performing the search, there are still some issues related to SLR completeness and repeatability that they did not evaluate. Kitchenham et al. (2011) presented a case study conducted to investigate the repeatability of the results provided by SLRs. Two research assistants (RAs) planned and conducted the same SLR topic, and their results were compared to each other. Their results were also compared with a previously published literature review on the same subject conducted by experienced researchers. Even though the same search period and libraries were used for all three SLRs, they reported different sets of primary studies for the same research topic. Kitchenham et al. conjectured that the lack of experience in the research topic and in the method, and the application of the inclusion/exclusion criteria can be the reasons for these differences.

More recently Carver et al. identified barriers to the SLR process (Carver et al. 2013). The authors gathered data from their experiences conducting SLRs, as well as from feedback of graduate students in an SLR course, and from authors of published SLRs. Among the most difficult tasks of the process are the ones related to selecting papers, extracting data and assessing the studies quality, and the most time-consuming tasks are the ones related to searching databases, choosing papers and extracting data. The authors' findings suggest the need for careful SLR planning, especially concerning scoping the research questions, and defining the inclusion/exclusion criteria. Also, their

study emphasizes the need for reviewing the whole planning (by experts) as well as taking advantage of teamwork to minimize bias and conflict resolution.

2.2 SLRs and Experts

Issues involving the use of SLRs in SE are not exclusive of students' participation. In 2009 Babar and Zhang performed an interview-based survey to identify the perceptions of research practitioners on conducting SLRs in the SE field (Babar and Zhang 2009). The authors selected 24 researchers identified as active practitioners in SLR executions from which 17 agreed to respond to their interview. Apart from the positive perceptions regarding the research method, the researchers reported some of the most challenging things in SLRs which included the effort involved in the whole process, the design of search strings, and the definition of research questions.

More aligned with the work of assessing the reliability of SLRs, MacDonell et al. (2010) investigated the consistency of the SLR process and the stability of its outcomes. Their study compared the results of two independent reviews (performed by groups with similar domain experience) undertaken with a common research question. In comparison to the work presented by Kitchenham et al. (2011), the reviewers have vast experience in the research topic (cross-company estimation models and within-company estimation models). Although the two groups conducted the SLR in different ways (search strings, review process), the findings were similar (from 11 primary studies identified by the two groups, nine were commonly identified). The main causes of the differences have been designated as: i) a lack of consensus on what constitutes a high-quality primary study, and; ii) misunderstandings as to what constitutes an appropriate response variable. The conclusion of the study indicates the robustness of SLR as a research method (considering groups with similar domain experience), although its repeatability can be compromised.

In a participant-observer case study, Kitchenham et al. (2012) performed a mapping study of unit testing and regression testing to investigate the completeness of general mapping studies. They compared it with other specific mapping studies, SLRs and an expert literature review aiming at investigating how well general mapping studies identify clusters of related studies and to what extent such clusters are complete. The authors identified differences between the general systematic mapping they performed and the expert literature review regarding included papers, showing their mapping study outperformed the expert review. They also found that in comparison to SLRs and more accurate mapping studies, general mappings can miss important and relevant works. During the comparison, the authors identified issues related to differences in the classification of selected studies between the literature reviews, and also inconsistencies in the selection of studies which led the authors to advise the use of clear explanations during the exclusion of studies.

Another work that compared the results from independent literature reviews is the one presented by Wohlin et al. (2013). The authors present a study about two systematic mapping studies on the same research topic aiming at evaluating their reliability. Although the two studies address the same research topic, significant differences were identified regarding the inclusion and the categorization of papers, indicating low similarity between them. Based on that, the paper presents four conjectures to be confirmed or rejected through future investigations: i) snowballing based on researcher expertise and knowledge of an area is more efficient than trying to find optimal search strings; ii) secondary studies will not find the same papers unless it is a study of a relatively narrow area with experts in the area conducting the study; iii)

secondary studies may come to the same general conclusions regarding an area even if the papers found are not the same, and; iv) secondary studies are not reliable per se; they rely heavily on the context of the secondary study.

In a more recent work, Hassler et al. presented a rank of barriers to the SLR process gathered from a community workshop (Hassler et al. 2014). Along with 37 composite obstacles to the SLR process, the authors also describe the impact of them on SLR methodology, researchers, authors, and consumers. Some of their findings share similarities with other previous studies, but new issues are also presented, such as the ones related to i) the presence of a sequential process for SLR instead of an iterative one; ii) the lack of support for interpretation and generalization of studies; iii) the misleading titles and abstracts; and iv) the lack of consistency of the SE terminology; among others.

The primary goal of this research is to characterize the reliability of SLRs by identifying similarities and differences in their processes and outcomes. Therefore, works such as (MacDonell et al. 2010; Kitchenham et al. 2011, 2012) and (Wohlin et al. 2013) are more closely related to the one presented in this paper. However, we decided not only to compare the included articles but also to compare search strings, inclusion/exclusion criteria, returned and excluded papers and also the outcome that was expected to answer the research question. Our expectation in applying this holistic view was to gather sources of comparison that would support us drawing a better conclusion on the points that make the SLR process more/less reliable. Also, we decided to provide instruments in our exploratory study to prevent the students from experiencing some of the difficulties previously mentioned. It allowed us to observe other challenges and pitfalls commonly faced by novices, as well as real problems with surveying evidence in the SE field that can be further used as a base to enhance the research on this topic.

3 The Exploratory Study Planning

Based on the previous discussions, this section presents the plan of our exploratory study on the reliability of SLR processes in the SE field. Detailed information on the materials and data collected during the study can be found in our study package available at http://lens-ese.cos.ufrj.br/appendices/EMSE/2016/StudyPackage.zip.

3.1 Goal

The research objective was set using the Goal-Question-Metric (GQM) template (Basili 1992), as Table 1 depicts.

Table 1 Research objective

Analyze	*quasi*-systematic literature reviews research protocols and reports
For the purpose of	characterization
With respect to	their SLR processes and outcomes similarities
From the point of view of	researchers and practitioners in software engineering
In the context of	Master and Doctoral students undertaking same goal SLRs as an assignment in the Experimental Software Engineering (ESE) course at UFRJ in the years of 2010 and 2012.

We intend to investigate the SLR process repeatability and outcome consistency based on the similarities and differences encountered in the research protocols and reports of seven SLRs dealing with the same research question and performed by similar teams of novice researchers (concerning mainly their inexperience in the research method).

3.2 Participants

The studies were executed during two years (2010 and 2012) in the ESE course at COPPE/UFRJ. The participants were graduate students (seven D.Sc. and 14 M.Sc.) in their first year of graduation (only taking disciplines at this period) in the System Engineering and Computer Science Program, and none of them had previous experience in the experimental topics taught in the course (Primary and Secondary Studies in SE), as can be seen in Fig. 1. The secondary study planning and execution were assignments given to the students – the main assignments used to grade the students in the module. We can highlight two main motivations for the students to participate in the study and be committed to it: i) first, many masters and doctorate students had expectations in executing an SLR in the context of their research (dissertations and thesis), which became true in many cases (see Table 2); ii) second, since the students were being marked on the assignment, they had to apply themselves in order not to fail the module. Otherwise, it could cost their standing in the graduation program.

We organized seven teams (three in 2010 and four in 2012) with three participants each aiming at reducing communication gaps and problems with the course commitment. Members of the Experimental Software Engineering Group at COPPE/UFRJ attending the course were grouped, and part-time participants were either placed in the same group or scattered consistently among the teams. Other characteristics such as the perceived knowledge (observed

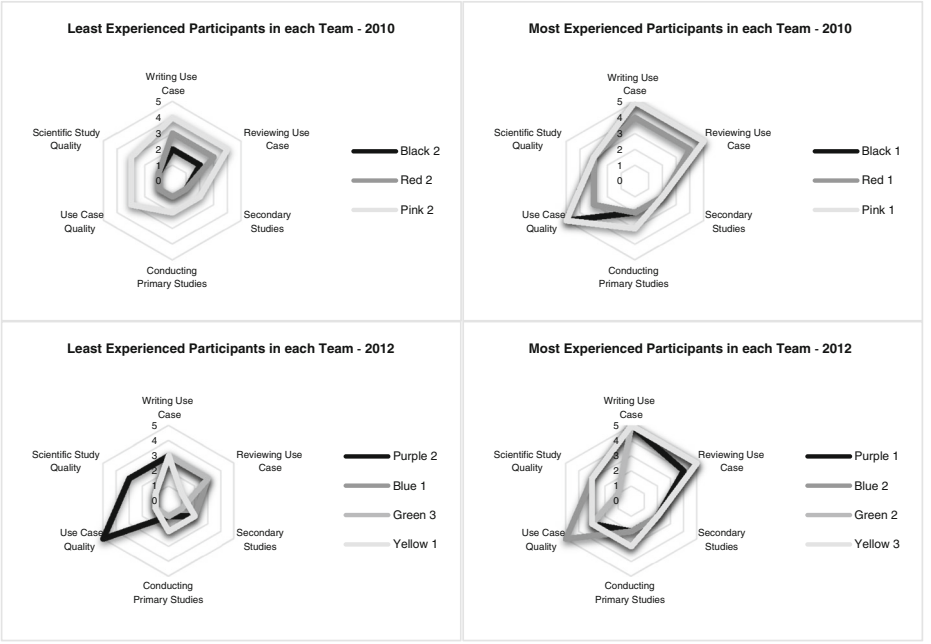

Fig. 1 Participants experience in the main topics related to the assignment

Table 2 Teams' names and characteristics

Year	Team' Name	# M.Sc. Students	# D.Sc. Students	ESE Group Members?	Part-time Students?	SLR Execution for their Research?[a]
2010	Black	2	1	Yes (1)	Yes (3)	Yes (1)
	Red	3	0	Yes (1)	Yes (1)	Yes (1)
	Pink	2	1	Yes (3)	No	Yes (2)
2012	Purple	1	2	No	Yes (1)	Yes (1)
	Blue	1	2	No	Yes (1)	Yes (1)
	Green	3	0	No	Yes (1)	No
	Yellow	2	1	Yes (2)	Yes (1)	Yes (2)

[a] Information gathered when searching for the students' articles, dissertation/thesis in the database of the graduation program

during the classes), declared knowledge (responses to a characterization form) of the topic under investigation – use cases – and academic experience were also used as drivers to organize the teams. For instance, no team was composed only of doctoral students, and all teams had participants with expertise in SE in practice (practitioners). It is important to emphasize that all participants had previous knowledge and, in some cases, expertise on use case descriptions either in academia or industry. Table 2 summarizes some of the characteristics of each team, including whether the participants conducted an SLR in the context of their research (see our study package for more information at http://lens-ese.cos.ufrj.br/appendices/EMSE/2016/StudyPackage.zip).

We are aware that ensuring two teams of researchers have similar knowledge and expertise (on the research method and topic) is laborious and subjective. Furthermore, the characteristics used to assess these features might not be sufficient to guarantee such assumption. Therefore, the participants were always grouped as much as possible guaranteeing similarities among the teams and ensuring a real commitment during the SLR planning and execution. The constraint of having at least one participant with experience in the software industry in each team was also a way to simulate a scenario in which a practitioner would perform an SLR.

3.3 Materials

To support this exploratory study, we prepared and used three materials: i) a consent form (written in Portuguese); ii) a characterization form (written in Portuguese), and iii) an initial research protocol (written in English). The students were not obligated to participate in the study, and for this reason, all of them received the consent form and were asked to sign it in the case of agreeing in taking part of the study. They knew they would be graded on the study and an alternative form of evaluation would be given to those that would not consent on participating in it. All of them signed the consent form.

After agreeing on the study, the participants filled in a characterization form, self-reporting their knowledge and experience (using a *Likert*-scale) in the following topics: English reading and comprehension, software development, requirements and use case, primary and secondary studies and quality appraisal of software artifacts and scientific papers. The stratification of the participants in different teams used this particular form.

The initial research protocol is the most important instrument of this study since it contains the main elements to guide the students on performing the SLR in the same topic. It contains the research question the teams should answer: "Which quality

attributes (and measurements used to evaluate such attributes) have been empirically studied for use cases?" The topic related to the use cases was suggested in the 2009 ISERN meeting, and it was used in our study because it is believed to be a grounded topic in the SE field in which the participants would have more knowledge and experience. Also, the participants would even feel more comfortable to work with it – which was the case of our study according to the characterization form responses.

Along with the research question, the following information was also provided in the initial research protocol:

(i) background information and perspectives of quality attributes regarding requirements specification, as presented in (Condori-Fernandez et al. 2009);
(ii) a request to extract from selected studies the approaches, templates or formats proposed to improve the use case quality;
(iii) some initial terms to support the search for studies;
(iv) definition of the search engines to be employed in the study – Scopus, Web of Science and IEEE Xplore;
(v) some initial criteria for the studies selection and evaluation, and;
(vi) a data extraction form suggestion.

The idea behind providing all this information to the students was to place them in the same perspective concerning the quality of use cases and also to prevent the main problems reported and highlighted in the related works. It is important to notice that, despite making available this information set, the teams should complete the SLR protocol and had the freedom to change some items, except for the research question.

3.4 Research Question and Assumptions on SLR Reliability

Driven by our main research question – do similar SLR protocols, executed by similar teams of novice researchers, lead to similar answers to the same research question? –, some behaviors concerning the SLR planning and outcome can be conjectured, as presented in Table 3.

Since SLRs provide a well-defined procedure to identify, analyze and interpret impartially and repetitively all kind of available evidence related to a specific research question (Biolchini et al. 2005) two of the behaviors presented in Table 1 are naturally expected to happen, especially considering the similarity of researchers' knowledge and experience executing the reviews. Yet, whether two SLR protocols are alike, and their execution (in terms of studies selection) and/or outcomes (in terms of answers to the research question) turn out to be

Table 3 Protocol and outcome similarities alternatives

		Outcome Similarity	
		High	Low
Protocol Similarity	High	Expected behavior	Low reliability in SLRs
	Low	High reliability in SLRs	Expected behavior

different, it might show that either some external factors influenced the selection of studies and the analysis of the results (e.g., existence of ambiguous information and various terminologies in the studies) or relevant information is missing from the research protocols. These issues hamper the SLR process repeatability and, thus, its reliability, as it might have been the cases reported in, (Kitchenham et al. 2011, 2012; Wohlin et al. 2013) and (Munir et al. 2014).

Conversely, if two SLR protocols are different and their outcomes turn out to be similar, this can reveal the existence of a similar terminology used to report the results and/or a similar researchers' point of view about the topic under investigation. The point of views can be expressed not exactly by the terms of the search strings and adopted selection criteria but by the intention of the search and selection of studies presented in these two elements (showing that some parts of the research protocol are particularly more important than others). It might have been the case reported in (MacDonell et al. 2010) concluding that SLRs are reliable since they can result in similar answers to the same research question even in the face of differences in their investigation processes.

3.5 Tasks and Procedures

All participants received equivalent lectures on SLRs and had about two months to execute the assignment and present the *quasi*-SLR results. The lectures involved topics related to primary and secondary studies in SE, and a secondary study planning and execution was one of the assignments given to the students – the main assignment used to grade the students in the module.

We asked the students to use the available guidelines for secondary studies executions to guide them in the planning and execution of their studies, and at any time they could ask questions on the SLR steps. Although the initial research protocol has been provided, the participants were free to fulfill it according to their understanding of the topic under investigation, as long as they would not modify the research question for the search engines (to not make us lose the baseline of comparison). To support the protocol refinement, especially the search string formulation using the Population-Intervention-Comparison-Outcome (PICO) strategy (Pai et al. 2004) we also advised them to identify control articles and to improve their search string based on them. As additional requests for the assignment, the students should use JabRef[1] for supporting the studies selection and data extraction, and should provide three main deliveries: i) the updated *quasi*-SLR plan; ii) the Bibtex featuring the studies selection and data extraction; iii) the complete *quasi*-SLR package which should include the final version of the research protocol, Bibtex, included/excluded papers and reports with the quality attributes for use cases extracted from the included articles.

3.6 Analysis Procedure

3.6.1 Research Protocol Similarity Analysis

We defined two similarity perspectives to analyze the agreement among the SLR plans: syntactic and semantic. In the syntactic perspective, we want to observe the similarity between

[1] http://www.jabref.org/

pairs of reviews regarding their search in the digital libraries. In the semantic perspective, we want to observe the similarity between pairs of reviews regarding participants' point of views about the research question. The following subsections present details about these two similarity perspectives.

Syntactic Perspective In this point of view, we want to observe the exact match between pairs of reviews in finding the same papers. To do so, we selected the Jaccard index – Eq. (1) (Jaccard 1912) – as a measure of similarity of two protocols (A and B) for two units of analysis: adopted search terms and papers returned in common, as described below:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (1)$$

In respect to the adopted terms in the search strings of two protocols (A and B), the Jaccard index expresses the portion of common terms between them (|A∩B|) in relation to the total of terms used in protocol A (|A|) plus the total of terms used in protocol B (|B|) excluding the common ones. To compare the similarity of two terms we had to apply some rules: i) we used the main search string (see Appendix 2) created by each team instead of using the three search strings tailored to each search engine; ii) we did not consider the use of quotation marks, that is, a term with or without them would be equivalent; iii) the terms were considered case insensitive; iv) we considered singular and plural forms (just the ones that add 's' at the end of a word) of a search term as equivalent; v) we considered the use of hyphens, that is, a search term with hyphen would be regarded as different from its similar without the hyphen. Since these rules simplify the search terms comparison, and the logics of the search strings were not considered in the described similarity calculation, it is wise to analyze the similarity of the adopted search terms along with the similarity of the returned papers, decreasing the threats to the validity of the syntactic perspective analysis.

Regarding the papers returned in common, the aim is to identify the portion of papers returned by both search strings (|A∩B|) in relation to all returned papers by the pair A (|A|) and B (|B|) (not considering duplicates – |A∩B|). As we will compare SLRs executed in different years, only returned papers up to 2010 should be discussed in the comparison of pairs of protocols from distinct years.

The analysis of the values distribution for each unit of analysis will support the identification of slight similarities (below the first quartile of the distribution) and almost perfect similarity (above upper third quartile of the distribution). A complete similarity is represented by 1.0.

Semantic Perspective In this point of view, we want to observe the similarities among the teams' intentions in searching and accepting the same papers, that is, their point of view about the research question. The units of analysis, in this case, are the main concepts embedded in the search string terms; the paper inclusion and exclusion criteria; and the included and excluded papers. Similarly to the syntactic perspective analysis, the semantic perspective one should take into consideration all mentioned units of analysis to support more reliable conclusions regarding the teams' similar/different intentions in searching and accepting the same papers.

To identify the main concepts (Appendix 3) embedded in the search strings, we needed to abstract them from the terms used in each research protocol by applying a coding technique

similar to the open coding provided by Grounded Theory (Corbin and Strauss 2007). One of the authors assembled and sorted alphabetically all the terms used in the seven search strings (Appendix 2), ignoring the logical structure of the search strings and aggregating the same name terms (by applying the rules mentioned in the previous subsection). Next, during a three-hour session, the three authors got together to identify the main concept of each term. For each of the 366 different search terms, the authors assessed its meaning based on the semantics of its words in the SE field and assigned a concept to it. Whenever a new concept was identified, we compared it to the existing concepts, avoiding the creation of different concepts with the same meaning. Overall 23 different concepts were identified and the same Jaccard index – Eq. (1) – could be used to measure the semantic similarity among teams in means of the concepts abstracted from their adopted search terms.

Concerning the paper inclusion and exclusion criteria, the semantic similarity can also be measured using the Jaccard index by checking the proportion of inclusion and exclusion criteria each pair of research protocols share. The last unit of analysis (included and excluded papers) is the one that relates the most to the teams' points of view about the research question, and it can be observed in the light of the teams' agreements and disagreements in including/excluding papers for data extraction. The Kappa coefficient – Eq. (2) (Cohen 1960) – can support the measurement of this feature, once it is used to measure the agreement in qualitative evaluations among different raters. In subjective interpretations, two observers will sometimes agree or disagree by chance; once no objective criterion is stated (Viera and Garrett 2005). Kappa coefficient intends to calculate the qualitative agreement among raters subtracting the probability the agreement might have happened by chance. To do so, it takes into account the relative agreement of raters (two teams in our case) in each of the analyzed categories (included and excluded papers in our case – qualitative perspective) – *po* – and the probability the agreement has happened by chance – *pe*, subtracting *pe* of *po*, as follows:

$$K = \frac{(po - pe)}{(1 - pe)} = 1 - \frac{(1 - po)}{(1 - pe)} \tag{2}$$

Although we could have used the Jaccard index presented previously to characterize the agreement on the inclusion and exclusion of papers, the Kappa coefficient is more robust to measure the agreement when making a qualitative evaluation, since it does not consider the agreement by chance – detailed information in (Viera and Garrett 2005). The consideration of papers published only up to 2010 for comparison of pairs of protocols from different years is also required in this case.

We used the work by Vieira and Garrett to identify the level of agreement for the obtained Kappa values, once it is commonly applied for this purpose (Viera and Garrett 2005). The confidence interval used was 95%. A perfect agreement is represented by 1.0.

In the end, two *quasi*-SLR protocols are considered similar if their syntactic and semantic perspectives have almost perfect similarity and agreement. Slight similarity and agreement emphasize that the protocols are quite different.

3.6.2 Outcomes Similarity Analysis

We can analyze the answers to the research question the teams can provide – quality attributes for use cases – upon two perspectives: the answers are correct, complete and consistent among

the seven SLRs, or they are incorrect, incomplete and inconsistent, and need to be revised and detailed so we can identify their actual match among the reviews, and, thus, perform the similarity analysis. In this regard, we decided to consider all answers as correct, complete and consistent. Otherwise, we would have to go through all the included papers from the seven SLRs and revise the teams' data extractions, which would result in a comparison of the authors' answers, not the teams' ones. Thus, we compared the quality attributes according to their syntax only; assuming whether the syntax is equivalent so is the meaning. We understand that this decision can make us overlook some answers particularities that would prevent us from matching the same name attributes with different meanings reported in various reviews; or even would make us match different name attributes with similar meanings. However, we made such decision to avoid biasing the *quasi*-SLR outcomes and teams' perspectives. The similarity of two answers (sets of quality attributes) is then calculated using the Jaccard index as we did for the search terms, returned papers, main concepts, and inclusion/exclusion criteria.

It is important to stress that even though the strategy used for extracting the main concepts from search terms can also be applied to the quality attributes for the use case, it would not result in a diversified group of concepts as happened previously since the scope, in this case, is narrower when compared to the adopted search terms.

4 Study Results

In this section, we report the similarities and agreements observed among the seven *quasi*-SLR research protocols and outcomes, highlighting some pitfalls (underlined throughout the section) that support the explanation for the observed divergent results. Figure 2 presents an overview of the teams' quantitative results divided by the protocols and outcomes similarity analysis perspectives – which will guide the report of this study results. As one can see, the selected search terms ranged from 11 terms used by the Black team and 215 terms by the Purple, while the returned papers ranged from 157 papers came back in the Pink search and 661 in the Black search. These different outcomes foretell the findings presented in this section.

4.1 Same Research Question and Different Protocols: Syntactic Perspective Analysis

Since the same research question has been addressed (with minor differences as it can be seen in Appendix 1) and an initial protocol was given to all the teams to ground their knowledge in the research topic and method, we expected some similarity among the research protocols. Surprisingly, we could not observe this behavior. Table 4 presents the Jaccard index (expressed in percentage) for the terms used in each pair of search strings.

Although the pairs Red-Green, and Pink-Yellow present the highest similarity index for terms in the search string, it is rather naive to assume any similarity given the value slightly above of 18% for their common terms. In the Pink-Yellow pair, the teams' characteristics might help us to explain this proximity when compared to the other teams. ESE group members mainly composed both the Pink and Yellow teams. In this case, almost a third (six out of 19) of similarities lie in the terms they usually used to search for empirical studies in SE – terms that they were more familiarized than the other participants due to their daily research activities in their research group.

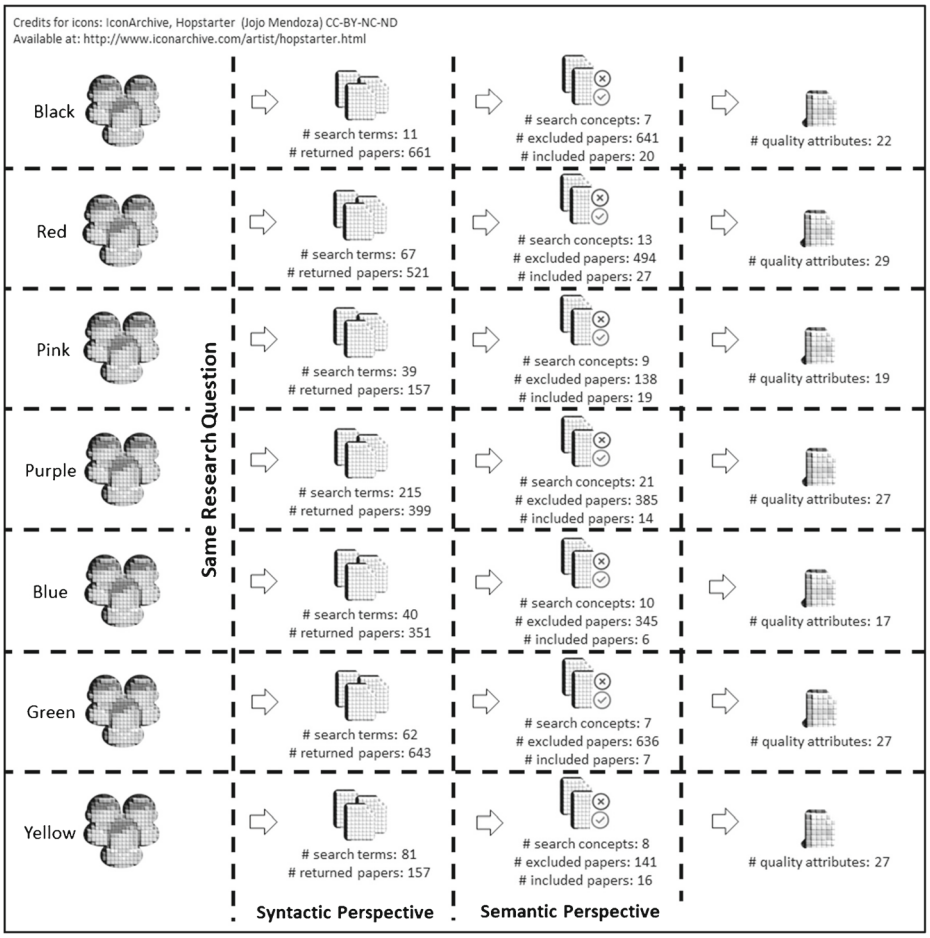

Fig. 2 Summary of teams' quantitative results

The Blue had noticeable divergence with almost all other teams (Fig. 3). A detailed analysis of the terms used in its search string (Appendix 2) reveals that the participants had preferred to use general over specific terms (keywords too high-level). For instance, instead of "quality attributes" and "quality characteristics" chosen by other teams, the Blue team decided for using

Table 4 Syntactic Perspective - Jaccard index (in %) for terms used in common between each SLR pair

	Red	Pink	Purple	Blue	Green	Yellow
Black	4.00[b]	5.88	1.35[b]	4.08	5.80	3.37[b]
Red	–	15.79[a]	7.22	8.08	18.35[a]	8.03
Pink	–	–	6.61	15.28[a]	12.90[a]	18.10[a]
Purple	–	–	–	3.24[b]	9.49	9.63
Blue	–	–	–	–	5.15	3.42[b]
Green	–	–	–	–	–	4.38

[a] Values above the third quartile of the distribution

[b] Values below the first quartile of the distribution

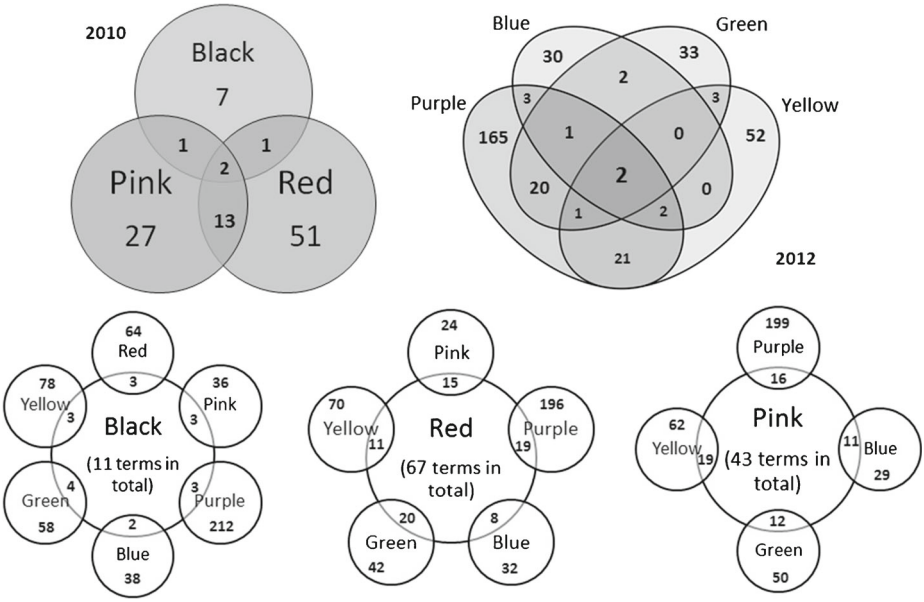

Fig. 3 Summary of the amount of terms in common among the teams

"attribute," "characteristic" and "quality" in its search. Although its quest would also return the papers that present "quality attributes" or "quality characteristics," we did not consider its terms as equal to the others during the comparison of the terms, since the other teams would not find the same papers returned by the Blue's search. An interesting observation on the identified search terms is that all teams but the Red chose to use words with no impact in the searches (unnecessary search terms); that is, they included plural terms and their singular versions, or compound terms that were already covered by simpler terms previously identified. As an example of these cases, some teams used "use case" and "use cases" in the same search string, or even "description template" and "template," among other examples.

Overall, the teams identified 366 distinct terms: no term was mentioned in all the seven search strings, three ("software development", "consistency" and "understandability") were used by six teams, eight ("system development", "use case", "quality characteristic", "quality factor", "quality feature", "completeness", "correctness", and "efficiency") by five teams, and three ("case study", "software project", and "quality attribute") by four teams. The remaining terms were used by less than half of the teams. From the search strings (see Appendix 3) we could notice that many teams tried to maximize the number of terms combination, disregarding whether they were valid search terms (many different combinations of terms producing noise return). Also, some terms had no relation to the research question, such as "testwarehouse", "program method", and "degree of functional encapsulation", not to mention other not typical terms for search, such as "desirable quality", "mistake free", and "wholeness" (inappropriate selection of search terms), stressing the difficulties in creating a search string to meet a research purpose.

As explained in Section 3.6.1, the analysis of similarities through the syntactic perspective should be done considering not only the terms used for the search but also the returned papers to take into account the logical expression of the search strings. The syntactic similarities of the returned papers were even lower than the similarity registered for the terms (Table 5).

Table 5 Syntactic Perspective - Jaccard index (in %) for papers returned in common between each SLR pair

	Red	Pink	Purple	Blue	Green	Yellow
Black	5.07	2.76	1.93[b]	1.69[b]	1.42[b]	2.33[b]
Red	–	12.25[a]	4.73	5.98	4.86	8.14[a]
Pink	–	–	5.27	7.86[a]	6.57	14.80[a]
Purple	–	–	–	1.76[b]	3.07	9.66[a]
Blue	–	–	–	–	2.79	2.42
Green	–	–	–	–	–	2.83

[a] Values above the third quartile of the distribution

[b] Values below the first quartile of the distribution

Aside from the differences in the logic of search strings (overuse of 'and' operators and distributive properties), common terms previously identified by the teams were differently organized in the search strings, even though all of the participants were instructed to use the PICO strategy (Pai et al. 2004) (misuse of the guidelines). Another observed detail regards how differently the teams configured the search engines, bounding the areas that should be excluded from the search (see Appendix 4). These differences have certainly affected the results provided by the search engines, explaining the lower percentages in Table 5.

As we can observe, no noticeable similarity can be seen from the syntactic perspective even though the same research question and an initial research protocol were given to the teams.

4.2 Same Research Question and Different Protocols: Semantic Perspective Analysis

The semantic perspective can help us to understand whether the teams' point of views have influenced the differences in the syntactic perspective and whether similar findings can be observed in the existing, though low, similarity. Out of the 366 distinct terms used in all seven *quasi*-SLR search strings, 23 main concepts were identified through the coding process: defect rate; evaluation; environment; general quality issue; product; project; quality features; requirement documents; requirement models; requirement representation policy; rework rate; scenario; scenario documents; scenario models; software life cycle; software technology; use case; use case concepts; use case documents; use case models; use case representation policy; user story documents; and user story models.

As examples of the coding process, terms such as "defect fee," "defect rate," "defect ratio," "error rate," "fault rate," and "mistake rate" were grouped in the main concept "defect rate." More specific terms such as "ambiguity," "clarity," "completeness," "comprehensibility," "concise," "correctness," "readability," "traceable," "understandability," "usability" were grouped in the main concept "quality features." To group terms such as "application method," "development approach" and "software technique" we used "software technology." Appendix 3 presents the complete list of concepts, their meaning and the respective terms that generated them.

Not all research protocols reported terms related to every main concept, although the similarity of concepts was a lot higher than the terms (Table 6). Out of the 23 concepts, two ("general quality issue" and "software life cycle") are present in all the seven strings. However, eight ("environment," "rework rate," "scenario documents," "scenario models," "use case concepts," "use case representation policy," "user story documents" and "user story models") are present in either the Red team ("use case concepts") or the Purple team (the other seven concepts). It made the Purple team, along with the Yellow team, present the worst similarity indexes for the concepts as depicted in Table 6.

Table 6 Semantic Perspective - Jaccard index (in %) for main concepts used in common between each SLR pair

	Red	Pink	Purple	Blue	Green	Yellow
Black	42.86	60.00[a]	27.27[b]	54.55	55.56	25.00[b]
Red	–	69.23[a]	47.83	76.92[a]	53.85	31.25
Pink	–	–	36.36	72.73[a]	60.00[a]	30.77[b]
Purple	–	–	–	40.91	27.27[b]	38.10
Blue	–	–	–	–	54.55	20.00[b]
Green	–	–	–	–	–	36.36

[a] Values above the third quartile of the distribution

[b] Values below the first quartile of the distribution

Upon these better results regarding the similarities of the main concepts, we had the expectation that the papers returned in common would be evaluated using a similar perspective, showing that even though the different searches did not retrieve the same papers, the teams had the intention to do so. While the analysis of the inclusion and exclusion criteria similarity leads to better results in comparison to previous similarity analysis (Table 7), a closer look at the actual matches among the teams highlights that they barely changed the inclusion and exclusion criteria given in the initial protocol (Appendices F and G). Furthermore, they diverged on papers they should include or exclude in many cases. As an example, an initial comparison among all seven *quasi*-SLRs revealed that from 2167 articles (up to 2010) only two papers were returned in common. One of these two articles (Losavio et al. 2004) was unanimously excluded because it does not relate to use case quality attributes, but to software architecture design. The other paper (Ramos et al. 2009), though, led to different decisions among the teams: four teams included the paper for data extraction (Black, Red, Pink, and Green), while three excluded it (Purple, Blue, and Yellow). Analyzing the paper, we could observe that the empirical study it presents might have caused the divergence among the decisions since its authors labeled the study as a case study, but the study description indicates to be a proof of concept (misunderstandings concerning empirical/experimental study strategies).

According to the research question, the quality attributes for use cases should be empirically studied to avoid reporting speculative attributes. Some teams (Pink, Purple, and Yellow) explicitly excluded papers (exclusion criteria – Appendix 7) that did not present any empirical study, or that presented either a toy example or a proof of concept, considering they would provide unreliable quality attributes. Some interesting issues were observed regarding it while

Table 7 Semantic Perspective - Jaccard index (in %) for inclusion/exclusion criteria used in common between each SLR pair

	Red	Pink	Purple	Blue	Green	Yellow
Black	16.67[b]	27.27[b]	30.77	16.67[b]	30.00[b]	15.38[b]
Red	–	63.64	50.00	63.64	70.00[a]	58.33
Pink	–	–	61.54	63.64	70.00[a]	72.73[a]
Purple	–	–	–	50.00	66.67	46.67
Blue	–	–	–	–	70.00[a]	72.73[a]
Green	–	–	–	–	–	63.64

[a] Values above the third quartile of the distribution

[b] Values below the first quartile of the distribution

analyzing the teams' reports and BibTeX (available in the study package). They show different expectations concerning the empirical aspect of the quality attributes (no explanation concerning the empirical focus used):

(i) The teams Black, Red, Pink, and Green included (Ramos et al. 2009) for evaluation considering it was a case study (as labeled by the paper's authors);
(ii) The Purple and Blue at first included (Ramos et al. 2009), but afterward excluded it. No explanation for the exclusion was provided (tacit knowledge regarding the study selection strategy).
(iii) The Yellow excluded (Ramos et al. 2009), considering that no study was described regarding quality attributes for the use case.

Table 8 presents the Kappa agreement on including and excluding papers in each pair of SLRs.

The negative values indicate that the probability of teams agreeing by chance while including and excluding papers is higher than their relative agreement. In the two particular cases in Table 8, neither the pair Purple-Green nor the pair Blue-Green had any common paper included. Figure 4 shows that most of the agreements in the pair Blue-Green lied in the papers they excluded in common (22). No paper (zero) was included in common by the teams, although both had included different papers they found in common: one paper was included only by Blue team and four papers were included only by Green team (see intersection).

Analyzing specifically the included papers (underlined) in the intersection of the pair Blue-Green (Fig. 4), we can observe that Green team included papers not completely related to the research question (misinterpretation of the research question). Two of the included papers – (Preiss et al. 2001) and (Rago et al. 2013) – were not about use cases quality attributes, but about quality characteristics expected for a software according to its requirements specifications (controversial understanding on the research topic). The first paper intends to use these features as the basis for a software development, while the second one intends to extract them from the specifications through mining. The only paper included by Blue team in the intersection – (Fantechi et al. 2002) – was indeed related to use cases quality attributes. The teams did not have the same perspective about the research question, as reinforced by the negative coefficient presented in Table 8; neither they had similar inclusion and exclusion criteria, although their similarity in Table 7 says the contrary (tacit knowledge regarding the study selection strategy).

Overall, the teams had a higher agreement regarding the semantic perspective when compared to the syntactic perspective. However, we did not find any reasonable explanation

Table 8 Semantic Perspective - Kappa coefficient (in %) for papers included and excluded between each SLR pair

	Red	Pink	Purple	Blue	Green	Yellow
Black	65.30	81.82[a]	51.90	33.33	100.0[a]	44.44
Red	–	47.07	42.77	45.22	72.97	39.23
Pink	–	–	56.14	53.23	72.67	67.02
Purple	–	–	–	26.15	−3.33[b]	73.42
Blue	–	–	–	–	−6.30[b]	100.0[a]
Green	–	–	–	–	–	62.07

[a] Values representing high agreement

[b] Values representing low agreement

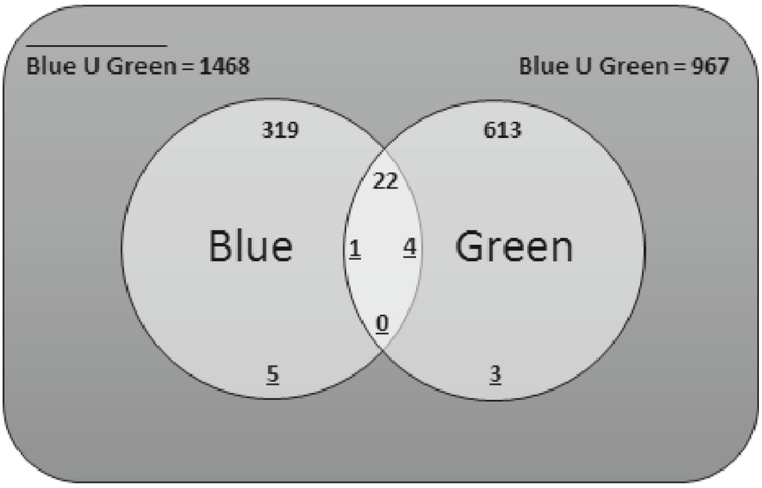

Fig. 4 Amount of papers returned, included (underlined), and excluded from each team in the pair Blue-Green

for this behavior because they barely changed the inclusion and exclusion criteria given in the initial protocol, and no other information could be obtained from their research protocols in order to support the understanding of such results. Table 9 summarizes the information concerning the returned and included papers in common per pair of teams, providing an overview of the findings and corroborating the previous results.

Table 9 Returned and included papers in common per pair of teams

			Red	Pink	Purple	Blue	Green	Yellow
Black	Returned papers in common		57	22	19	15	17	18
	Included Papers	total in the intersection	20	12	8	6	3	14
		# common	12	10	4	2	3	9
		% common	60.00	83.33	50.00	33.33	100.00	64.29
Red	Returned papers in common		–	74	39	43	50	49
	Included Papers	total in the intersection	–	18	11	6	5	17
		# common	–	7	4	2	3	6
		% common	–	38.89	36.36	33.33	60.00	35.29
Pink	Returned papers in common		–	–	25	29	44	37
	Included Papers	total in the intersection	–	–	11	7	5	13
		# common	–	–	6	3	3	8
		% common	–	–	54.54	42.86	60.00	61.54
Purple	Returned papers in common		–	–	–	13	31	49
	Included Papers	total in the intersection	–	–	–	4	2	8
		# common	–	–	–	1	0	5
		% common	–	–	–	25.00	0.00	62.50
Blue	Returned papers in common		–	–	–	–	27	12
	Included Papers	total in the intersection	–	–	–	–	5	2
		# common	–	–	–	–	0	2
		% common	–	–	–	–	0.00	100.00
Green	Returned papers in common		–	–	–	–	–	22
	Included Papers	total in the intersection	–	–	–	–	–	4
		# common	–	–	–	–	–	2
		% common	–	–	–	–	–	50.00

4.3 Same Studies and Different Outcomes

Each one of the seven teams elaborated a list of quality attributes for use cases. Thus, 83 distinct quality attributes (complete list in Appendix 11) were extracted from the seven *quasi*-SLRs. From this total, 29 (~30%) were presented in at least two lists, and just five quality attributes (*consistency, correctness, completeness, readability, and understandability*) were presented in all seven lists. In addition to the overall analysis involving the seven *quasi*-SLRs, we also compared their lists in pairs. Table 10 summarizes the percentage of quality attributes found in common between each pair. It is important to observe that the comparison among the teams Black, Red, and Pink, and Purple, Blue, Green, and Yellow was accomplished considering quality attributes found exclusively in the papers published up to 2010.

The percentages presented in Table 10 indicate a low level of agreement regarding the quality attributes for use cases, once no pair of teams could identify at least 50% of quality attributes in common. This fact can be partially explained by the low level of agreement regarding the papers selected by each team, that is, in most cases, the SLR teams analyzed different papers.

It is possible to accomplish an additional analysis: what is the level of similarity considering only the papers included in common by two teams? Therefore, the final report of each team was analyzed to extract: i) the papers included in common by two teams, and; ii) the quality attributes identified in each one of these common papers. Regarding item (ii), unfortunately, the Red and Pink teams did not report the quality attributes per paper (imprecise reports), which made it impossible to compare the findings of these two teams with the others. Thus, the comparison was accomplished between each pair composed of the teams Black, Purple, Blue, Green, and Yellow. The terms were compared by their syntax, not exactly by their meanings, since not all the reports presented a complete and detailed information on the attributes gathered from the selected studies (incomplete reports).

Table 11 summarizes the number of papers included in common, the total amount of quality attributes for use cases identified in these papers and the number and percentage of the quality attributes determined in common by each pair of teams.

Again, the percentages highlighted in Table 11 indicate the low level of agreement between each pair of teams regarding quality attributes for use cases, even when these teams analyzed the same set of papers. In the worst case, the Green team had no paper in common with Purple and Blue and just two quality attributes in common with Black and Yellow. A thorough investigation of the quality attributes extracted by Green revealed that most of these attributes are not related to use cases, such as accessibility, the complexity of source code, safety, pluggability, portability and support for parallel development, among others (report of

Table 10 Jaccard index (in %) for quality attributes found in common between each SLR pair

	Red	Pink	Purple	Blue	Green	Yellow
Black	34.2	46.4[a]	36.1	39.3[a]	19.5	28.9
Red	–	37.1[a]	33.3	31.4	21.7	19.1[b]
Pink	–	–	39.4[a]	44.0[a]	17.9[b]	31.4
Purple	–	–	–	29.4	20.0	22.7
Blue	–	–	–	–	18.9[b]	18.9[b]
Green	–	–	–	–	–	10.2[b]

[a] Values above the third quartile of the distribution

[b] Values below the first quartile of the distribution

Table 11 Quality attributes in common per pair of teams

			Purple	Blue	Green	Yellow
Black	Included papers in common		4	2	3	9
	Quality	total	12	13	10	21
	Attributes	# common	10	4	2	9
		% common	83.3	30.8	20.0	42.9
Purple	Included papers in common		–	1	0	5
	Quality	total	–	4	0	16
	Attributes	# common	–	2	0	8
		% common	–	50.0	0.0	50.0
Blue	Included papers in common		–	–	0	2
	Quality	total	–	–	0	14
	Attributes	# common	–	–	0	2
		% common	–	–	0.0	14.3
Green	Included papers in common		–	–	–	2
	Quality	total	–	–	–	11
	Attributes	# common	–	–	–	2
		% common	–	–	–	18.2

information not related to the research topic). The Green team also presented some great divergences with other teams during the semantic analysis (Table 8 and Fig. 4), and these last results acknowledge their difficulties in interpreting the research question and the research protocol itself (misinterpretation of the research question). Interesting enough, Green did not seem to have problems with the search string elaboration (next section), different from other groups.

It was also observed that a specific paper reported the "7C's of communicability" (Phalp et al. 2007) as a group of attributes related to use cases quality (*coverage*, *cogent*, *coherent*, *consistent abstraction*, *consistent structure*, *consistent grammar* and *consideration of alternatives*). The teams that selected this paper extracted and reported the seven attributes individually, but the Blue team extracted just one attribute (*communicability*) from the same paper, that is, *communicability* was used as a surrogate for the seven other attributes. We are aware that there is no rule concerning the way the studies should be synthesized. However, explanations regarding the perspectives used for data extraction and synthesis are necessary to allow the study understanding and replication (subjectivity of the research synthesis strategy).

Moreover, it is also possible to observe that the Black and Purple teams had high convergence degree regarding the quality attributes for use cases when they analyzed the same set of papers. However, the data collected from their protocols do not allow us to conjecture why this convergence came up, since the levels of syntactic and semantic agreement between these teams are low, and the sets of selected papers are quite divergent.

4.4 Study Conclusion

The results presented in sections 4.1 to 4.3 show that the same research question led to different protocols, considering the syntactic and semantic perspectives, as well as at various outcomes. These results indicate that similar groups of novice researchers can elaborate distinct protocols and, consequently, obtain different outcomes when trying to answer the same research question. As this conclusion affects the reliability of SLRs conducted by novices, some of the pitfalls presented in this section can also be experienced by practitioners that have never undertaken an SLR before.

As mentioned previously, we understand that practitioners can be seen as more experienced than novice researchers concerning SE topics, which might prevent them from generating

some of the mentioned pitfalls. However, the differences in SE terminology adopted in the industry and academia can bring some difficulties to practitioners even in this regard, which forces us to discuss challenges on surveying evidence in SE as a way of making this research tool more feasible/reliable for both researchers and practitioners.

Prior to discussing the challenges of SLR planning and execution in SE, the next section presents the quality assessment we performed on the seven *quasi*-SLR protocols and reports to identify additional issues the novices had that might led them to give the divergent results herein presented.

5 *quasi*-SLR Research Protocols and Report Grades: Quality Assessment

The differing results led us to question about the quality of SLR search protocols and reports. Therefore, we decided to assess them using two different strategies: i) reviewing each research protocol and report based on a set of criteria adapted from the Database of Abstracts of Reviews of Effects (DARE) criteria (NHS Centre For Review And Dissemination 2002) and ii) calculating the precision and recall (Diest et al. 2009) of each search string.

5.1 Assessing the Protocols and Reports through a DARE Criteria Adaptation

We adapted the set of criteria from DARE – which is used to evaluate SLRs in the medical field – to support the team's assignment evaluation. DARE provides five questions related to the existence and/or the quality of inclusion/exclusion criteria, search for evidence, selected results assessment, selected results details, and selected results synthesis. Having these as inspiration, we created a scoring (ranging from 0 to 10) to suit the context of the given assignment which consisted of: i) checking whether the teams did not change the initial protocol (5), changed it to be better (10) or to be worse (0) in the case of the planning; ii) checking whether they applied their planning; and; iii) checking whether their final report has a reasonable level of detail. Assessing the completeness of their planning and report would require an oracle protocol and an oracle report, which were not the case, and for this reason, we decided not to use this perspective for the assessment. Table 12 presents the criteria utilized for the protocols assessment along with reasonable judgments concerning each criterion and their respective scores for the assignment.

The three authors assessed the seven research protocols individually according to the criteria above, using the mean to assign a protocol score and taking notes on the issues whenever necessary. Each final team score was given by the average of all three evaluations. Table 13 presents the final score of each team concerning their research protocol, and it also includes the main pitfalls identified during the assessment.

Along with the evaluation of the research protocols, we accomplished the evaluation of the reports. Differently from the protocol, for the case of the reports, we had no baseline for comparison, so we based our assessment on the amount of useful and understandable information their report provided. Table 14 presents the criteria used for the report's assessment along with possible judgments concerning each criterion and their respective scores for assignment.

Similarly, the authors assessed each report individually and assigned their scores and comments based on the criteria above. The same calculation used for the protocols final scores were performed for the reports. Table 15 details the assessment results.

Table 12 DARE criteria adapted to assessing the research protocols

Protocol Evaluation Criteria	Judgment Description	Assigned Score
Research Question	The Research Question was not changed	5
	The Research Question was changed for better	10
	The Research Question was changed for worse	0
Search Strings	The Search String was not structured using PICO strategy	0
	The Search String was incompletely/incorrectly structured using PICO strategy	5
	The Search String was completely/correctly structured using PICO strategy	10
Control Articles	Control Articles were not identified	0
	Control Articles were identified, but they are not all correct control articles	5
	Control Articles were identified, and they are all correct control articles	10
Inclusion Criteria	The Inclusion Criteria were not changed	5
	The Inclusion Criteria were changed for better	10
	The Inclusion Criteria were changed for worse	0
Exclusion Criteria	The Exclusion Criteria were not changed	5
	The Exclusion Criteria were changed for better	10
	The Exclusion Criteria were changed for worse	0
Study Selection Strategy	A Study Selection Strategy was not identified	0
	A Study Selection Strategy was identified, but it is not complete/adequate	5
	A Study Selection Strategy was identified, and it is complete/adequate	10
Quality Assessment Criteria	The Quality Assessment Criteria were not changed	5
	The Quality Assessment Criteria were changed for better	10
	The Quality Assessment Criteria were changed for worse	0

Table 13 Research protocols scores

	Score	Main Issues Identified During the Assessment
Black	2.92	The team simplified the first protocol in each evaluated criterium by reducing the scope of the research question, deleting some inclusion and exclusion criteria, some fields from the extraction form, and the entire quality assessment. Also, some terms of the search string were mixed through the different perspectives PICO, and their control articles were not in the selected search engines.
Red	4.42	The team did not change the initial protocol much, and although they used the PICO strategy, they mixed search terms from different perspectives. This team also identified control articles that were not in the selected search engines.
Pink	6.66	The team improved the initial research protocol a lot, expanding inclusion and exclusion criteria, and extraction and quality assessment forms. However, its search string is very confusing with lots of "AND" operations, and no control article was identified.
Purple	5.83	The biggest problem with the protocol is its inconsistency. Every time some information about the search string is given, it is presented differently. Also, it did not follow the PICO strategy. Even though the team expanded the inclusion and exclusion, they added some particular exclusion criteria. Additionally, one of its control articles is not a paper we (authors) selected for answering the research question.
Blue	6.25	The PICO strategy has not been fully followed, and although the team expanded the exclusion criteria, they are not too much different from the initial protocol. As an interesting note, the team selected six papers overall at the end, being four of them control articles.
Green	6.25	The team is among the few that used the PICO strategy correctly. However, the only selected control article was not among the ones we (authors) selected for answering the research question. The students did expand the inclusion and exclusion criteria.
Yellow	7.38	This team also used the PICO strategy correctly and presented a lot of different control articles. Its selection strategy is very detailed, and many new fields were created to report the studies. The drawback in the planning concerns the quality assessment that they changed completely, deleting some necessary items defined previously.

Table 14 DARE criteria adapted to assessing the reports

Protocol Evaluation Criteria	Judgment Description	Assigned Score
Studies Synthesis	The included studies were not synthesized	0
	The included studies were synthesized, but they are not adequate	5
	The included studies were synthesized, and they are adequate	10
Information Detailed	No information was detailed concerning the included studies	0
	Not enough information was detailed concerning the included studies	5
	Enough information was detailed concerning the included studies	10
Quality Assessment Use	The quality assessment criteria were not used	0
	The quality assessment criteria were used but not completely/correctly	5
	The quality assessment criteria were used completely/correctly	10

As one can see from the protocols and reports scores, there is a tendency that low scores protocols could lead to low score reports. Likewise, high scores protocols could result in high score reports. This type of assessment does not consider the correctness of the results because there should be an oracle for protocol and report comparison. Thus, it evaluates whether the relevant information is presented and whether they are detailed enough for further analysis and/or aggregation. We did try to make some adaptations on DARE to fit our needs, however, to assess the correctness of the SLRs we went for a more appropriate analysis: the precision and recall.

5.2 Evaluating the Searches through Precision and Recall Analysis

To accomplish the precision and recall of each search strategy we first needed to have a baseline of the relevant papers that can answer the research question. To do so, the three authors had to go through all the 2435 papers returned by the seven *quasi*-SLRs and evaluate them according to their perspective on the research question, taking two issues into consideration: to guarantee a common empirical study focus (experimental and empirical studies would be accepted); and a common use case quality focus (use case quality can be observed while constructing use case diagrams and descriptions and while inspecting either of them).

Table 15 Report scores

	Score	Main Issues Identified During the Assessment
Black	3.89	Everything was simplified. There is little definition concerning the quality attributes. No quality assessment was made.
Red	4.44	Descriptions of the studies show partial information. It is the only group that planned, executed and used the quality assessment result to rank the quality attributes.
Pink	7.22	It presented the selected studies in a very detailed way. Although the team planned and executed the quality assessment, they did not use it for anything other than list the quality of the selected paper.
Purple	6.66	It presented very detailed information regarding the studies, and similar to others, the team performed the quality assessment, but it was not used for anything.
Blue	6.11	A very detailed report. No use for the quality assessment, however.
Green	5.00	The summary of the papers is not much detailed. The quality assessment was not performed.
Yellow	7.22	It did present detailed information on the studies, but there is no actual use for the quality results of the evaluation.

For the selection strategy, we followed these two steps:

1. Each author read the title and abstract of the papers, evaluating them according to his/her understanding of the research question and the inclusion/exclusion criteria described in the initial protocol. Each paper was rated as I (Include) or E (Exclude);

   a. A consensus would define the final status of the paper;
   b. Whenever two authors decided for the exclusion of the paper, the paper would be excluded;
   c. Whenever two authors decided for the inclusion of a paper and the remaining author for its exclusion, the paper should be marked for a double check analysis.

2. Papers marked for second check analysis should be evaluated once more, now after reading the paper (not necessarily a full reading, though).

   a. The majority of the decisions would define the final status of the paper.

From the 2435 papers, we (authors) agreed to include 32 papers (29 up to 2010, and three from 2011 to 2012 – Appendix 10) and exclude 2318 papers, which means an agreement of 96.5%. The remaining 85 papers did not achieve a consensus regarding inclusion or exclusion, so they were excluded. This result allowed us to define the precision and recall of each search (Diest et al. 2009), using the papers we selected as the universe of relevant papers for answering the research question (*quasi*-golden standard (Zhang et al. 2011)). Table 16 presents the precision and recall of each search, separating the papers up to 2010 from those up to 2012.

Comparing the results from the previous subsection with the precision and recall of the searches, we could notice that the best protocols and reports also presented the highest precision and recall. However, low score regarding protocol and report did not directly lead to low precision and recall and vice-versa. For instance, we expected that Black would have presented very low precision and recall, and Green would have presented very low protocol and report scores, which is not true in comparison to other teams. The coincidences in the two different types of assessment seem to be more related to the effort the teams put into the assignment. Pink and Yellow presents the highest concentration of ESE group members whose supervisor happens to be the professor of the discipline. Also, Red and Green teams, which presented low score for protocol and report, and for precision and recall, respectively, are only composed by master students, even though there is a presence of practitioners in both of them.

Table 16 Precision and recall of each search in %

Papers up to 2010							
	Black	Red	Pink	Purple	Blue	Green	Yellow
Precision	2.57	2.69	12.10	2.63	4.98	0.72	13.08
Recall	58.62	48.28	65.52	31.03	41.38	13.79	58.62
Papers up to 2012							
Precision	–	–	–	2.26	3.42	0.78	12.74
Recall	–	–	–	28.13	37.50	15.63	62.50

6 Challenges and Pitfalls on SLR Planning and Execution in the SE Field

The observed results were different from our expected results as similarities in the research questions and protocol did not lead to similar outcomes (see Fig. 5). We are aware that the vast differences in the search strings induced the high quantity of different returns, but even when analyzing the same set of papers included by the teams, the quality attributes for use cases were not the same. Still, this study allowed us to identify some pitfalls of planning and reporting the SLR studies (underlined in the previous sections) that probably caused the differences identified throughout this exploratory study. As we are going to discuss in this section, six main reasons can be highlighted as challenges for conducting SLRs in SE and might explain the mentioned pitfalls; they are the lack of:

(i) experience in the investigated topic that caused the novice researchers to misuse terms in the search string, include studies and give answers not related to the research question;
(ii) the experience of novice researchers in systematic reviews promoting inconsistencies in their review protocol and execution, and making them do unnecessary work and not report relevant information;
(iii) a common terminology regarding use cases, requirements and quality attributes that made novice researchers search for studies using a variety of different terms and report the results inconsistently;
(iv) clearness and completeness of the papers that might have caused their inconsistent inclusion/exclusion of papers among the reviews;
(v) verification procedures to support the identification of inconsistencies throughout the *quasi*-SLR process, and;
(vi) commitment or interest with the research topic that caused the novice researchers to overlook important features of SLRs, not report significant decisions made during its process neither report details on the results.

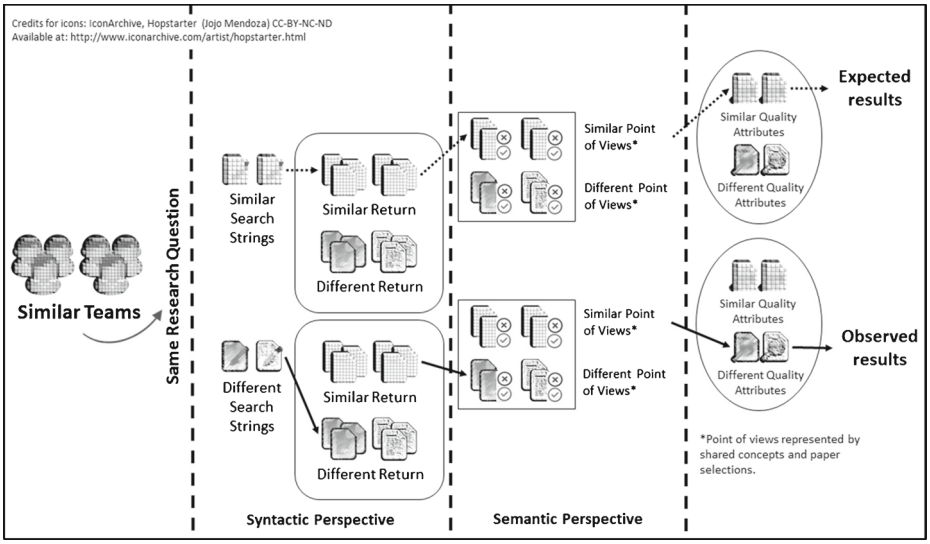

Fig. 5 Expected results versus observed results

The next subsections present a couple of observations supporting these believed main challenges. They might not be representative of all SLR experiences in SE, but they might help us understand some common issues identified in our field mainly when novice researchers (especially concerning the research method) perform secondary studies. Along with the challenges, we will also present some proposals that might be used to overcome them.

6.1 Lack of Experience in the Topic

"Keywords too high-level," "inappropriate selection of search terms," "report of information not related to the research topic" and "controversial understanding of the research topic" may be related to lack of experience in the topic or even lack of knowledge about the topic terms that are used by academia. As previously mentioned "use cases" was chosen as a subject for investigation, and more specifically "quality attributes for use cases," because we believed it is a ground topic in the SE field in which the novice researchers would have little misunderstandings. Also, to make the teams more similar to practitioners applying the evidence-based software engineering approach, we included in each group at least one participant with high experience in the software industry. However, we could observe some misunderstandings possibly related to the experience of the novices in the *quasi*-SLR topic.

Regarding use cases, it was possible to notice that Green extracted from the selected papers several attributes of quality not related to use cases, as described in Section 4. A thorough analysis of Green members' profiles reveals that one of them has a poor background regarding use cases. However, the other two members have enough experience to avoid these mistakes. As we do not know how the SLR tasks were distributed among the members, we can just speculate other causes, such as lack of verification procedures or lack of team commitment, as will be discussed later.

The differences concerning the applied inclusion/exclusion criteria (most of them not described in the protocols) aside from representing issues on describing important decisions, stress different perspectives regarding the SLR topic among the teams, evidencing the impact that knowledge and experience in the topic might affect the results. Hence, these findings indicate that expertise in the SLR topic seems to play an important role, especially concerning the keywords definition, the selection of works and the extraction of the results. Therefore, it is important to have a look at some seminal works in the research topic to get in touch with the vocabulary used in the area to minimize the effects of these issues. A proper selection of control articles also contributes to align the expectations concerning what to look for and what to select as included papers.

6.2 Lack of Experience in the Method

"Keywords too high-level," "unnecessary search terms," "many different combinations of terms producing noise return," "overuse of 'and' operators and distributive properties," "misuse of the guidelines," "no explanation concerning the empirical focus used," "tacit knowledge regarding the study selection strategy," "misinterpretation of the research question," "imprecise reports," "incomplete reports" and "subjectivity of the research synthesis strategy" may be related to lack of experience in the research method. The vast differences identified among the returned papers from the seven *quasi*-SLRs, even when some agreement regarding the search string was found, revealed the importance of organizing the search string properly.

It was observed among the teams that although they shared some concepts and terms, the terms were placed in different parts of the logical structure of the search string, leading to worse similarities results concerning the returned papers when compared to the terms similarity. All teams were instructed to use the PICO structure (Population AND Intervention AND Comparison AND Outcome; no Comparison expected), but we observed the abstraction of this structure was not properly applied in the *quasi*-SLRs. Some teams did not follow the PICO structure, but played with the search string logical structure, creating an additional dimension and even chaining logical operators as Pink. Other teams placed terms of population along with the intervention (Purple and Green) and others, terms of population along with outcome (Black, Red, and Purple). We expected the novice researchers to use as population, articles describing software development projects at the stage of requirements and empirical/experimental studies related to requirements; as intervention, use case descriptions or diagrams, or formats/guidelines/standards for its description; and as an outcome, quality attributes for use cases. In this case, no comparison was expected since we did not want to make any comparison between the intervention and a specific use case description/modeling way.

Other observed issues with the research protocols analysis were the insertion of papers satisfying exclusion criteria during the *quasi*-SLR execution and the evidence of missing information that might have affected the novices' decision. Purple evolved the exclusion criteria using too specific criteria, for instance: "articles about product line reporting quality attributes for use cases," and "articles about techniques that lead analysts to elaborate use cases such as prototyping, UI/GUI technologies, task models, sketching and mock-ups." Also, in many cases, we could neither identify criteria for including and excluding some specific papers nor find any explanation about the analysis and synthesis procedure used by the teams, facts that could have biased the studies selection. These behaviors made us wonder whether the teams had followed the research protocol and whether they had seen the importance of planning as much as possible before reviewing to avoid biasing the findings. As mentioned in previous works, novices might take advantage of more iterative approaches for SLR executions (Oates and Capper 2009; Lavallée et al. 2014), since they can improve novices' understanding concerning the method and information that should be described in order to consistently continue the SLR tasks, avoiding biasing the selection and reports.

Additionally, knowing the selected search engines' properties beforehand can support a better use of them, optimizing the search strategy by cutting additional search terms and, and thus reduce the noise in return.

6.3 Lack of Clearness and Completeness of the Papers

"Misunderstandings concerning empirical/experimental study strategies" may be related to lack of clearness and completeness of the papers under evaluation. The case in which four out of seven teams decided for the inclusion of the same returned paper, and the others, for its exclusion (section 5.2) raised the question whether the empirical study was well explained and categorized in the article. The paper does not report many details about the described case study (Ramos et al. 2009) and this fact might have hampered the team's inclusion/exclusion judgment.

The teams' knowledge and experience in empirical studies might have also contributed to different results, but we did notice teams Red, Green and Yellow excluded many papers after

reading them in full (according to their BibTeX), that is, they could not decide on the inclusion/exclusion just by reading the title and abstracts of the papers.

It is one more indication that we must keep spreading the need for using structured abstracts and reporting the guidelines utilized in the performed primary studies when writing scientific papers, as well as writing for the synthesis of evidence as advised by Wohlin (2014).

6.4 Lack of a Common SE Terminology

"Keywords too high-level," "many different combinations of terms producing noise return," "inappropriate selection of search terms," "misunderstandings concerning empirical/experimental study strategies," "controversial understanding of the research topic" may be related to lack of a common SE terminology.

The Black team tried to define general terms such as "scenario," "guidelines," "quality attributes," and "requirements engineering," creating the smallest term list among all teams, with only 11 terms. On the other hand, the Purple tried to specify precise terms and completed the task with the biggest term list among all the teams: 215 terms. Analyzing these 215 terms, we observed that the Purple used synonyms and tried to cover all possible combinations, such as "use case engineering," "use case modeling," "user scenario engineering," "user scenario modeling," "user story engineering," and "user story modeling," and the same strategy was applied to the other terms. It was also observed that the other teams had adopted the same strategy of Purple, but without trying to exhaust all combinations. In summary, except for the Black, all the others, to a greater or lesser degree, sought to use synonyms at some point during the keywords definition. However, it is not hard to observe that there is a significant difference among the synonyms adopted by each team, which can explain why each SLR returned such distinct sets of papers.

The choice of different synonyms can be a side effect of a lack of knowledge/experience in use cases or other topic related to the SLR, but we do not believe this is the case once the chosen synonyms, with some exceptions, are related to use cases, quality attributes, and empirical studies. In this particular case, possibly an adequate explanation is the lack of a common SE terminology. As there is not a minimum agreement about the appropriate terminology on investigation topics, the reviewers tend to adopt generic terms or run the risk of choosing a set of terms that is not a common sense among SE researchers.

Thus, it is important to search for taxonomies in the topic of research using them to suppnort the searching for studies and the report of results. This last case is crucial to facilitate future aggregations of two or more studies. Additionally, control papers can be useful in this instance as well.

6.5 Lack of Verification Procedures

"Unnecessary search terms," "inappropriate selection of search terms," "misuse of the guidelines," "misunderstandings concerning empirical/experimental study strategies," "misinterpretation of the research question," "controversial understanding of the research topic," "imprecise reports," "incomplete reports," "report of information not related to the research topic" may be related to lack of verification procedures. Most of the issues identified in the assessment of the research protocols and reports could have been overcome with a use of verification procedures, as simple pair reviews.

Many planning and outcome inconsistencies can be observed in the research protocols and reports, such as misalignment among the research questions and search strings (e.g. "fault free" and "testwarehouse"); PICO description not in conformance with the search strings; lack of fields in the extraction forms (Appendix 9) to support the extraction of information to answer the research questions (Appendix 1) or even to assess the quality of the selected papers (Appendix 8); answers not aligned to the research question, among others. These issues highlight the importance of referring to the study research question in each research protocol step while justifying the decisions made.

The studies selection process was also another issue identified in the *quasi*-SLRs. Some articles are inconsistent with the inclusion and exclusion criteria reported by the teams. For instance, the novice researchers in the Green team should have excluded at least two articles included if they had followed their exclusion criterion: "papers not presenting features about use case diagrams or specifications."

Regarding the data extraction, we could not find in any of the four research protocols (teams Red, Purple, Blue, and Green) fields to hold information to support the answering of quality assessment questions, although all protocols had data extraction fields concerned with quality attributes for use cases and their evaluation studies. This observation must mean that the teams either did not assess the studies – case of Purple and Green – or had to read the included papers just to evaluate them. The Black team did not plan a study quality assessment in its research protocol; and the Pink, Blue, and Yellow, although planned and reported the evaluation of all included papers, did not use it for any purpose. Only the Red team planned, reported and used the quality assessment to rank the quality attributes for use cases that had been found.

It is important to keep track of the main research question throughout the research protocol elaboration and follow it during the review execution. A general cross-checking, including the results, is advisable to check the plan and outcomes consistency and to increase confidence in the results. Zhang and Babar in (Zhang and Babar 2010) and Zhang, Babar and Tell in (Zhang et al. 2011) provide an interesting approach to be used in order to gather relevant studies during the search step, called *quasi*-gold standard, being an attractive alternative also to assess the quality of the search strategy used. Likewise, Petersen and Ali (2011) suggest interesting strategies for the study selection that can help in the planning phase of SLRs, mitigating some inconsistency risks during the study selection phase.

6.6 Lack of Commitment to the SLR

"No explanation concerning the empirical focus used;" "tacit knowledge regarding the study selection strategy;" "imprecise reports;" and "incomplete reports" may be related to lack of teams' commitment. Since the reviews were accomplished in the context of an ESE course, some participants might have faced this task as something purely related to obtaining a mark.

However, another variable to be considered is the time available to accomplish the SLR. Two months might not be enough to internalize the concepts related to the method and apply them in practice. In this case, the unexpected outcomes may be derived from the time pressure to deliver the final report, which in turn led the teams to give up the needed rigor in critical stages of the *quasi*-SLRs. We could notice that so they could reduce the effort and optimize the time on conducting the review, the participants split

the work among themselves, especially concerning the step of reading the returned papers' title and abstract. This can be observed in the teams' BibTeX in which in many cases a single paper had the evaluation of only one participant in the team. We also noticed that the limited time might be the cause of the poorly detailed reports (previous section). The amount of detail regarding the identified quality attributes was quite low, and in many cases, no definition for them was reported (Appendix 11).

SLRs demand high team commitment since it is a time-consuming task, its planning requires focus, dedication, and the selection and extraction phases involve a careful reading of hundreds or thousands of papers and a detailed cross-checking. Without this commitment, team members may try to shorten paths and minimize the rigor needed to obtain relevant results. Perhaps the best way to get this undertaking is to include in the team only people who have a direct interest in the SLR outcomes. On the other hand, involving people only to increase labor availability may negatively influence the overall results.

7 Threats to Validity

During this exploratory study, several threats to validity could be identified. Concerning the construct validity, the authors might not have considered all the main features to observe similarities among research protocols and answers, as well as the impact of the former in the latter, although we did consider most of the elements described in an SLR research protocol to accomplish the comparison. We understand that when an SLR research protocol is not well planned, it might lack important information for guiding the SLR execution. Likewise, as we conjectured in the previous subsections, information might be missing from the protocols – such as selection criteria and analysis procedure – because either the students did not update them frequently, or they did not make their reasoning explicit in the SLR plan and report. Still, for simplification matters, we decided to take all the reviews as complete, including the papers included and excluded, and the reported quality attributes. We understand that this decision might have made us overlook some answers and particularities that would have prevented us from comparing the research protocols and reported attributes in different reviews. We could have taken advantage of information gathered from the novice researchers to understand their decisions during the review, but we assumed the review was systematic and repeatable. About the coding process used to measure the similarity of search concepts across the reviews – presented in Section 3.6.1 –, the three researchers involved in it have significant theoretical knowledge and practical experience in the subject (use cases and quality attributes). Also, the coding was held at a meeting where the researchers discussed the different points of view to reach a consensus.

About internal validity, we identified some pitfalls and challenges in the previous sections that might have represented risks to the observed results. We did try to anticipate some problems that occurred in similar studies undertaken with students – presented in Section 2 – in this exploratory study. Therefore, the existing initial protocol was used as a starting point for all teams, and they were advised to use SLR guidelines to support the research protocol evolution. Also, the participants were organized into teams according to their knowledge and experiences in software development and experimentation. Still, all these efforts were not enough to prevent some other uncontrolled factors from

happening, as previously presented. Even though all participants received lectures and extra explanations concerned with the SLR executions, the lectures and even the material the students received (including the existing guidelines) might have left room for misunderstanding concerning the assignment.

The use of a non-native language can also be seen as a threat to this study validity. It is possible to conjecture that the elaboration of the search string in English and the reading of papers in a non-native language caused some difficulty that led to some of the previously commented mistakes. We do not believe in this possibility because all the participants are used to reading and writing technical papers and assignments in English, which considerably minimizes this kind of confusion factor. The short time for the SLR execution (two months) is also another threat. Nevertheless, most of the planning was given to the participants in the initial protocol, meaning they did not have to plan everything from scratch.

Regarding external validity, we cannot generalize this study mainly because we observed the *quasi*-SLRs planned and carried out by novice researchers (especially concerning the method) during a limited time. However, we understand literature reviews have been used as starting points in a lot of SE research executed by researchers (most of them novice ones such as graduate students), thus the investigation of this research strategy used by novice researchers is worth it. Also, considering that in industrial settings practitioners might not be used to reading and synthesizing papers, many of the findings can also be seen as reasonable in these contexts, even though the expertise in SE topics might help them to avoid some of the mentioned pitfalls. One thing we could conclude from this experience with novices: lack of knowledge and expertise in the topic and/or method can either lead to divergent results (most probably) or convergent results by chance.

As for the conclusion validity, the authors performed a coding process on the search terms to identify the main concepts presented in each search string that add some subjectivity to the comparison of the SLRs, and that might hamper this study conclusion and replication. Our intention with this process was to find more generic terms that would support us in identifying some similarities among the protocols, as a way to capture not only the exact terms used for the search but also the intention the teams had on searching for works on similar topics. Thus, the main concepts intend to offer generic representations of the search terms. Still, even in this case, it is not possible to see many similarities. One might think that the same process could be used for the outcomes (quality attributes for use cases) as well, and while this is true, we believe it would not add much to the similarity discussions in this particular work, since the generic term (concept) we could abstract from the outcomes would be a single one (quality features), leading to 100% similarity among all teams.

An interesting coding process to apply to the outcomes is one to identify similar quality attributes with different names, and different quality attributes with similar names. We were not able to perform such analysis since not all reported quality attributes have an associated definition, as can be observed in Appendix 11, making the coding process unfeasible. Any attempt at capturing the quality attributes definitions from the included papers would make the authors interfere the teams' answers to the research question, biasing the comparison among them. Hence, the comparison of the quality attributes reported by each team was made without any interpretation, that is, the comparison was made through the terms reported with almost no inference. For example:

"complex" and "complexity" were considered the same quality attribute, while "size" and "small size" were found to be different ones. The assignment did not specify anything about the level of granularity at which the quality attributes should be reported. Thus, the terms used to report the quality attributes were either quite different or quite similar, as Table 11 (Sectionx 4.3) allows observing. This last remark is related to another important assumption we made through the execution of this study; we assumed that the SLR packages the novice researchers provided were correct and complete. This was a way to properly assess the reliability of SLRs (process repeatability and outcomes consistency) without taking the risk of interfering in the teams' answers.

8 Conclusions

This work presents and discusses the planning, execution, and the results of SLRs performed by novice researchers, trying to evaluate similarities and differences among the protocols and the sets of selected studies. Although SLR protocols with low similarity have generated results with low similarity, as expected, protocols with some similarity (search string, returned and included studies, and so on) led to different results as well. This result makes us conjecture that SLRs are not reliable (having a repeatable process and leading to consistent results) as we might think, mainly when performed by novices. This discrepancy can be partially explained by the researcher's inexperience in the SLR's method and domain since several papers reported in the related works section highlight the importance of the researcher's experience in practice and the domain as a success factor for the SLR's repeatability and consistency in results.

On the other hand, the evidence-based software engineering (EBSE) promotes SLRs as the most important instrument to collect relevant information regarding a particular technology aiming at support practitioners in their decision making in SE. This way, the successful planning and execution of SLRs performed by professionals is a critical factor to EBSE.

These scenarios bring many pitfalls and challenges related to SLRs conducted by novice researchers and also practitioners, since there is an inherent difficulty involving this kind of participant, in addition to other factors that cause problems even for the most experienced researchers. We observed that missing information from the research protocols and reports of results mainly related to the adopted analysis procedure could compromise the repeatability of SLRs and the consistency of results. Another observation made through this exploratory study was that whether the main research objective is not used to guide the SLR planning activities, the researchers might bias the findings in the face of new information they encounter during the study selection. If changes are necessary during the process, then the process must be redone. Iterative approaches for SLR executions (Lavallée et al. 2014) combined with verification procedures can support the use of this research strategy by novices (Oates and Capper 2009). Furthermore, an in-depth understanding of the terminology employed in the topic under investigation and the *quasi*-golden standard is required to support the elaboration of appropriate search strings (regarding their precision and recall) and the reporting of results.

One issue deserves further discussion: would removing the novice researchers, replacing them with more experienced researchers, solve the problem? Could the inclusion of a senior researcher with experience in the SLR topic be able to make the SLR process

reliable as mentioned earlier? It is inevitable that there is some degree of subjectivity in the application of the inclusion/exclusion criteria, and especially in the information extraction. This subjectivity is complemented with the previous researchers' experience, that is, with their perspectives on that topic. This scenario seems to be similar to that discussed in the software artifact inspection studies. Perspective-based reading techniques (Shull et al. 2000) are good examples of the application of multiple perspectives aiming at evaluating a topic from different points of view. In this case, inspectors with various interests and backgrounds increase the likelihood of detecting defects, since different perspectives are explored during the inspection process. In the light of these results, the previous question can be rephrased: does the combination of different angles during the evaluation of studies returned by the search engines play a vital role regarding the repeatability of SLRs and, hence, their reliability; or will a rigorous protocol with senior researcher support be enough? Future investigations have to be conducted to look into the boundaries between what a research protocol can provide to support the reliability of an SLR when the researchers become essential in this process, which strategies can be adopted to mitigate this issue and how these parts can be joined in an efficient manner.

Acknowledgments We thank Daniela Cruzes, Marcela Genero, Martin Höst, Natalia Juristo, Nelly Condori-Fernandez, Oscar Dieste and Oscar Pastor for the initial discussions at ISERN 2009 that started this work; Vitor Faria Monteiro for his contribution to the original protocol planning; David Budgen for suggestions regarding an earlier version of this study report; all students for their engagement during the Experimental Software Engineering course in 2010 and 2012, and also the CNPq and CAPES for supporting this research. Prof. Travassos is a CNPq Researcher.

Appendix 1

Table 17 Teams' Research Questions

Team's Name	Research Question
Black	What are the existing quality attributes in use case specifications that are used as the object of studies in primary studies?
Red	1) Which quality attributes (and measurements used to evaluate such attributes) have been empirically studied for use cases? 2) What are the attributes (and measurements/factors) regarding the quality of use cases? • Is there any description format or standard to describe use cases that assure or maximize their quality?
Pink	Which quality attributes (and measurements used to evaluate such attributes) have been empirically studied for use cases? • In the case of finding empirically studied quality attributes for use cases, what was the description format used to materialize the use case model?
Purple	Which quality attributes (and measurements used to evaluate such attributes) have been empirically studied for use cases?
Blue	Which quality attributes (and measurements used to assess such attributes) have been empirically studied for use cases? • Is there some description format that can promote the use case quality?
Green	Which quality attributes (and measurements used to evaluate such attributes) have been empirically studied for use cases?
Yellow	Which quality attributes have been empirically studied for use cases? • Which measurements have been used to evaluate such attributes?

This appendix presents the research question reported by each team in their research protocol

Appendix 2

Table 18 Teams' Main Search Strings

Team's Name	Main Search String
Black	("requirements engineering" OR "requirements specification" Or "software development" OR "systems development" OR "software specification" OR "development projects" OR "development process") **AND** ("use case" OR "use cases" OR "scenario" or "scenarios") **AND** (guidelines OR "quality attributes" Or "quality attribute")
Red	("controlled experiment" OR "case study" OR "software project" OR "software development" OR "software requirement" OR "software elicitation" OR "software modeling" OR "software engineering" OR "software description" OR "software specification" OR "software diagram" OR "software application" OR "software experiment" OR "software implementation" OR "system project" OR "system development" OR "system requirement" OR "system elicitation" OR "system modeling" OR "system engineering" OR "system description" OR "system specification" OR "system diagram" OR "system application" OR "system experiment" OR "system implementation") **AND** ("use case") **AND** ("quality characteristic" OR "quality feature" OR "quality factor" OR "understandability" OR "efficiency" OR "correctness" OR "defect rate" OR "completeness" OR "consistency" OR "readability" OR "usefulness" OR "ease of learning" OR "traceable" OR "acceptability" OR "usability" OR "testability" OR "simulation" OR "level of abstraction" OR "communication" OR "plausibility" OR "consistency of structure" OR "alternative flow" OR "misinterpretation" OR "ease of construction" OR "cost effectiveness" OR "checkability" OR "soundness" OR "verifiability" OR "perceived ease of use" OR "intention to use" OR "precision" OR "appropriateness" OR "ease of use" OR "ease of analyze" OR "reuse" OR "quality model" OR "authoring" OR "guidelines" OR "format" OR "template")
Pink	(Software Project OR Software Development OR Software Development Project OR System Development OR Application Development OR Application Project OR Empirical Study OR Experiment OR Experimentation OR Empirical Assessment OR Empirical Evaluation OR Experimental Study OR Case Study OR survey OR Pilot Study **AND** (Use Case **AND** (description OR specification OR diagram OR model OR authoring))) **AND** (Use case quality OR Quality of use case OR Quality of the use case OR (Use case **AND** (quality characteristic OR quality feature OR quality factor OR quality attribute OR understandability OR comprehensibility OR effectiveness OR efficiency OR correctness OR completeness OR clarity OR consistency))) **AND** (Description format OR Description template OR template)
Purple*	(((((("Software Application" OR "Software System" OR "Software Program" OR "Software Project" OR "Software Engineering" OR "Software Design" OR "Software Development" OR "Project Engineering" OR "Project Design" OR "Project Development" OR "Software Process" OR "Software method" OR "Software methodology" OR "Software technique" OR "Software approach" OR "Application Process" OR "Application method" OR "Application methodology" OR "Application technique" OR "Application approach" OR "System Process" OR "System method" OR "System methodology" OR "System technique" OR "System approach" OR "Program Process" OR "Program method" OR "Program methodology" OR "Program technique" OR "Program approach" OR "Project Process" OR "Project method" OR "Project methodology" OR "Project technique" OR "Project approach" OR "Engineering Process" OR "Engineering method" OR "Engineering methodology" OR "Engineering technique" OR "Engineering approach" OR "Design Process" OR "Design method" OR "Design methodology" OR "Design technique" OR "Design approach" OR "Development Process" OR "Development method" OR "Development methodology" OR "Development technique" OR "Development approach" OR "Requirement engineering" OR "Software Factory" OR "TestWareHouse" OR "Software House" OR "SoftwareHouse") **AND** ("Use case modeling" OR "Use case modeling" OR "Use case engineering" OR "Use case description" OR "Use case specification" OR "Use case diagram" OR "Use-case modeling" OR "Use-case modeling" OR "Use-case engineering" OR "Use-case description" OR "Use-case specification" OR "Use-case diagram" OR "User Story modeling" OR "User Story modeling" OR "User Story engineering" OR "User Story

Table 18 (continued)

Team's Name	Main Search String
	description" OR "User Story specification" OR "User Story diagram" OR "User Scenario modeling" OR "User Scenario modeling" OR "User Scenario engineering" OR "User Scenario description" OR "User Scenario specification" OR "User Scenario diagram" OR "Requirement modeling" OR "Requirement modeling" OR "Requirement description" OR "Requirement specification" OR "Requirement diagram" OR "Use case model" OR "Use-case model" OR "User Story model" OR "User Scenario model" OR "Requirement model" OR "user-system interaction scenario")) **AND** (("Case study" OR "experience report" OR "experimental study" OR "empirical study" OR "action research" OR "action-research" OR "survey" OR "evaluation study" OR "experimental evaluation" OR "empirical evaluation" OR "proof of concept" OR "randomized experiment" OR "pseudo-randomized experiments" OR "experiment" OR "evidence-based experiment" OR "industrial study report" OR "industrial study" OR "industrial report" OR "industrial experiment" OR "empirical report" OR "research report" OR "method evaluation" OR "laboratory experiment" OR "application of the approach" OR "application of the method" OR "application of the process" OR "application of the framework" OR "application of the technique" OR "application of the methodology") **AND** ("Use case pattern" OR "description format" OR "description template" OR "quality characteristics" OR "quality features" OR "quality factors") OR ("Volatility" OR "degree of functional encapsulation" OR "fragmentation errors" OR "syntactic quality" OR "semantic quality" OR "pragmatic quality" OR "acquisition" OR "understandability" OR "understandable" OR "comprehension" OR "reusability" OR "reusable" OR "unambiguous" OR "ambiguity" OR "internally consistent" OR "modifiable" OR "precise" OR "not redundant" OR "concise" OR "repeatability" OR "comprehensibility" OR "comprehensiveness" OR "responsiveness" OR "intelligibility" OR "efficiency" OR "efficient" OR "effectiveness" OR "efficiently" OR "efficiencies" OR "correctness" OR "correct" OR "soundness" OR "defect rate" OR "defect fee" OR "defect ratio" OR "error rate" OR "error fee" OR "error ratio" OR "fault rate" OR "fault fee" OR "fault ratio" OR "rework rate" OR "rework fee" OR "rework ratio" OR "less rework" OR "more rework" OR "lower rework" OR "increasing rework" OR "decreasing rework" OR "decrescent rework" OR "crescent rework" OR "mistake rate" OR "mistake fee" OR "mistake ratio" OR "completeness" OR "completion OR "complete" OR "wholeness" OR "fullness" OR "consistency" OR "consistence" OR "consistent" OR "solidity" OR "stability" OR "fastness" OR "hardness" OR "validity" OR "substantiality" OR "appropriateness" OR "adequacy" OR "adequation" OR "suitability" OR "intention to use" OR "intention to apply" OR "intention to analyze" OR "ease of analysis" OR "ease of find" OR "ease of discover" OR "perceived ease of use" OR "characterization" OR "classification" OR "summarization" OR "verifiability" OR "verifiable" OR "checkability" OR "traceability" OR "traced" OR "traceable" OR "readability" OR "granularity"))))
Blue	(("use case" OR "use cases") **AND** ("software" OR "application" OR "system" OR "program" OR "project" OR "development" OR "description" OR "narrative" OR "narratives" OR "elicitation" OR "specification" OR "construction" OR "model" OR "models" OR "modeling" OR "modeling" OR "diagram" OR "diagrams" OR "templates" OR "format" OR "writing" OR "guidelines") **AND** ("quality" OR "characteristics" OR "properties" OR "features" OR "quality improvement" OR "quality factors" OR "quality assessment" OR "quality evaluation" OR "understandability" OR "efficiency" OR "effectiveness" OR "correctness" OR "defect rate" OR "completeness" OR "consistency") **AND** ("criteria" OR "attribute" OR "attributes" OR "measurement" OR "measure" OR "metric" OR "metrics"))
Green	(software project OR software development OR software engineering OR software design OR application project OR application development OR application engineering OR application design OR system project OR system development OR system engineering OR system design OR program project OR program development OR program engineering OR program design OR software analysis OR software method OR software methodology OR software technique OR software approach OR application analysis OR application method OR application methodology OR application technique OR application approach OR system analysis OR system method OR system methodology OR system technique OR system approach OR program analysis OR program method OR program methodology OR program technique OR

Table 18 (continued)

Team's Name	Main Search String
	program approach) **AND** (use case OR software requirement OR software elicitation OR software description OR software specification OR application requirement OR application elicitation OR application description OR application specification OR system requirement OR system elicitation OR system description OR system specification OR program requirement OR program elicitation OR program description OR program specification) **AND** (quality characteristics OR quality features OR quality factors OR quality attributes OR quality measurement OR quality evaluation OR understandability OR efficiency OR consistency)
Yellow	(("development of software" OR "software development" OR "software project" OR "system development" OR "systems development" OR "requirements engineering" OR "action research" OR "case study" OR "case studies" OR "experiment" OR "experiments" OR "exploratory study" OR "exploratory studies" OR "empirical analysis" OR "empirical assessment" OR "empirical comparison" OR "empirical evaluation" OR "empirical evidence" OR "empirical evidences" OR "empirical study" OR "empirical studies" OR "empirical validation" OR "empirical work" OR "empirical works" OR "experimental evaluation" OR "experimental study" OR "experimental studies" OR "study report" OR "experience report") **AND** ("textual use case" OR "textual use cases" OR "use case analysis" OR "use cases analysis" OR "use case approach" OR "use cases approach" OR "use case approaches" OR "use cases approaches" OR "use case driven approach" OR "use cases driven approach" OR "use case authoring" OR "use cases authoring" OR "use case description" OR "use cases description" OR "use case descriptions" OR "use cases descriptions" OR "use case diagram" OR "use cases diagram" OR "use cases diagram" OR "use cases diagrams" OR "use case method" OR "use cases method" OR "use case methods" OR "use cases methods" OR "use case model" OR "use cases model" OR "use case models" OR "use cases models" OR "use case modeling" OR "use cases modeling" OR "use case modeling" OR "use cases modeling" OR "use case scenarios" OR "use cases scenarios" OR "use case specification" OR "use cases specification" OR "use case technique" OR "use cases technique" OR "use case techniques" OR "use cases techniques" OR "use case writing" OR "use cases writing") **AND** ("quality attribute" OR "quality attributes" OR "quality characteristic" OR "quality characteristics" OR "quality feature" OR "quality features" OR "desirable quality" OR "desirable qualities" OR "use case quality" OR "use cases quality" OR "quality of the use case" OR "quality of the use cases" OR "quality of use case" OR "quality of use cases" OR "unambiguity" OR "understandability" OR "understandable" OR "comprehension" OR "clarity" OR "correctness" OR "correct" OR "correctly" OR "completeness" OR "complete" OR "consistency" OR "consistent" OR "readability" OR "readable" OR "traceability" OR "usability"))

This appendix presents the search strings conceived by each team to perform the search in the engines. These search strings represent the main ones each team used and adapted for searching in each of the three search engines

*Due to several inconsistencies presented in the team's research protocol, we had to use the search string they applied in Scopus engine

Appendix 3

Table 19 Main Search Concepts – Search Strings Coding

Main Concept	Search Terms Related	Teams that Presented the Concept
Defect Rate – it refers to different ways of measuring defect	"defect fee"; "defect rate"; "defect ratio"; "error fee"; "error rate"; "error ratio"; "fault fee"; "fault rate"; "fault ratio"; "fragmentation errors"; "mistake fee"; "mistake rate"; "mistake ratio"	Red, Purple and Blue
Evaluation – it refers to different ways of assessment, especially empirical studies strategies	"action research"; "action-research"; "case studies"; "case study"; "controlled experiment"; "empirical analysis"; "empirical assessment"; "empirical comparison"; "empirical evaluation"; "empirical evidence"; "empirical report"; "empirical studies"; "empirical study"; "empirical validation"; "empirical work"; "evaluation study"; "evidence-based experiment"; "experience report"; "experiment"; "experimental evaluation"; "experimental studies"; "experimental study"; "experimentation"; "exploratory studies"; "exploratory study"; "industrial experiment"; "industrial report"; "industrial study"; "industrial study report"; "laboratory experiment"; "method evaluation"; "pilot study"; "proof of concept"; "pseudo-randomized experiments"; "randomized experiment"; "research report"; "simulation"; "software experiment"; "study report"; "system experiment"; "survey"	Red, Pink, Purple and Yellow
Environment – it refers to different settings of software development companies	"software factory"; "software house"; "softwarehouse"; "testwarehouse"	Purple
General Quality Issue – it refers to general terms related to quality	"attribute"; "characteristics"; "criteria"; "desirable qualities"; "desirable quality"; "features"; "measure"; "metric"; "pragmatic quality"; "properties"; "quality"; "quality attribute"; "quality characteristic"; "quality factor"; "quality feature"; "quality of use case"; "quality of the use case"; "semantic quality"; "syntactic quality"; "use case quality"; "use cases quality"	Black, Red, Pink, Purple, Blue, Green and Yellow
Product – it refers to different terms used to name software products	"application"; "program"; "software"; "software application"; "software program"; "software system"; "system"; "system application"	Red, Purple, and Blue
Project – it refers to different terms used to name software projects	"application project"; "development projects"; "program project"; "project"; "software project"; "system project"	Black, Red, Pink, Purple, Blue, and Green
Quality Features – it refers to different characteristics used to evaluate quality	"acceptability"; "adequacy"; "adequation"; "ambiguity"; "appropriateness"; "checkability"; "clarity"; "complete"; "completeness"; "completion"; "comprehensibility"; "comprehension"; "comprehensiveness"; "communication"; "concise"; "consistence"; "consistency of structure"; "consistency"; "consistent"; "correct"; "correctly"; "correctness"; "cost effectiveness"; "degree of	Red, Pink, Purple, Blue and Yellow

Table 19 (continued)

Main Concept	Search Terms Related	Teams that Presented the Concept
	functional encapsulation"; "ease of analysis"; "ease of analyze"; "ease of construction"; "ease of discover"; "ease of find"; "ease of learning"; "ease of use"; "effectiveness"; "efficiencies"; "efficiency"; "efficient"; "efficiently"; "fastness"; "fullness"; "granularity"; "hardness"; "intelligibility"; "intention to analyze"; "intention to apply"; "intention to use"; "internally consistent"; "level of abstraction"; "misinterpretation"; "modifiable"; "not redundant"; "perceived ease of use"; "plausibility"; "precision"; "precise"; "readability"; "readable"; "repeatability"; "responsiveness"; "reusable"; "reusability"; "reuse"; "solidity"; "soundness"; "stability"; "summarization"; "substantiality"; "suitability"; "testability"; "traceability"; "traceable"; "traced"; "unambiguity"; "unambiguous"; "understandability"; "understandable"; "usability"; "usefulness"; "validity"; "verifiability"; "verifiable"; "volatility"; "wholeness"	
Requirements Documents – it refers to software requirements documents	"application description"; "application requirement"; "application specification"; "description"; "narrative"; "program description"; "program requirement"; "program specification"; "requirement description"; "requirement specification"; "software description"; "software requirement"; "software specification"; "specification"; "system description"; "system requirement"; "system specification"	Black, Red, Pink, Purple, Blue and Green
Requirements Models – it refers to different ways of depicting requirements	"diagram"; "model"; "requirement diagram"; "requirement model"; "software diagram"; "software model"; "system diagram"; "system model"	Red, Pink, Purple and Blue
Requirements Representation Policy – it refers to different ways of guiding requirements representation	"description format"; "description template"; "format"; "guideline"; "template"	Black, Red, Pink, Purple and Blue
Rework Rate – it refers to different ways of measuring rework	"crescent rework"; "decreasing rework"; "decrescent rework"; "increasing rework"; "less rework"; "lower rework"; "more rework"; "rework fee"; "rework rate"; "rework ratio"	Purple
Scenario – it refers to different terms used to name scenarios	"scenario"; "use case scenarios"; "use cases scenarios"; "user-system interaction scenario"	Black, Purple, and Yellow
Scenario Documents – it refers to scenario documentation	"user scenario description"; "user scenario specification"	Purple
Scenario Models – it refers to different ways of depicting scenarios	"user scenario diagram"; "user scenario model"	Purple
Software Life Cycle –		

Table 19 (continued)

Main Concept	Search Terms Related	Teams that Presented the Concept
it refers to the different phases of a software development process	"acquisition"; "application of the approach"; "application of the framework"; "application of the method"; "application of the methodology"; "application of the process"; "application of the technique"; "application analysis"; "application design"; "application development"; "application elicitation"; "application engineering"; "application process"; "authoring"; "characterization"; "classification"; "construction"; "design process"; "development"; "development of software"; "development process"; "elicitation"; "engineering process"; "measurement"; "modeling"; "modeling"; "program analysis"; "program design"; "program development"; "program engineering"; "program elicitation"; "program process"; "project design"; "project development"; "project engineering"; "project process"; "quality assessment"; "quality improvement"; "quality evaluation"; "quality measurement"; "requirement engineering"; "requirements engineering"; "requirement modeling"; "requirement modeling"; "software analysis"; "software design"; "software development"; "software development project"; "software elicitation"; "software engineering"; "software implementation"; "software process"; "system analysis"; "system design"; "system development"; "systems development"; "system elicitation"; "system engineering"; "system implementation"; "system process"; "use case analysis"; "use cases analysis"; "use case authoring"; "use cases authoring"; "use case engineering"; "use-case engineering"; "use case modeling"; "use cases modeling"; "use-case modeling"; "use case modeling"; "use cases modeling"; "use-case modeling"; "use case writing"; "use cases writing"; "user scenario engineering"; "user scenario modeling"; "user scenario modeling"; "user story engineering"; "user story modeling"; "user story modeling"; "writing"	Black, Red, Pink, Purple, Blue, Green and Yellow
Software Technology – it refers to different types of technologies used to support the software development	"application approach"; "application method"; "application methodology"; "application technique"; "design approach"; "design method"; "design methodology"; "design technique"; "development approach"; "development method"; "development methodology"; "development technique"; "engineering approach"; "engineering method"; "engineering methodology"; "engineering technique"; "program approach"; "program method"; "program methodology"; "program technique"; "project approach"; "project method"; "project methodology"; "project technique"; "quality model"; "software approach"; "software method"; "software methodology"; "software technique"; "system approach"; "system method"; "system methodology"; "system technique"; "use case approach"; "use case approaches"; "use case driven approach"; "use case method"; "use case technique"; "use cases approach"; "use cases approaches"; "use cases driven approach"; "use cases method"; "use cases technique"	Red, Purple, Green and Yellow

Table 19 (continued)

Main Concept	Search Terms Related	Teams that Presented the Concept
Use Case – it refers to different terms used to name use cases	"use case"	Black, Red, Pink, Blue, and Green
Use Case Concepts – it refers to concepts regarding use cases	"alternative flow"	Red
Use Case Documents – it refers to use case documentation	"use case description"; "use cases description"; "use-case description"; "use case specification"; "use cases specification"; "use-case specification"	Purple and Yellow
Use Case Models – it refers to different ways of depicting use cases	"use case diagram"; "use cases diagram"; "use-case diagram"; "use case model"; "use cases model"; "use-case model"; "textual use case"	Purple and Yellow
Use Case Representation Policy – it refers to different ways of guiding use case representation	"use case pattern"	Purple
User Story Documents – it refers to user story documentation	"user story description"; "user story specification"	Purple
User Story Models – it refers to different ways of depicting user stories	"user story diagram"; "user story model"	Purple

This appendix presents the main concepts that were extracted from the terms used in the teams' search strings

Appendix 4

Table 20 Teams' Search Engines Configuration

Team's Name	IEEExplore	Scopus	Web of Science
Black	✓	✓	The search string was refined to returned only papers from "computer science"
Red	✓	The search string was refined to exclude articles from "medicine" and "biochemistry"	✓
Pink	✓	✓	✓
Purple	The team removed this search engine due to its limitation on the number of search terms	✓	✓
Blue	✓	✓	✓
Green	✓	✓	✓
Yellow	✓	✓	✓

This appendix presents the search engines selected by the teams and the particular configurations that some of them did to them

Appendix 5

Table 21 Teams' Control Articles

Paper	Black	Red	Pink	Purple	Blue	Green	Yellow
El-Attar, M. & Miller, J. **A subject-based empirical evaluation of SSUCD's performance in reducing inconsistencies in use case models**. *Empirical Software Engineering,* 2009*, 14*, 477–512					✓		
Menzel, I.; Mueller, M.; Gross, A. & Doerr, J. **An experimental comparison regarding the completeness of functional requirements specifications**. *Proceedings of the 2010 18th IEEE International Requirements Engineering Conference, RE2010,* 2010, 15–24							✓
Anda, B.; Hansen, K. & Sand, G. **An investigation of use case quality in a large safety-critical software development project**. Information and Software Technology, 2009, 51, 1699–1711		✓		✓			
Fantechi, A.; Gnesi, S.; Lami, G. & Maccari, A. **Application of linguistic techniques for Use Case Analysis**. *Requirements Engineering, 2002. Proceedings. IEEE Joint International Conference on,* 2002, 157–164					✓		
Phalp, K.; Vincent, J. & Cox, K. **Assessing the quality of use case descriptions**. *Software Quality Journal,* 2007*, 15*, 69–97							✓
Cox, K., Phalp, K., Shepperd, M. **Comparing Use Case Writing Guidelines**. 7th International Workshop on Requirements Engineering: Foundation for Software Quality, 2001.	✓						✓
España, S.; Condori-Fernandez, N.; Gonzalez, A. & Pastor, O. **Evaluating the completeness and granularity of functional requirements specifications: A controlled**				✓			

Table 21 (continued)

Paper	Black	Red	Pink	Purple	Blue	Green	Yellow
experiment. Proceedings of the IEEE International Conference on Requirements Engineering, 2009, 161–170							
Ben Achour, C.; Rolland, C.; Maiden, N. & Souveyet, C. **Guiding use case authoring: results of an empirical study**. Proceedings of the IEEE International Conference on Requirements Engineering, 1999, 36–43.	✓	✓					
Chandrasekaran, P. **How Use Case Modeling Policies Have Affected the Success of Various Projects (or How to Improve Use Case Modeling)**. Addendum to the 1997 ACM SIGPLAN conference on Object-oriented programming, systems, languages, and applications, 6–9		✓					
Kamalrudin, M; Hosking, J & Grundy, J. **Improving requirements quality using essential use case interaction patterns**. Software Engineering (ICSE), 2011 33rd International Conference on, IEEE, 2011, 531–540						✓	
Phalp, K.; Vincent, J. & Cox, K. **Improving the quality of use case descriptions: Empirical assessment of writing guidelines** *Software Quality Journal,* 2007*, 15*, 383–399					✓		✓
Anda, B.; Sjøberg, D. & Jørgensen, M. Knudsen, J. L. (Ed.). **Quality and Understandability of Use Case Models**. Object-Oriented Programming: 15th European Conference Budapest, Hungary, June 18--22, 2001 Proceedings, Springer Berlin Heidelberg, 2001, 402–428	✓						
Ramos, R. b.; Castro, J.; Alencar, F.; Araújo, J.; Moreira, A. & Penteado, R. **Quality improvement for use case model**. SBES 2009 - 23rd Brazilian Symposium on Software Engineering, 2009, 187–195		✓					
Cox, K. & Phalp, K. **Replicating the CREWS use case authoring guidelines experiment**. Empirical Software Engineering, 2000, 5, 245–267	✓	✓					✓
Phalp, K & Cox, K. **Supporting Communicability with Use Case Guidelines: An Empirical Study**. Keele; Keele University, 2002, 8–10.							✓
Cherfi, S.-S.; Akoka, J. & Comyn-Wattiau, I. **Use case modeling and refinement: A quality-based approach**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2006*, 4215 LNCS*, 84–97					✓		
Issa, A. & Al-Ali, A. **Use case patterns driven requirements engineering**. 2nd International Conference on Computer Research and Development, ICCRD 2010, 2010, 307–313				✓			

This appendix presents the control articles selected by each team

Appendix 6

Table 22 Teams' Inclusion Criteria

Inclusion Criteria	Black	Red	Pink	Purple	Blue	Green	Yellow
To talk about requirements engineering		✓	✓	✓	✓	✓	✓
To analyze requirements elicitation with respect to use cases		✓	✓	✓	✓	✓	✓
To present characteristics of use cases diagram or specification		✓	✓	✓	✓	✓	✓
To describe some description format for use cases		✓	✓	✓	✓	✓	✓
To present a guideline or road map to writing use cases		✓					
To talk about quality attributes of use case	✓			✓		✓	
To present some empirical study involving the quality characteristics	✓		✓	✓			
To compare use cases writing		✓					
The documents have to be available online	✓						

This appendix presents the inclusion criteria used by each team

Appendix 7

Table 23 Teams' Exclusion Criteria

Exclusion Criteria	Black	Red	Pink	Purple	Blue	Green	Yellow
Not concerned with requirements engineering		✓	✓	✓	✓	✓	✓
Not talking about use cases diagram or specification	✓	✓	✓	✓	✓	✓	✓
About product line reporting use case quality attributes				✓			
About techniques that lead analysts to elaborate use cases such as prototyping, UI/GUI techniques, task models, sketching and mock-ups				✓			
Not presenting features about uses case diagram or specification	✓	✓	✓	✓	✓	✓	✓
Not presenting any empirical evaluation involving quality characteristics/features/attributes of use case			✓				✓
Use proof of concept as the study validation				✓			
If two papers publish the same empirical results, one of them is excluded					✓		
Not written in English							✓
Any paper that is not accessible					✓		✓

This appendix presents the exclusion criteria used by each team

Appendix 8

Table 24 Teams' Quality Assessment Criteria

Quality Assessment Criteria	Black	Red	Pink	Purple	Blue	Green	Yellow
Were the aims and objectives clearly reported (including a rationale for why the study was undertaken)?			✓				
Was there an adequate description of the context in which the research was carried out?			✓	✓			
Was the research design appropriate to address the aims of the research?			✓				
Is there any description about the size of the population that joined the study?				✓			
Was there an adequate description of the sample used and the methods for how the sample was identified and recruited?			✓				
Were appropriate data collection methods used and described?			✓				
Is there any statistical result?				✓			
Was there an adequate description of the methods used to analyze data and whether appropriate methods for ensuring the data analysis was grounded in the data?			✓				
Did the study provide clearly stated findings with credible results and justified conclusions?			✓				
Is there any description of the threats to validity?				✓			
Is there any description of the study generalization?				✓			
Is there any definition for the quality features identified?							✓
Is there any description about how the quality features have been identified?		✓	✓	✓	✓	✓	✓
Is there any description of restrictions and conditions where the quality features were observed?		✓	✓	✓	✓	✓	
Is there any description of how the quality features can be measured?		✓	✓	✓	✓	✓	✓
Is the measurement procedures fully described?							✓
Does the paper describe any adaptation/evolution of pre-existent approach?		✓		✓	✓	✓	
Is the description format referenced/evaluated/used in other works?		✓	✓	✓	✓	✓	
Does the paper describe an application of the description format used to its evaluation?		✓	✓	✓	✓	✓	
Does the paper evaluate the UC diagram or the use case description format through a well-described example application?							✓
Is there any empirical/experimental result regarding the description format (or the UC diagram)?		✓	✓	✓	✓	✓	✓
Is it possible to identify for which types of system the description format can be used?		✓	✓	✓	✓	✓	
Is it possible to evaluate which quality features the description (or the UC diagram) format can promote?		✓	✓	✓	✓	✓	✓

Appendix 9

Table 25 Teams' Extraction Forms

Extraction Form Fields	Black	Red	Pink	Purple	Blue	Green	Yellow
Title	✓	✓	✓	✓	✓	✓	✓
Authors	✓	✓	✓	✓	✓	✓	✓
Year of publication	✓	✓	✓	✓	✓	✓	✓
Source of publication	✓	✓	✓	✓	✓	✓	✓
Abstract		✓	✓	✓	✓	✓	✓
Study strategy	✓		✓		✓		✓
Study goal	✓		✓				
Study context	✓		✓				
Study findings	✓		✓				✓
Use Case description format		✓	✓	✓	✓	✓	✓
Use Case description format source			✓				
Content guidelines details							✓
Minor guidelines details							✓
Style guidelines details							✓
Template guideline details							✓
Types of system in which the description format has been fpplied to (optional)	✓	✓	✓	✓	✓	✓	✓
Tool/Technique to improve quality attributes		✓					
Restrictions on the UC description format							✓
Quality focus							✓
Quality attributes	✓	✓	✓	✓	✓	✓	
Quality attribute description							✓
Quality attribute source			✓				
Quality attribute measurement		✓	✓	✓	✓	✓	✓

This appendix presents the extraction form used by each team

Appendix 10

Table 26 Included Papers by the Teams and the Authors

Paper	Black	Red	Pink	Purple	Blue	Green	Yellow	Authors
Diaz, I. b. c.; Losavio, F.; Matteo, A. & Pastor, O. **A specification pattern for use cases**. *Information and Management,* 2004*, 41*, 961–975		✓						
El-Attar, M. & Miller, J. **A subject-based empirical evaluation of SSUCD's performance in reducing inconsistencies in use case models**. *Empirical Software Engineering,* 2009*, 14*, 477–512	✓	✓	✓	✓	✓		✓	✓
Condori-Fernandez, N.; Daneva, M.; Sikkel, K.; Wieringa, R.; Dieste, O. & Pastor, O. **A systematic mapping study on empirical evaluation of software requirements specifications techniques**. *2009 3rd International Symposium on Empirical Software Engineering and Measurement, ESEM 2009,* 2009, 502–505				✓				
Jain, A. & Chaudhary, B. **A Use Case Driven Formal Approach to Check Consistency between UI Requirement and Implementation**. *Industrial and Information Systems, 2008. ICIIS 2008. IEEE Region 10 and the Third International Conference on,* 2008, 1–6			✓					
Yue, T.; Briand, L. & Labiche, Y. **A use case modeling approach to facilitate the transition towards analysis models: Concepts and empirical evaluation**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2009*, 5795 LNCS*, 484–498	✓	✓						✓
Denger, C. c.; Paech, B. d. & Freimut, B. e. **Achieving high quality of use-case-based requirements**. *Informatik - Forschung und Entwicklung,* 2005*, 20*, 11–23	✓	✓						✓
El-Attar, M. & Miller, J. **AGADUC: Towards a more precise presentation of functional requirement in use case models**. *Proceedings - Fourth International Conference on Software Engineering Research, Management and Applications, SERA 2006,* 2006, 346–353	✓	✓						
Espana, S.; Condori-Fernandez, N.; Gonzalez, A. & Pastor, O. **An empirical comparative evaluation of requirements engineering methods**. *Journal of the Brazilian Computer Society,* 2010, 1–17			✓		✓			✓
Torner, F.; Ivarsson, M.; Pettersson, F. & Ohman, P. **An Empirical Quality Assessment of Automotive Use Cases**. *Requirements Engineering, 14th IEEE International Conference,* 2006, 89–98								✓
Hadar, I.; Kuflik, T.; Perini, A.; Reinhartz-Berger, I.; Ricca, F. & Susi, A. **An empirical study of requirements model understanding: Use Case vs. Tropos models**. *Proceedings of the ACM Symposium on Applied Computing,* 2010, 2324–2329			✓	✓			✓	✓
Cox, K.; Aurum, A. & Jeffery, R. **An experiment in inspecting the quality of use case descriptions**. *Journal of Research and Practice in Information Technology,* 2004*, 36*, 211–229								✓
Mustafa, B. Fujita, H. *(Ed.)* **An experimental comparison of use case models understanding by novice and high knowledge users**. *Frontiers in Artificial Intelligence and Applications,* 2010*, 217*, 182–199							✓	

Table 26 (continued)

Paper	Black	Red	Pink	Purple	Blue	Green	Yellow	Authors
Menzel, I.; Mueller, M.; Gross, A. & Doerr, J. **An experimental comparison regarding the completeness of functional requirements specifications**. *Proceedings of the 2010 18th IEEE International Requirements Engineering Conference, RE2010,* 2010, 15–24				✓			✓	✓
Cox, K.; Aurum, A. & Jeffery, R. **An experiment in inspecting the quality of use case descriptions.** *Journal of Research and Practice in Information Technology,* 2004*, 36*, 211–229	✓		✓					
Loconsole, A. & Borstler, J. **An industrial case study on requirements volatility measures**. *Proceedings - Asia-Pacific Software Engineering Conference, APSEC,* 2005*, 2005*, 249–256				✓				
Anda, B.; Hansen, K. & Sand, G. **An investigation of use case quality in a large safety-critical software development**. Project *Information and Software Technology,* 2009*, 51*, 1699–1711	✓	✓	✓	✓			✓	✓
Fantechi, A.; Gnesi, S.; Lami, G. & Maccari, A. **Application of linguistic techniques for Use Case Analysis**. *Requirements Engineering, 2002. Proceedings. IEEE Joint International Conference on,* 2002, 157–164					✓			✓
Florez-Larrahondo, G. & Haddock, W. **Aspect oriented programming with hidden Markov models to verify design use cases**. *Proceedings of the 8th ACM International Conference on Aspect-Oriented Software Development, AOSD'09,* 2009, 223–228			✓					
Phalp, K.; Vincent, J. & Cox, K. **Assessing the quality of use case descriptions**. *Software Quality Journal,* 2007*, 15*, 69–97	✓	✓	✓				✓	✓
Richards, D. & Boettger, K. **Assisting Decision Making in Requirements Reconciliation**. *Proceedings of the International Conference on Computer Supported Cooperative Work in Design,* 2002*, 7*, 345–350		✓						
Subramaniam, K.; Far, B. & Eberlein, A. **Automating the transition from stakeholders' requests to use cases in OOAD**. *Canadian Conference on Electrical and Computer Engineering,* 2004*, 1*, 0515–0518	✓	✓						
El-Ansary, A. **Behavioral pattern analysis: Towards a new representation of systems requirements based on actions and events**. *Proceedings of the ACM Symposium on Applied Computing,* 2002, 984–991		✓						
Alchimowicz, B.; Jurkiewicz, J.; Ochodek, M. & Nawrocki, J. **Building benchmarks for use cases**. *Computing and Informatics,* 2010*, 29*, 27–44								✓
Åšmialek, M.; Bojarski, J.; Nowakowski, W.; Ambroziewicz, A. & Straszak, T. **Complementary use case scenario representations based on domain vocabularies**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2007*, 4735 LNCS*, 544–558		✓						
Alspaugh, T.; Sim, S.; Winbladh, K.; Diallo, M.; Naslavsky, L.; Ziv, H. & Richardson, D. **Clarity for stakeholders: Empirical evaluation of ScenarioML, use cases, and sequence diagrams.** *5th International Workshops on Comparative Evaluation in Requirements Engineering, CERE,* 2007, 1–10			✓					✓
Mustafa, B. **Comparing the effect of use case format on end user understanding of system requirements**. *Journal of Information Technology Research,* 2010*, 3*, 1–20							✓	

Table 26 (continued)

Paper	Black	Red	Pink	Purple	Blue	Green	Yellow	Authors
Dahan, M.; Shoval, P. & Sturm, A. **Comparing the impact of the OO-DFD and the Use Case methods for modeling functional requirements on comprehension and quality of models: a controlled experiment**. *Requirements Engineering,* 2012, 1–17							✓	✓
Li, X.; Liu, Z. & He, J. **Consistency checking of UML requirements**. *Proceedings of the IEEE International Conference on Engineering of Complex Computer Systems, ICECCS,* 2005, 411–420			✓					
Torner, F.; Ivarsson, M.; Pettersson, F. & Ohman, P. **Defects in automotive use cases**. *ISESE'06 - Proceedings of the 5th ACM-IEEE International Symposium on Empirical Software. Engineering,* 2006*, 2006*, 115–123	✓							✓
Mavin, A.; Wilkinson, P.; Harwood, A. & Novak, M. **EARS (Easy Approach to Requirements Syntax)**. *Proceedings of the IEEE International Conference on Requirements Engineering,* 2009, 317–322				✓				
Hnatkowska, B. & Grzegorczyn, M. **Empirical comparison of comprehensibility of requirement specification techniques based on natural languages and activity diagrams**. *Proc. of the 10th Int. Workshop on Modeling, Simulation, Verification and Validation of Enterprise Information Systems, MSVVEIS 2012 and 1st Int. Workshop on, WEBI 2012,in Conj. with ICEIS 2012,* 2012, 27–36								✓
Bernárdez, B.; Durán, A. & Genero, M. **Empirical evaluation and review of a metrics-based approach for use case verification**. *Journal of Research and Practice in Information Technology,* 2004*, 36*, 247–25								✓
Durán, A.; Bernárdez, B. b.; Genero, M. & Piattini, M. **Empirically driven use case metamodel Evolution**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2004*, 3273*, 1–11								✓
Bolloju, N. & Sun, S. **Enhancing the quality of use case models and activity diagrams using a differential quality model**. *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, DESRIST '09,* 2009		✓						
España, S.; Condori-Fernandez, N.; Gonzalez, A. & Pastor, O. **Evaluating the completeness and granularity of functional requirements specifications: A controlled experiment**. *Proceedings of the IEEE International Conference on Requirements Engineering,* 2009, 161–170	✓		✓	✓				✓
Bolloju, N. & Sun, S. **Exploiting the Complementary Relationship between Use Case Models and Activity Diagrams for Developing Quality Requirements Specifications**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2008*, 5232 LNCS*, 144–153	✓							
Rolland, C. & Ben Achour, C. **Guiding the construction of textual use case specifications**. *Data and Knowledge Engineering,* 1998*, 25*, 125–160	✓						✓	✓
Ben Achour, C.; Rolland, C.; Maiden, N. & Souveyet, C. **Guiding use case authoring: results of an empirical study**. *Proceedings of the IEEE International Conference on Requirements Engineering,* 1999, 36–43	✓	✓	✓			✓	✓	✓

Table 26 (continued)

Paper	Black	Red	Pink	Purple	Blue	Green	Yellow	Authors
Jagielska, D.; Wernick, P.; Wood, M. & Bennett, S. **How natural is natural language?: How well do computer science students write use cases?** *Proceedings of the Conference on Object-Oriented Programming Systems, Languages, and Applications, OOPSLA,* 2006*, 2006*, 914–924	✓	✓	✓			✓	✓	✓
Mustafa, B. & Deris, S. **How prior knowledge affects user's understanding of system requirements?** *World Academy of Science, Engineering and Technology,* 2009*, 38*, 18–26							✓	
Kaiya, H.; Tanigawa, M.; Suzuki, S.; Sato, T.; Osada, A. & Kaijiri, K. **Improving reliability of spectrum Purple2lysis for software quality requirements using TCM**. *IEICE Transactions on Information and Systems,* 2010*, E93-D*, 702–712				✓				
Kamalrudin, M; Hosking, J & Grundy, J. **Improving requirements quality using essential use case interaction patterns**. *Software Engineering (ICSE), 2011 33rd International Conference on, IEEE,* 2011, 531–540						✓		
Phalp, K.; Vincent, J. & Cox, K. **Improving the quality of use case descriptions: Empirical assessment of writing guidelines** *Software Quality Journal***,** 2007*, 15*, 383–399	✓		✓		✓		✓	✓
El-Attar, M. & Miller, J. **Improving the quality of use case models using anti-patterns**. *SOFTWARE AND SYSTEMS MODELING, SPRINGER HEIDELBERG,* {2010}*, {9}*, {141–160}								✓
Sinnig, D.; Chalin, P. & Khendek, F. **LTS semantics for use case models**. *Proceedings of the ACM Symposium on Applied Computing,* 2009, 365–370	✓							
El-Attar, M. & Miller, J. **Matching anti-patterns to improve the quality of use case models**. *Proceedings of the IEEE International Conference on Requirements Engineering,* 2006, 96–105	✓						✓	✓
D'Amorim, F. & Borba, P. **Modularity analysis of use case implementations**. *Journal of Systems and Software,* 2012*, 85*, 1012–1027						✓		
Preiss, O.; Wegmann, A. & Wong, J. **On quality attribute based software engineering**. *PROCEEDINGS OF THE 27TH EUROMICRO CONFERENCE - 2001: A NET ODYSSEY, IEEE COMPUTER SOC,* {2001}, {114–120}						✓		
Geneva, G.; Llorens, J.; Metz, P.; Prieto-Diaz, R. & Astudillo, H. **Open issues in industrial use case modeling**. *Lecture Notes in Computer Science,* 2005*, 3297*, 52–61		✓						
Condori-Fernandez, N.; Daneva, M.; Sikkel, K. & Herrmann, A. **Practical relevance of experiments in comprehensibility of requirements specifications**. *Proceedings - 1st International Workshop on Empirical Requirements Engineering, EmpiRE 2011,* **2011**, 21–28				✓				
El-Attar, M. & Miller, J. **Producing robust use case diagrams via reverse engineering of use case descriptions.** *Software and Systems Modeling***,** 2008*, 7***,** 67–83	✓	✓						
Bolloju, N. & Sugumaran, V. **Quality dependencies among use case models and sequence diagrams developed by novice systems analysts**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2007*, 4802 LNCS*, 96–105								✓

Table 26 (continued)

Paper	Black	Red	Pink	Purple	Blue	Green	Yellow	Authors
Ramos, R. b.; Castro, J.; Alencar, F.; Araújo, J.; Moreira, A. & Penteado, R**. Quality improvement for use case model**. SBES 2009 - 23rd Brazilian Symposium on Software Engineering, 2009, 187–195	✓	✓	✓			✓		✓
Cox, K. & Phalp, K. **Replicating the CREWS use case authoring guidelines experiment**. *Empirical Software Engineering,* 2000*, 5*, 245–267	✓	✓	✓	✓			✓	✓
Johnson, R.; Roussos, G. & Tagllati, L. d. **Requirements analysis for large scale systems**. *Journal of Object Technology***,** 2008*, 7*, 119–137		✓						
Henderson-Sellers, B.; Zowghi, D.; Klemola, T. & Parasuram, S. Bellahsene, Z.; Patel, D. & Rolland, C. *(Eds.)* **Sizing use cases: How to create a standard metrical approach** *OBJECT-ORIENTED INFORMATION SYSTEMS, PROCEEDINGS,* 2002*, 2425*, 409–421		✓						
Udomchaiporn, A.; Prompoon, N. & Kanongchaiyos, P. **Software Requirements Retrieval Using Use Case Terms and Structure Similarity Computation**. *Software Engineering Conference, 2006. APSEC 2006. 13th Asia Pacific,* 2006, 113–120		✓		✓				
Kassab, M.; Constantinides, C. & Ormandjieva, O. **Specifying and separating concerns from requirements to design: A case study**. *Proceedings of the Second IASTED International Multi-Conference on Automation, Control, and Information Technology, ACIT 2005,* 2005, 18–27				✓				
Whittle, J. **Specifying precise use cases with use case charts**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2006*, 3844 LNCS*, 290–301		✓						
Sinha, A.; Sutton, S. & Paradkar, A. **Text2Test: Automated Inspection of Natural Language Use Cases**. *Software Testing, Verification and Validation (ICST), 2010 Third International Conference on,* 2010, 155–164		✓						
Phalp, K.; Adlem, A.; Jeary, S.; Vincent, J. & Kanyaru, J. **The role of comprehension in requirements and implications for use case descriptions**. *Software Quality Journal,* 2011*, 19*, 461–486								✓
Anda, B. & Sjøberg, D. **Towards an inspection technique for use case models**. *ACM International Conference Proceeding Series,* 2002*, 27*, 127–134			✓					✓
Böttger, K.; Schwitter, R.; Mollá, D. & Richards, D. **Towards reconciling use cases via controlled language and graphical models**. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science),* 2003*, 2543*, 115–128		✓						
Kealey, J. & Amyot, D. **Towards the Automated Conversion of Natural-Language use Cases to Graphical use Case Maps**. *Electrical and Computer Engineering, 2006. CCECE '06. Canadian Conference on,* 2006, 2377–2380		✓						
Belgamo, A.; Fabbri, S. & Maldonado, J. **TUCCA: Improving the effectiveness of use case construction and requirement analysis**. *2005 International Symposium on Empirical Software Engineering, ISESE 2005,* 2005, 266–275			✓					✓

Table 26 (continued)

Paper	Black	Red	Pink	Purple	Blue	Green	Yellow	Authors
Rago, A. b.; Marcos, C. c. & Diaz-Pace, J. b**. Uncovering quality-attribute concerns in use case specifications via early aspect mining**. *Requirements Engineering,* 2011, 1–18						✓		
Gemino, A. & Parker, D. **Use case diagrams in support of use case modeling: Deriving understanding from the Picture**. *Journal of Database Management,* 2009*, 20*, 1–24							✓	
Hornbæk, K.; Høegh, R.; Pedersen, M. & Stage, J. **Use case evaluation (UCE): A method for early usability evaluation in software development**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2007*, 4662 LNCS*, 578–591		✓						
Cherfi, S.-S.; Akoka, J. & Comyn-Wattiau, I. **Use case modeling and refinement: A quality-based approach**. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics),* 2006*, 4215 LNCS*, 84–97					✓			✓
Issa, A. & Al-Ali, A. **Use Case Patterns Driven Requirements Engineering**. *Computer Research and Development, 2010 Second International Conference on,* 2010, 307–313			✓	✓				
Juarez-Ramirez, R.; Licea, G. & Cristobal-Salas, A. **Using Tree Diagram Concepts to model Use Case Flows**. *Current Trends in Computer Science, 2007. ENC 2007. Eighth Mexican International Conference on,* 2007, 157–164		✓			✓			

This appendix presents the papers included by each team and by the authors

Appendix 11

Table 27 Teams' Answers to the Research Questions – Quality Attributes for Use Case

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
Abstraction	–	–	Considered a specification issue in that there is a danger of mixing problem domain/internal design detail into the specification task where it is not always appropriate	–	ND	–	–
Accuracy	–	ND	–	ND	–	–	ND
Accessibility	–	–	–	–	–	This attribute is related to the understandability attribute as well. Improving accessibility in use case descriptions can improve communication on projects for non-technical stakeholders	–
Action completeness	–	–	–	–	–	–	"Complete action description" means a single event that is written in the mode of the content guideline
Ambiguity	–	ND	There should be no ambiguities in the use case descriptions or in the use of terminology	ND	ND	–	There should be no ambiguities in the use case descriptions or in the use of terminology
Analytical	There should be no concept related to computational solution, including user interface	ND	It should only describe what the system should do. This includes the exclusion of any design or implementation decisions,	ND	ND	–	–

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
			including interface details. Except those explicitly defined by the customer.				
Appropriate action description	–	–	–	–	–	–	ND
Appropriateness	–	ND	–	–	–	–	–
Availability	–		–	–	–	This attribute is present in models where a use case action is always available to be performed in a system	–
Changeability	ND	ND	–	–	–	This attribute is a feature present in use case models which do not demand a huge effort to modify their implementations	–
Clarity	–	ND	Consist on quality attribute that permits the detection of problems with requirements and that structure is most clear for the stakeholders	–	–	–	–
Cogent	ND	–	The description of the flow of the use case should follow a logical path with events being stated in the correct order in the description; it should be a complete end-to-end transaction, and it should provide a rational answer to a problem	–	–	–	The cogent facet contains three elements: text order, dependencies, and rational answer
Coherence	ND	–	Description of each event should repeat a noun from the last or a previous event	ND	–	–	This facet implies good local coherence and global

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
							coherence for better understanding of the text
Communicability	–	ND	Ability to extract information from the descriptions	–	ND	–	System responses not well described to make clear what information was intended to be transmitted to the user
Completeness		ND	The underlying requirements must correctly be represented by the UC diagram and textual descriptions. This means that all information and facts that are expected to be in the UC descriptions and diagram must be present	ND	ND	This attribute is relevant to avoid problems like missing interactions on use case models. It can also be divided into two features: action completeness and UC completeness. The action completeness attribute is focused only on the action flow presented in the use case, but the UC completeness is concerned with the whole use case model characteristics completeness	All information and facts that are expected to be in the UC descriptions and diagram must be present
Complexity	ND	ND	–	ND	ND	–	–
Complexity of source code	–	–	–	–	–	This attribute is related to the level of use case implementations modularization that can impact on the complexity of source code, increasing the effects of coupling, cohesion and size metrics	–

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
Compliance	–	ND	–	–	–	–	–
Comprehension effectiveness	–	–	–	–	–	–	How well the model is understood, as reflected by the number of total correct answers in the comprehension tasks
Comprehension efficiency	–	–	–	–	–	–	The effort required to understand the model, as reflected by the time taken to perform the comprehension tasks
Conformity	–	–	Right functional requirements have been described	–	–	–	
Consideration of alternatives	ND	–	–	–	–	–	Consideration of alternatives suggests there be separation (a separate section) for alternative and exceptional flows. The viability of these exceptions and alternatives is important. Numbering is a check to make sure that the alternative and exceptional paths listed correspond exactly to their equivalent in the main flow
Consistency	Diagram and textual descriptions must be consistent with each other	ND	There should be consistency among all the elements of the use case model. The structure of the use cases and the use of language,	ND	ND	This feature is important to avoid issues like redundancy and conflicting or nonsensical interactions between use cases in models.	Consistency comprises: • Consistent Structure: consistent structure requires that variations be kept out of the main flow of events.

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
			grammar and terminology should be consistent across all use cases			Inconsistency in requirements also occurs when there are incorrect actions, requirements clashes, and bad dependencies	• Consistent Grammar: the structure of the use cases and the use of language and grammar should be consistent across all use cases; use of present simple tense without adverbs, adjectives, pronouns, synonyms, and negatives
Consistent Abstraction	ND	–	–	–	–	–	–
Consistent grammar	ND	–	–	–	–	–	–
Consistent level of abstraction	–	–	–	–	–	–	The model should not contain any design or implementation decisions, including interface details
Consistent structure	ND	–	–	–	–	–	–
Content required	–	–	–	–	–	–	There should be no unnecessary details about user interface or internal design. Each event should be atomic, that is, sentences with more than two clauses should be avoided
Correctness	Diagram and textual descriptions must represent system requirements correctly	ND	Underlying requirements must correctly be represented by the UC diagram and textual descriptions. This means that all information and facts that are expected to be in the UC descriptions and diagram must be present	ND	–	This feature is recognizable in projects where there are no wrongly sequenced interactions or wrong interactions in use case structures	The underlying requirements must be represented correctly without information that misrepresents the requirements. The use case modeling technique is used correctly

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
Coverage	ND	–	–	–	–	–	The notion of coverage introduces attributes of scope (too much information) and span (not enough information). These can be viewed as finer-grained completeness, a suggested quality good use cases should portray
Defect rate	–	ND	–	ND	ND	–	–
Dependability	–	–	–	–	–	This feature reunites important characteristics in use case models, like the availability, reliability, survivability and fault tolerance. These features indicate the level of dependence in use case interactions	–
Deployability	–	–	–	–	–	This quality feature involves concepts like distributability and configurability. Such concerns can also be related to the changeability attribute and the importance to manage these characteristics in use case implementations	–
Easy of learning	–	–	–	–	–	–	–
Easy to automate	–	ND	–	–	–	–	–
Efficiency	–	–	–	ND	–	–	–
Feasible completeness	–	–	–	–	ND	–	–

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
Formality	–	ND	–	–	–	–	–
Granularity	–	–	Consist in the size of model unity	ND	ND	–	–
Homogeneous	–	–	–	–	–	This attribute is related to the small size feature. It recommends the implementations of homogeneous use cases, when they handle approximately the same number of steps and requirements, to avoid the concentration of requirements in a single, unbalanced use case	–
Integrability	–	–	–	–	–	This feature highlights the capacity of use case models to be available to adaptation, what demands means to support interoperability and composability of components	–
Interoperability	–	–	–	ND	–	–	–
Interdependency	–	–	–	ND	–	–	–
Lack of notion of atomicity	–	ND	–	–	–	–	–
Learnability	–	–	–	ND	–	–	–
Level of abstraction	–	ND	–	–	–	–	–
Maintainability	–	ND	The use case model should be such that changes to it can be made completely and consistently without	ND	ND	This attribute can be related to others quality attributes but is especially related to the changeability feature	–

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
			changing the structure and style of the use cases				
Maturability	–	–	–	ND	–	–	–
Modifiability	–	–	–	ND	–	–	ND
Operability	–	–	–	ND	–	–	–
Over-specification	ND	–	–	–	–	–	–
Perceived comprehensibility	–	–	–	–	–	–	The participants' opinion regarding the ease of understanding the diagrams and descriptions
Perceived ease of use	–	–	–	–	ND	–	–
Perceived usefulness	–	–	–	–	ND	–	–
Performance	–	–	–	–	–	The performance of use case models is achieved when their implementations are concerned with accurate response, and there are features like schedulability and scalability	–
Plausibility	–	–	How realistic are the use cases	–	–	–	The flow of events was realistic, that is, the events follow a logical and complete sequence, and it is clearly stated where variations can occur
Pluggability	–	–	–	–	–	This feature is especially highlighted in projects where the effort to plug/unplug a use case in the system is not high	–

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
Portability	–	–	–	–	–	This quality attribute can be related to the reusability feature, where a single use case can be used in another requirements specification	–
Precision	–	ND	–	–	–	–	–
Problem solving	–	–	–	–	–	–	It's the ability to use knowledge gained to answer related problems not directly answerable from presented material
Readability	–	ND	The use case descriptions should be precise and use a common terminology. There should be no duplicated information, very large or very small use cases, or presumptions regarding how an actor will perform its own internal tasks. Each use case should follow a logical path to a clear goal	ND	ND	This feature is directly related to the understandability attribute. Improving readability of use cases for non-technical people can significantly improve communication on projects	The use case descriptions should be precise and use a common terminology
Recoverability	–	–	–	ND	–	–	–
Redundancy	ND	ND	There should not be superfluous information	ND	–	–	The model should also not contain repeated information as this may lead to confusion
Retention	–	–	–		–	–	Inception level of information. It's the comprehension over time of material being presented

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
Reusability	–	ND	Consist in reuse the models on another project	ND	–	This attribute can be an additional indicator of quality in use case models. It shows the ability of a specification to use the same use case in different projects	–
Safety	–	–	–	–	–	This feature is recognizable in use cases where the risk of occurrence of errors is minimum	–
Security	–	–	–	ND	–	This attribute implies the necessity of adding means to avoid and recover from runtime faults	–
Separation of concerns	–	–	–	–	–	This feature is related to the degree of scattering over classes (DOSC), the concern diffusion over components (CDC), and the lines of code (LOC) metrics for all use case implementations in the system for each implementation	–
Single diagram	–	–	–	–	–	–	The use case model should include one single diagram showing all the actors and use cases
Size	–	ND	–	–	–	–	–
Small size	–	–	–	–	–	This feature tries to focus on the recommendation to avoid	–

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
						large use cases in models. This is considered a problem because a large use case is trying to handle several different requirements at the same time or there are many alternative flows and steps	
Specification completion	–	–	–	–	ND	–	–
Stability	–	–	–	ND	–	–	–
Support for parallel development	–	–	–	–	–	This feature is related to the amount of code that is shared between all use cases in the system. In a situation where the parallel development is better supported, use cases code is highly spread over the system components	–
Terminology correction	–	–	–	–	–	This feature highlights the importance of avoiding terminology errors such as synonyms, homonyms, and ambiguous use of pronouns	–
Testability	–	–	–	ND	–	This feature is not observable at runtime but over the product lifecycle. It is also related to the maintainability quality attribute	–
Traceability	ND	ND	–	–	ND	This attribute is related to the necessary effort to trace changes in use case	–

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
						specification from requirements to implementation	
Understandability	It must be legible and unambiguous	ND	The model must be presented in a readable form. The information contained in the UC descriptions must be precise and unequivocal. The model should also not contain repeated information as this may lead to confusion. All stakeholders must share a common understanding of the presented functional requirements	ND	ND	This attribute highlights the importance of the use of appropriate terms, especially to non-technically minded stakeholders. The more details appear in use cases, the more confusing and difficult to understand the descriptions become for non-technical stakeholders	All stakeholders must share a common understanding of the presented functional requirements
Usability	–	ND	–	–	–	This quality attribute is related to the administrability of use cases, the level of management a developer has when dealing with a use case implementation	–
Vagueness	–	ND	–	–	–	–	–
Verifiability	ND	ND	The use case model should be such that it can be checked, automatically or by humans, whether the finished system adheres to the use case model	ND	–	–	–
Verification effectiveness	–	–	–	–	–	–	

Table 27 (continued)

Quality Attribute	Quality Attribute Definition						
	Black	Red	Pink	Purple	Blue	Green	Yellow
							How well the model is understood, as reflected by the number of total correct answers in the verification tasks
Verification efficiency	–	–	–	–	–	–	The effort required to understand the model, as reflected by the time taken to perform the verification tasks

This appendix concerns the use cases quality attributes the teams reported

ND means that the team listed the quality attribute in its final report, but provided no definition for it

References

Babar, M. A., Zhang, H (2009) Systematic literature reviews in software engineering: preliminary results from interviews with researchers. Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement. Lake Buena Vista: IEEE

Basili VR (1992) Software modeling and measurement: the goal/question/metric paradigm. Technical Report. University of Maryland at College Park: College Park, MD, p 24

Biolchini J et al (2005) Systematic review in software engineering. Federal University of Rio de Janeiro. Rio de Janeiro, p 31. (RT-ES 679/05). Available at: http://www.cos.ufrj.br/uploadfile/es67905.pdf. Accessed 17 Aug 2017

Brereton P (2011) A study of computing undergraduates undertaking a systematic literature review. IEEE Trans Educ 54(4):558–563

Carver JC et al (2013) Identifying barriers to the systematic literature. Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. Baltimore: IEEE, p 203–213

Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46

Condori-Fernandez N et al (2009) A systematic mapping study on empirical evaluation of software requirements specifications techniques. Proceedings of the 3rd International Symposium on Empirical Software Engineering and Measurement. Lake Buena Vista: IEEE, p 502–505

Corbin, J.; Strauss, A. (2007) Basics of qualitative research: techniques and procedures for developing grounded theory. 3. ed. [S.l.]: Thousand Oaks, SAGE Publicationse. ISBN 978-1412906449

Dias Neto AC et al (2007) Characterization of model-based software testing approaches. PESC/COPPE/UFRJ. Rio de Janeiro. (ES-713/07). Available at: http://www.cos.ufrj.br/uploadfile/1188491168.pdf. Accessed 17 Aug 2017

Diest O, Grimán A, Juristo N (2009) Developing search strategies for detecting relevant experiments. Empir Softw Eng 14(5):513–539

Dybå T, Kitchenham B, Jørgensen M (2005) Evidence-based software engineering for practitioners. IEEE Softw 22(1):58–65

Fantechi A et al (2002) Application of linguistic techniques for use case analysis. Proceedings of the IEEE Joint International Conference on Requirements Engineering. Essen: IEEE, p 157–164

Garousi V, Eskandar MM, Herkiloglu K (2016) Industry–academia collaborations in software testing: experience and success stories from Canada and Turkey. Softw Qual J:1–53. https://doi.org/10.1007/s11219-016-9319-5

Hassler E et al (2014) Outcomes of a community workshop to identify and rank barriers to the systematic literature review process. Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. London: ACM. No. 31

Jaccard P (1912) The distribution of the flora in the alpine zone. New Phytol 11(2):37–50

Kasoju A, Petersen K, Mäntylä MV (2013) Analyzing an automotive testing process with evidence-based software engineering. Inf Softw Technol 55(7):1237–1259. https://doi.org/10.1016/j.infsof.2013.01.005

Kitchenham B; Charters S (2007) Guidelines for performing systematic literature reviews in software engineering. Keele University and University of Durham. Keele/Durham, p 65. (EBSE-2007-01)

Kitchenham B et al (2011) Repeatability of systematic literature reviews. Proceedings of the 15th International Conference on Evaluation and Assessment in Software Engineering. Durham: IEEE, p 46–55

Kitchenham B; Brereton P; Budgen, D (2012) Mapping study completeness and reliability - a case study. Proceedings of the 16th International Conference on Evaluation and Assessment in Software Engineering. Ciudad Real: IET, p 126–135

Kuhrmann M, Fernández DM, Daneva M (2017) On the pragmatic design of literature studies in software engineering: an experience-based guideline. Empir Softw Eng, Springer, US, pp 2852–2891. https://doi.org/10.1007/s10664-016-9492-y

Lavallée M, Robillard P-N, Mirsalari R (2014) Performing systematic literature reviews with novices: an iterative approach. IEEE Trans Educ 57(3):175–181

López L, Costal D, Ayala CP, Franch X, Annosi MC, Glott R, Haaland K (2015) Adoption of OSS components: A goal-oriented approach. Data Knowl Eng 99:17–38. https://doi.org/10.1016/j.datak.2015.06.007

Losavio F et al (2004) Designing quality architecture: incorporating ISO standards into the unified process. Inf Syst Manag 21(1):27–44

MacDonell S et al (2010) How reliable are systematic reviews in empirical software engineering? IEEE Trans Softw Eng 36(5):676–687

Munir H, Moayyed M, Petersen K (2014) Considering rigor and relevance when evaluating test driven development: a systematic review. Inf Softw Technol 56(4):375–394

NHS Centre for Reviews and Dissemination, University of York (2002) The Database of Abstracts of Reviews of Effects (DARE). Effect Mat 6(2):1–4

Oates BJ, Capper G (2009) Using systematic reviews and evidence-based software engineering with masters students. Proceedings of the 13th International Conference on Evaluation and Assessment in Software Engineering. Durham: British Computer. Society:79–87

Pai M et al (2004) Systematic reviews and meta-analyses: an illustrated, step-by-step guide. Natl Med J India 17(2):86–95

Petersen K; Ali NB (2011) Identifying strategies for study selection in systematic reviews and maps. Proceedings of the 5th International Symposium on Empirical Software Engineering and Measurement. Banff: IEEE, p 351–354

Petersen K et al (2008) Systematic mapping studies in software engineering. Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering. Bari: British Computer Society

Petersen K, Vakkalanka S, Kuzniarz L (2015) Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 64(1):1–18

Phalp KT, Vincent J, Cox K (2007) Assessing the quality of use case descriptions. Softw Qual J 15(1):69–97

Preiss O; Wegmann A; Wong J (2001) On quality attribute based software engineering. Proceedings of the 27th Euromicro Conference. Warsaw: IEEE, p 114–120

Rago A, Marcos C, Diaz-Pace JA (2013) Uncovering quality-attribute concerns in use case specifications via early aspect mining. Requir Eng 18(1):67–84

Rainer A, Hall T, Baddoo N (2006) A preliminary empirical investigation of the use of evidence based software engineering by undergraduate students. Proceedings of the 10th International Conference on Evaluation and Assessment in Software Engineering. Keele: British Computer. Society:91–100

Ramos R et al (2009) Quality improvement for use case model. Proceedings of the 23rd Brazilian Symposium on Software Engineering. Fortaleza: IEEE, p 187–195

Riaz M et al (2010) Experiences conducting systematic reviews from novices' perspective. Proceedings of the 14th International Conference on Evaluation and Assessment in Software Engineering. Swinton: British Computer. Society:44–53

Shull F, Rus I, Basili V (2000) How perspective-based reading can improve requirements. Computer 33(7):73–39

Travassos GH et al (2008) An environment to support large scale experimentation in software engineering. Proceedings of the 13rd IEEE International Conference on Engineering of Complex Computer Systems. Belfast: IEEE, p 193–202

Ulziit B, Warraich ZA, Gencel C, Petersen K (2015) A conceptual framework of challenges and solutions for managing global software maintenance. Journal of Software: Evolution and Process 27(10):763–792. https://doi.org/10.1002/smr.1720

Viera AJ, Garrett JM (2005) Understanding interobserver agreement: the kappa statistic. Fam Med 37(5):360–363

Wohlin C (2014) Writing for synthesis of evidence in empirical software engineering. Proceedings of the 8th ACM/ IEEE International Symposium on Empirical Software Engineering and Measurement. Torino: ACM. No. 46

Wohlin C et al (2013) On the reliability of mapping studies in software engineering. J Syst Softw 86(10):2594–2610

Zhang H, Babar MA (2010) On searching relevant studies in software engineering. Proceedings of the 14th International Conference on Evaluation and Assessment in Software Engineering. Swindon: ACM, p 111-120

Zhang H, Babar MA, Tell P (2011) Identifying relevant studies in software engineering. Journal. Inf Softw Technol 53(6):625–637

Talita Vieira Ribeiro is a D.Sc. student working in the Experimental Software Engineering Group at the Federal University of Rio de Janeiro (UFRJ). She holds a masters degree in Systems Engineering and Computer Science from UFRJ, and a B.Sc. degree in Computer Science from the Federal University of Para (UFPA). Her current research interest are Evidence-based Software Engineering, and Source Code Review and Quality. Her work involves Knowledge Translation in the Software Engineering field.

Jobson Massollar is a postdoctoral reasercher at COPPE/UFRJ, Rio de Janeiro, Brazil. He is also an assistant professor at Veiga de Almeida University, Rio de Janeiro, Brazil. He currently teaches various courses related to Software Engineering. His research interests include Requirements Modeling, Model Drive Development, Software Verification and Validation, and Context Aware Systems.

Guilherme Horta Travassos (GHT) is a Professor of Software Engineering at COPPE/UFRJ and a CNPq (Brazilian Research Council) 1D Researcher. He holds a D.Sc. in Systems Engineering and Computer Science from COPPE/UFRJ, with a post-doc in Experimental Software Engineering at the University of Maryland/ College Park - USA. He heads the Experimental Software Engineering Group at COPPE/UFRJ and is a member of ISERN, SBC and ACM. Apart from that, he is an associate editor of Elsevier - Information and Software Technology (IST), and takes part in the editorial board of World Scientific – International Journal of Software Engineering and Knowledge Engineering (IJSEKE), Springer Open - Journal of Software Engineering Research and Development (JSERD), and e-Informatica Software Engineering Journal (EISEJ). Further information regarding his research interests and publications can be obtained at http://www.cos.ufrj.br/~ght.