Quality Assessment of Systematic Reviews in Software Engineering: A Tertiary Study

You Zhou
Software Institute
Nanjing University
Nanjing,Jiangsu,P.R.China
mg1332020
@software.nju.edu.cn

He Zhang
Software Institute
Nanjing University
Nanjing,Jiangsu,P.R.China
hezhang@nju.edu.cn

Xin Huang
Software Institute
Nanjing University
Nanjing,Jiangsu,P.R.China
mg1432004
@software.nju.edu.cn

Song Yang
Software Institute
Nanjing University
Nanjing,Jiangsu,P.R.China
ysong12
@software.nju.edu.cn

Muhammad Ali Babar
School of Computer Science
University Adelaide
Australia
ali.babar
@adelaide.edu.au

Hao Tang
Software Institute
Nanjing University
Nanjing,Jiangsu,P.R.China
mg1232011
@software.nju.edu.cn

ABSTRACT

Context: The quality of an Systematic Literature Review (SLR) is as good as the quality of the reviewed papers. Hence, it is vital to rigorously assess the papers included in an SLR. There has been no tertiary study aimed at reporting the state of the practice of quality assessment used in SLRs in Software Engineering (SE).

Objective: We aimed to study the practices of quality assessment of the papers included in SLRs in SE.

Method: We conducted a tertiary study of the SLRs that have performed quality assessment of the reviewed papers.

Results: We identified and analyzed different aspects of the quality assessment of the papers included in 127 SLRs.

Conclusion: Researchers use a variety of strategies for quality assessment of the papers reviewed, but report little about the justification for the used criteria. The focus is creditability but not relevance aspect of the papers. Appropriate guidelines are required for devising quality assessment strategies.

Categories and Subject Descriptors

D.2.0 [**Software Engineering**]: General

General Terms

Experimentation

Keywords

Systematic (Literature) Review, Quality Assessment, Software Engineering

EASE '15 April 27 - 29, 2015, Nanjing, China

http://dx.doi.org/10.1145/2745802.2745815

1. INTRODUCTION

Evidence-Based Software Engineering (EBSE) [15] was introduced by Kitchenham, Dyba, and Jorgensen almost 11 years ago (i.e., 2004). Several hundreds of Systematic Literature Reviews (SLRs) on different topics of Software Engineering (SE) have been published since 2004. An SLR is only as good as the evidence they are based on. A central issue in SLR is the level of confidence that can be placed in the conclusions and recommendations arising from an SLR.

One of the reliable mechanisms of increasing the level of confidence in the findings of an SLR is to rigorously assess the quality of the primary studies included in an SLR. There are a few guidelines for designing and applying rigorous quality assessment of the papers included in an SLR in SE. There is no existing systematic review of the quality assessment mechanisms for SLRs in SE. Many researchers report the quality assessment of the primary studies included in an SLR as one of the components of reporting an SLR without having a good understanding of how to effectively perform quality assessment. It can be a significant challenge to design and apply a rigorous quality assessment to the papers included in an SLR.

In order to understand the quality assessment practices used in the reported SLR, we have carried out a review of the existing SLRs which claim to have assessed the quality of the papers included in those SLRs. We identified the SLRs that provide detailed information about their quality assessment (e.g., quality assessment instrument), and analyzed the existing quality assessment instruments and the number of criteria included in the quality assessment instruments. Their specific purposes of the quality assessments were classified into a few common purposes. We also did an analysis of the aspects that quality assessment focuses on when imple-

menting quality assessment in SLRs and compiled a list of the most often used criteria for each aspect. Due to the fact that SLR and Systematic Mapping Studies (SMS) may have different ways of assessing the quality of the papers included in an SLR or SMS, this research analyzes the purposes of the quality assessment and different aspects of the quality assessment in SLR and SMS separately. The possible differences between SLR and SMS could offer advice when researchers perform quality assessment in SLR and SMS.

The main contribution of this research is that it presents a ten-year overview of the use of quality assessment in SLRs by SE researchers. All quality assessment systems (e.g., checklists, guidelines, and quality questions for reuse) used in SLRs in SE have been enumerated in this paper. Hence, this work offers a valuable reference for the future systematic reviewers. To be specific, we also identified and classified the purposes of quality assessment of the existing SLRs, and analyzed their popularity among the published SLRs. We have also extracted and reported the common aspects found in different quality assessment instruments. For each aspect, researchers can gain the knowledge of the most concerned criteria, which is particularly helpful in developing and applying quality assessment instruments for future SLRs. Our research aims to help researchers to understand the state-of-the-practice of quality assessment in SE and enable their knowledgeable development of rigorous quality assessment systems, which can inevitably improve the quality and credibility of the findings and conclusions of SLRs.

The rest of this paper is structured as follows. In Section 2, the related work of this paper is introduced. Section 3 elaborates the research methodology of this SLR. In Section 4, the selected SLRs in this study are described. Section 5 answers the research questions based on the data extracted from the SLRs, and highlights the important findings. Section 6 and 7 discuss the threats to validity and draw the conclusions respectively.

2. RELATED WORK

2.1 Tertiary Studies

Since SLR is a relatively new research methodology in SE, there have not been many tertiary studies of SLRs in SE. We have identified two generic categories of important series of tertiary studies as well as some other studies that can be considered related to our work reported in this paper.

Kitchenham et al. conducted two tertiary studies to provide an overview of the secondary studies related to EBSE in 2007 [13] and 2009 [14]. Later, da Silva et al. updated the tertiary studies conducted by Kitchenham and her colleagues until the end of 2009 [3]. These three articles reviewed 120 SLRs in total. They show the changes of SLRs in SE over time, including the coverage of topics, the concentration on a few topics, the number of researchers and organizations that became more globally distributed. Also, their findings address several limitations with the use of SLR in SE.

Compared to these tertiary studies that mainly focus on the general methodology of SLR, there are some other tertiary studies that have addressed some particular aspects of SLRs. Cruzes and Dybå performed another tertiary study for assessing the types and methods of research synthesis used in SLRs in SE [2]. They restricted their search scope and selection criteria to investigate the SLRs explicitly influenced by EBSE [7, 15] and SLR guidelines [11, 12]. These guidelines provide a standard procedure for guiding researchers to perform a high-quality SLR in SE.

Santos and da Silva carried out a research on critically appraising the use of SLR with respect to the types of research questions asked in the published reviews; they were interested in finding out how the questions were presented in the reports, and how the questions were used to guide the search of primary studies [4]. In order to understand the study selection strategies used by researchers to reduce bias and resolve disagreements in SLRs, Petersen and Ali contributed a set of strategies that can be followed by researchers [18] for making informed decisions. Our research reported in this paper focuses on another important component of an SLR, i.e., quality assessment of the papers reported in the published SLRs.

2.2 Quality Assessment

For assessing the quality of individual primary studies and grading the overall strength of a body of evidence, there exists many influential scales in software engineering, medicine, and sociology. Among them, some are used to assess experiment and randomized controlled trials. One of the most well-known and widely used scales for assessing randomized controlled trials is the scale developed by Jadad et al.[10]. Another assessment tool, the Critical Appraisal Skills Programme (CASP) which is at the Public Health Resource Unit in Oxford, has been widely adapted in both medicine and healthcare [9]. This appraisal tool contains three broad issues when appraising the report of a systematic review: a) *Are the results of the review valid?* b) *What are the results?* c) *Will the results help locally?*

It contains ten questions which are not all suitable for software engineering. Dybå and Dingsøyr[5] developed eleven criteria based on the CASP and principles of good practice for conducting empirical research in SE proposed by Kitchenham et al.[16]. Sjoberg et al.[20] discussed measures to increase the quality of empirical studies in SE in general, while Kitchenham et al.[16] have proposed a set of more concrete guidelines to assist researchers in performing empirical studies. The revised ten questions are universal questions and could be easily applied in SE. systematic reviewers can form their own quality assessment checklists by selecting and revising several questions from the original questions. A large number of systematic reviewers choose to develop their own checklists based on the CASP.

The Grade of Recommendation Assessment, Development and Evaluation (GRADE) Working Group has developed a system for grading the quality of evidence and strength of recommendations [17]. The approach defines the quality of a body of evidence as four grades of evidence where one can be confident that an estimate of effect or association is correct. The GRADE system initially categorizes evidence concerning study design by assigning randomized experiments a high grade and observational studies a low grade. However, by considering the quality, consistency, and directness of the studies in the evidence base, the initial overall grade could be

increased or decreased. Compared to CASP and the revised version by Dybå and Dingsøyr, the overall grade is determined by several aspects. As a result, GRADE assesses the evidence in a more precise way and is a kind of quantitative approach. To use this system, researchers should combine several basic components. GRADE is a complicated approach and may be difficult for new SLR researchers.

3. METHOD

We adapted the comprehensive guidelines specified by Kitchenham and Charters [11] and employed Quasi-Gold Standard (QGS) in developing the search strategy. QGS is an approach to devising a rigorous search strategy for improving the validity and reliability of an SLR's search phase[23]. The review was mainly conducted by four researchers. The three research students independently searched, selected SLRs, and extracted the information of the quality assessment of the selected SLRs. Their supervisor played the role of expert who dealt with divergences between the three students during the process.

3.1 Research Questions

According to our research objective, the primary Research Question (RQ) of this tertiary study is "*What is the state-of-the-art of quality assessment in systematic literature reviews in software engineering?*" This question can be decomposed into the following more answerable research questions.

- RQ1: What are the existing quality assessment systems applied in systematic literature reviews in software engineering?
- RQ2: What are the purposes of quality assessment in systematic literature reviews?
- RQ3: What aspects do the existing systematic literature reviews focus on when assessing the quality of primary studies?

3.2 Search Strategy

Even a well designed and performed search process is less likely to be able to collect relevant studies completely. In this review, the approach of Quasi-Gold Standard (QGS) proposed by Zhang et al.[23] was adopted. The approach systematically integrates manual and automated search strategies and suggests a relatively objective and rigorous approach to improve the quality of the search strategy for an SLR in terms of sensitivity and precision. Our search consists of three stages: manual search, automatic search, and snowballing. The detail of each search stage is described as follows.

Manual Search

The manual search was initiated in June 2014. At this stage, the authors jointly chose the venues (e.g., journals, conference proceedings) recognized as highly specific to Empirical Software Engineering (ESE) and EBSE as well as highly reputed generic publication venues in SE. After carefully considering the venues available in SE community, the authors selected six of them for manual search: EMSE, ESEM, EASE, TSE, IST and JSS.

We reused the list of the SLRs identified by Zhang et al. [22] which ranged between 2004 and 2010. The reason that we selected the study list of Zhang et al.[22] as our starting set rather than [3] is the quality issues of the latter tertiary study. Note that the purpose of the manual search is establishing an effective quasi-gold standard for improving the follow-up automated search instead of striving to capture as many SLRs as possible, thus the selected venues are slightly different to the one used in Zhang et al.'s tertiary study. A few previously searched (but not specific to ESE and EBSE) venues were ignored at the manual search stage in this study.

During the search, we found that the SLRs published between 2004 and 2010 have some different characteristics from those published later. Using the same search string from the previous study, we missed some SLRs, most of which are mapping studies. Considering that the search string from the former review stage was validated with high quasi-sensitivity, we just modified the string and recoded the following search string into equivalent forms to match the syntax of each digital library. Also, the quasi-sensitivity was calculated to evaluate the string and adjust it if needed.

> (software AND (((systematic OR controlled OR structured OR exhaustive OR comparative OR evidence) AND (review OR survey OR (literature search) OR map)) OR (mapping study) OR (scoping study) OR (systematic map) OR (tertiary study) or meta-analysis))

Sensitivity is an important metric for evaluating the quality and efficiency of a search strategy. It refers to the proportion of relevant studies covered by the QGS and can be calculated as:

$$Sensitivity = \frac{Number\ of\ relevant\ studies\ retrieved}{Total\ number\ of\ relevant\ studies} 100\%$$

In the evaluation of the above search query, a few studies were missed. The corresponding quasi-sensitivity reached 85.6% and quasi-precision reached 1.55%. It confirms the updated search string is valid for carrying out automated search.

Automated Search

The automated search was conducted through four of the major publishers' digital library portals in [21]: IEEE Xplore, ACM Digital Library, ScienceDirect, and SpringerLink. These data sources were used in order to maximize the number of candidate papers to be located and collected. Because the seminal paper of EBSE [15] and the first guidelines (technical report) for performing SLRs in SE [12] were both published in 2004, we restricted the search scope between 2004 and 2013 inclusive. The search was performed by searching the fields of *title*, *keyword* and *abstract* of the publications.

Snowballing

Table 1: Quality assessment guidelines in SLRs

Ref	Name of guidelines
[15]	Evidence-based software engineering
[7]	Evidence-based software engineering for practitioners
[7]	Procedures for undertaking systematic reviews
[11]	Guidelines for performing systematic literature reviews in software engineering (version 2.3)
[1]	Systematic review in software engineering

The four digital libraries used for automated search are also the most important publishers of software engineering research. Although the automated search covered the majority of SE publications, we might have still missed some studies despite using the QGS strategy. To identify as many SLRs as possible, we further employed the snowballing strategy to seek even more SLRs. We used Google Scholar to check the papers associated with the three EBSE seminal papers and two versions of the guidelines on SLRs in SE. We checked the studies list which cited them for the assumption that an SLR in SE with sufficient quality would cite them. These papers and guidelines are listed in Table 1.

The search strategy was developed by three research students and reviewed by their supervisor to ensure not to miss the studies. These research students performed the manual search, automated search, and snowballing. We assigned the workload to make sure that each study was checked by at least two researchers independently to minimize the potential impact of any bias. The disagreements on search and selection opinions were solved in joint discussion. Any disagreements that could not be solved were escalated to the supervisor.

3.3 Inclusion and Exclusion

A set of inclusion and exclusion criteria were specified based on the analysis needs and the quality of the found papers to guarantee that only SLRs having quality assessment were included. A study identified after applying the search strategy would be selected if it met all the predefined inclusion criteria, or be eliminated if it met any of the predefined exclusion criteria. With respect to the objectives and research questions of this tertiary study, the following inclusion and exclusion criteria were applied:

Inclusion criteria:

1. The abstract or title has to explicitly state that the article is a type of systematic literature review.
2. The paper is in the area of software engineering.
3. The paper is peer reviewed (journal article, conference paper).
4. The paper is a regular paper or a full paper.
5. The paper includes a part of quality assessment.
6. The full-text of the paper can be accessible.

Exclusion criteria:

1. The paper is not written in English.
2. The paper's full-text is not accessible.
3. Any gray publication without peer-review, e.g., technical reports.
4. The paper is explicitly a short paper or with less than six pages.

All of these criteria were used to filter the SLRs which contain quality assessment in SE. For this reason, we excluded grey literature (i.e., non peer reviewed). The papers with less than six pages were also excluded because we observed that those papers could not include sufficient details about the procedure of performing SLR, in particular quality assessment.

3.4 Data Extraction

The data extracted from each selected secondary study (SLR) are listed in Table 2.

Table 2: Data extraction form

Attribute	Description
Title	The title of the selected papers.
Year	The publication year of the selected papers.
Country	The countries where authors' affiliations are situated.
Study type	The type of the study:meta analysis, SLR or mapping study.
Guideline(s)	The guidelines which they adapted the quality assessment from.
Criteria	The criteria they used to assess the quality.
Number	The number of the criteria they used.
Purpose(s)	Whether they were willing to accomplish the purpose. (cf. P1-P5 in Section 5.2)
Aspect(s)	For each aspect, which criterion was used to assess the aspect. (cf. A1-A4 in Section 5.3)

Data extraction has three main constituents: 1) general citation information of the publication; 2) the information about study type to find whether different study types made differences in their results; 3) the detailed information about the checklists or guidelines used in the quality assessment. The Purpose field refer to the five purposes (originally defined in [11]) that researchers may intend to accomplish by quality assessment. The aspect attribute includes the common aspect(s) that the quality assessment may focus on. (More detail about the *purposes* and *aspects* can be found in Section 5.2 and 5.3)

4. RESULTS

4.1 Selected Studies

Based on the list of studies, 2004-2010, included in [22], we retrieved 96 SLRs between 2011 and 2013 in manual search stage. And the results of the automated search stage are listed in Table 3. In snowballing stage, we identified 260 SLRs. After inclusion and exclusion stage, there were 127 SLRs left. All of these SLRs claim to have performed the quality assessment of the papers included in their respective SLRs. However, after extracting the data from these SLRs, we found that 17 SLRs that did not report their quality assessment criteria inside the paper or somewhere available to public. Therefore, we divided the 127 SLRs into two lists: 110 SLRs with explicit quality assessment criteria reported, listed in Appendix A, and the other 17 papers (without sufficient information about the quality assessment claimed to have been used) are listed in Appendix B. The data extracted from the 110 SLRs was used to answer the research questions for this SLR as reported in the next section.

Table 3: Summary of search results

Database	No. of retrieved SLRs	No. of selected SLRs
IEEE Xplore	989	133
ACM DL	129	38
ScienceDirect	244	73
Springer	18274	61
Total	19636	305

Considering the publication year of these SLRs, We have observed that an increasing number of SLRs report the details of the quality assessment performed on the included papers except 2011 (shown in Figure 1).

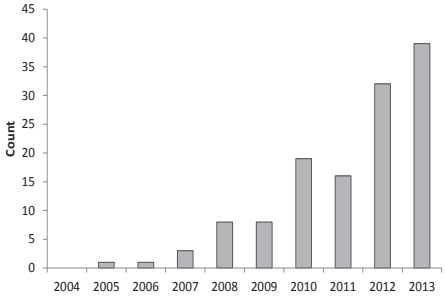

Figure 1: Distribution of Publication Year of SLRs

Among the selected 110 SLRs, 75 studies are systematic literature reviews (SLRs) and 35 are systematic mapping studies (SMSes). Since quality assessment is not a necessary component of of SMSs [11], researchers did not tend to carry out quality assessment in SMSes. Thus, compared to SLRs, it was not expected that many SMSes would have performed the quality assessment of the papers included in the SMSes.

In terms of the paper type, 57 papers are journal articles and 53 papers are conference or workshop papers. Considering their publication venues, the top five venues are IST(31), ESEM(7), EASE(5) JSS(4), and TSE(3). These five venues are all recognized as highly specific to empirical and evidence-based software engineering or highly reputed generic publication venues in SE. In particular, the journal of Information and Software Technology (IST) dominated in all SE publication venues.

Among the 17 exceptional SLRs, six claimed to have performed the quality assessment; but we failed to find the criteria they had used ([SP12] [SP15][SP4][SP14][SP10][SP17]). For instance, the authors in [SP17] cross-checked other results but included no quality assessment question list.

4.2 Study Distribution

The demographic data provides an overview of the population of the selected SLRs with quality assessment. We only collected the affiliation information of the first author per SLR, which reaches 25 countries. The information about the continental distribution of the organizations is illustrated in Figure **??**. It shows that the European researchers prefer to have quality assessment included in their SLRs more than the authors from other regions. The European researchers were involved in 44 SLRs. Looking into Europe, the most SLRs with quality assessment come from Sweden(14) and Spain(14), followed by UK(9). However, the Brazilian researchers contributed 16 SLRs with quality assessment, which is the largest number. For individual researchers, we identified the top four researchers who contributed more to this topic than others. They are Richard Torkar (6) and Robert Feldt (5) whose institutions affiliate to Sweden, and Barbara Kitchenham (5) and Mahmood Niazi (6) who worked for Keele University in UK. All of them came from European.

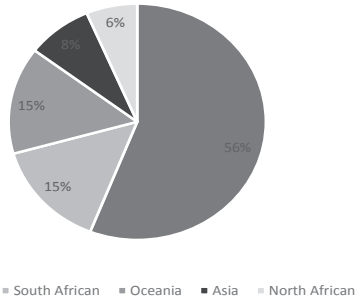

Figure 2: Continent Distribution of Publications

5. DISCUSSION

In this section, we discuss the research questions based on the data from the 110 papers listed in Appendix A.

5.1 RQ1:What QA Systems?

We classified these papers according to the checklists from the guidelines related to the quality assessment in ESE and EBSE. The result is shown in Table 4.

Table 4: Guidelines for quality assessment

Ref	Guideline	No. of SLRs
[5]	Dybå and Dingsøyr 2008	26
[11]	Kitchenham andCharters2007	15
[12]	Kitchenham 2004	9
[6]	Dybå,Dingsøyr,Hanssen2007	5
[S31]	Dybå and Dingsøyr 2008	2
[13]	Kitchenham and Brereton2009	2
[9]	CASP	2
	Others	14

When it comes to the papers which refer checklists from the guidelines, we found many papers developed their quality assessment checklists based on previously developed and used quality assessment checklists that can also serve as some sort of guidelines. The most cited checklist was presented by Dybå and Dingsøyr [5]. That quality assessment checklist has eleven criteria which were informed by those proposed for the Critical Appraisal Skills Programme(CASP) (in particular, those for assessing the quality of qualitative research [8]) and by principles of good practice for conducting empirical research in software engineering [16]. These eleven criteria cover four main aspects pertaining to the study quality which need to be considered: Reporting, Rigor, Credibility and Relevance (cf. Section 5.3). There are 26 SLRs which refer to the checklists proposed in [5]. We further found that more than half of these papers adopted entire or part of the checklist from [5] without any change, including seven papers ([S39][S23][S26][S51][S57][S87][S104]) adopting the entire list and six papers ([S19][S29][S14][S28][S106][S56]) with a part of the eleven criteria.

Kitchenham et al. [12] [11] suggested the quality relating the extent to which the study minimizes bias and maximizes internal and external validity. Furthermore, they introduced how to derive the checklists using the quality instrument. Compared with [S31], they provide more details and instruments about the quality assessment criteria.

Although most of the papers conducting the quality assessment refer to the checklists from the guidelines, some researchers chose to define checklists for quality assessment themselves, such as [S17][S44]. These researchers may have their special preferences and concerns about study quality

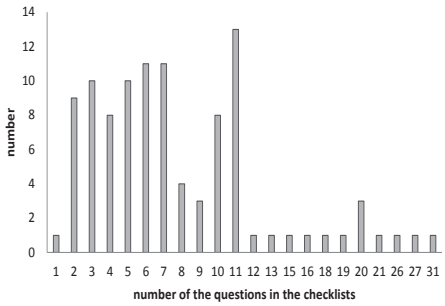

Figure 3: Number of questions of quality checklists

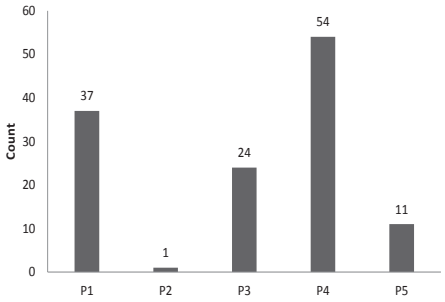

Figure 4: Number of SLRs of Each Purpose

that differ from others' choices.

There are also some other interesting findings in our research. We carried out an analysis on the number of the questions in the checklists which is shown in Figure 3. Most of the selected SLRs setup their quality checklists with two to seven questions. In some extreme cases, one paper includes one question only in its checklist [S63] – '*Is it clear how the factors for establishing/building trust between clients and vendors were identified in offshore software outsourcing relationship?*'. Such a question, however, may need plentiful information from the paper. In contrast, another paper [S37] which presents 31 questions in the checklist that pays more attention to the details of the primary studies, such as "*Is there a rationale for why the study was undertaken?*","*Do the authors define/describe all treatments and all controls?*" and so on. This paper divides the checklist into eight parts which assess different aspects of the primary studies. ***Questions relevant to the special topics being reviewed are scarce to present. Researchers prefer to use questions which focus on the general condition of the primary studies to assess the quality.***

We carried out a survey on how the questions in the checklists are presented. As a result, nearly all of the questions in the checklists are close-end. Only two selected papers ([S38][S81]) present open-ended questions in their checklists. The reason might be that the answers to open-ended questions are mostly qualitative, it is harder to gain a quantitative assessment for comparison than closed-ended questions. Gonz'alez et al. [S38] conducted the quality assessment and data extraction at the same time using open-ended questions. The other paper [S81] presents open-ended questions to describe how rigorous the process of the study is, so that the trust of industry practitioners and other researchers can be gained.

Apart from the guidelines listed in Table 4, we also found three papers which define their checklists by referring to some other SLRs. Catal et al. [S21] adopted the quality assessment criteria developed in Afzal and Torkar's SLR [S2] without any change. Liu [S55] adapted the contexts and scoring rule of the checklist from Zhang et al.'s SLR [S110] (six criteria from Zhang et al.'s SLR and the other six criteria defined by himself). Svahnberg et al. [S95] presented the quality assessment criteria based on the recommendation in Kitchenham et al.'s study [S49]. In these three cases, the referring SLRs have the similar topics to the referred previous SLRs. ***If there exist any former SLRs on the similar topic, researchers may consult these SLRs in developing their criteria for quality assessment.***

5.2 RQ2: Why Quality Assessment?

The purpose of quality assessment refers to what researchers try to accomplish when carrying out the quality assessment of the papers included in an SLR. Not all the papers present the purposes of their quality assessment. We extracted the data of the purposes of quality assessment by identifying what an SLR claims to achieve or what is the actual use of the assessment results. After the data extraction, we found that 96 papers explicitly present their purposes or imply their uses. They were classified in terms of the purposes defined by Kitchenham [11] as follows:

- P1 (Selection): To provide still more detailed inclusion/exclusion criteria.
- P2 (Difference): To investigate whether quality differences provide an explanation for differences in study results.
- P3 (Weighting): As a means of weighting the importance of individual studies when results are being synthesized.
- P4 (Interpretation): To guide the interpretation of findings and determine the strength of inferences.
- P5 (Recommendation): To guide recommendations for further research.

We identified the purpose(s) of quality assessment in each SLR, and groups the SLRs per purpose as shown in Figure 4.

Selection. P1 becomes the most popular purpose among the five purposes except P4. A relatively large number of researchers use the quality assessment for paper selection. For most SLRs, the inclusion/exclusion activity normally happens in a few rounds of scanning of the candidate papers without further reading. After the initial selection, the full text should be checked when applying more specific quality criteria. This enables researchers to reach a more informative judgment for inclusion and exclusion. Although it is unknown whether most papers selected studies based on the quality assessment results, interestingly we found ten studies explicitly claiming not to exclude studies with quality criteria. For the mapping studies, they did not exclude studies due to the relatively low quality because of their objective

of providing a broad view of the studied research topic or a research area's immaturity ([S46][S36][S22]). The other systematic reviews give the following reasons: 1) they just indicated the overall quality of the current work available ([S32]); 2) the authors wanted to include as much research as possible regardless the immaturity of the area ([S41, S37, S86]); and 3) the study's perspective was the aggregation of the reported empirical literature ([S75, S61]).

Difference. In the Figure 4, only one paper fulfils the P2 and the authors are Kitchenham and her team ([S49]). Given the five purposes were proposed in Kitchenham et al.'s guidelines[11, 15], we may wonder the feasibility of P2 in software engineering because this purpose might cost too much effort to achieve. This purpose was originally adapted from other disciplines like medicine, but has not been confirmed by the researchers' practice in SE.

Weighting. We found 24 SLRs applied quality assessment for this purpose. Most of them used three-point scale questions, i.e., the score of individual question can be either '*yes*' (1), '*to some extent*' (0.5), or '*no*' (0), such as [S5, S80, S36, S66]. The others simply adopted the binary score (1 for *yes* and 0 for *no*). Compared to the three-point scale, the binary score is a relatively weak scoring system because it provides very limited assessment without knowing any detail about the quality falling in between '*yes*' and '*no*'. In some special cases, researchers assigned different weights in order to highlight the importance of certain criteria in their scoring systems [S55].

Interpretation. As defined in [11], "*The study quality assessment can be used to devise a detailed inclusion/exclusion criteria and/or to assist data analysis and synthesis*". Figure 4 indicates that about half of the SLRs with quality assessment reported aimed to guide the interpretation of findings. Among the 54 SLRs, only 14 papers explicitly claimed that they carried out the quality assessment for the purpose of P4, e.g., [S17, S93, S110]. By scanning the details of the criteria and looking for the actual use of the assessment results, we further inferred that the other 40 SLRs assessed the quality of the included papers for the similar purpose. The number is larger than the numbers of the SLRs with other purposes. The reason for its popularity might be that quality assessment, as the complement of data extraction, could provide more information for further analysis and synthesis of the data extracted, and is also able to reveal the credibility of findings and conclusions. ***Quality assessment could enable the readers to better understand the review results and gain more confidence about the conclusions drawn from the review***.

Recommendation. For P4, there are 11 SLRs that contain the questions like "*Does the paper discuss limitations or validity?*" [S41] for how the issues of current work was perceived and reported, and "*Did the study recommend the further continuous research?*" [S110] about how the future continuous research was considered. Of the all SLRs with quality assessment, this purpose is addressed in a relatively small portion but could provide important information to other researchers. ***When implementing the quality assessment, systematic reviewers could consider the criterion of the recommendations for future research, which may influence the ongoing work in a research area***.

Table 5: Purpose Matrix of SLRs

	P2	P3	P4
P1	[S93][S10][S18] [S104][S2][S74] [S51][S45][S60]	[S67][S14][S81]	[S95][S87][S42] [S52] [S40][S39] [S53][S16]
P2		[S23][S7][S33]	

There are also SLRs which completed multiple purposes. The Table 5 shows the SLRs with two quality assessment purposes. Furthermore, only three addressed three purposes, which are [S110, S19, S77]. No SLR was found with more than three purposes.

In addition to the above five purposes, researchers use quality assessment to achieve other purposes. One is to classify primary studies. For example, Pino [S73] classified each study with the extraction of the design type of each study according to the study design hierarchy for software engineering present in [12]. In [S18], a series of criteria were related to an evaluation of the eligibility of the study type, which was considered to be quantitative, qualitative and mixed-method scholarly research and grey literature. Another extra purpose we notice is to extract and structure the most useful information in [S38]. During our research, we only discovered these two extra purposes.

Among the 110 studies we analyzed, there were 14 studies that did not explicitly claim their purposes or we could not make any inference about the purposes of the quality assessment of the papers used in those studies from the information provided in the papers, e.g., [S73, S105]. They merely assessed the study quality and listed the criteria and results. Their discussions, however, did not relate to the quality assessment. ***Some reviews report the quality assessment as a component but with neither purposes given nor perceivable contribution to the review.***

5.3 RQ3: What Aspects to Assess?

According to Dybå and Dingsøyr [5], four main aspects can be considered when assessing the quality of the primary studies. They are *Reporting* (A1), *Rigor* (A2), *Credibility* (A3), and *Relevance* (A4), and are described as below:

- **Reporting**: The quality of the reporting of a study's rationale, aims, and context.
- **Rigor**: The rigor of the research methods employed to establish the validity of data collection tools and the analysis methods.
- **Credibility**: The credibility of the study methods for ensuring that the findings were valid and meaningful.
- **Relevance**: The relevance of the study for the software industry at large and the research community.

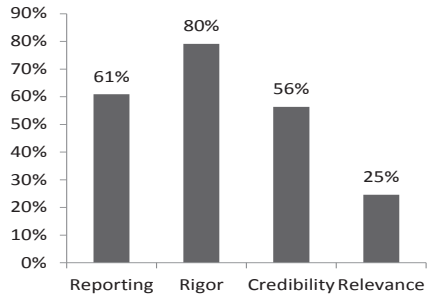

Figure 5: Percentage of SLRs per aspect

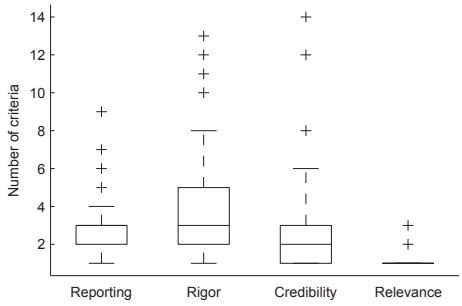

Figure 6: Distribution of quality criteria per aspect

Reporting often represents the minimum quality threshold of a review and is used to exclude non-empirical research papers. The other three main aspects pertain to the quality that need to be considered when appraising studies for a review [S31]: *Rigor* focuses on whether the research method used is thorough and appropriate; *Credibility* is related to whether the findings are well-presented and meaningful; and *Relevance* is the usefulness of findings to software industry and research community. The difference between *Rigor* and *Credibility* is that the former mainly emphasizes the validity and the description of the research methods which include data collection and data analysis methods; whereas, the latter concentrates on the statement of findings and the limitations of results. These aspects may cover a variety of quality characteristics of almost the whole research in general. In this review, we adopted these four aspects and the quality criteria reported in the selected SLRs are discussed according to these four aspects.

According to the extracted data, *Rigor* is identified as the quality aspect with the most concerns. About 80% of the selected SLRs assess the *rigor* of the research methods of the primary studies. In the meantime, *reporting* and *credibility* were also considered important in quality assessment. Over half of the selected SLRs assessed *reporting* (61%) and *credibility* (56%). The *relevance* of the primary studies, however, was considered by approximately one fourth of the SLRs only. The specific percentage of the SLRs' quality focus is shown in Figure 5. Moreover, Figure 6 shows the distribution of the number of criteria for each aspect. The mean values of the criteria number of *Reporting*, *Rigor*, *Credibility* and *Relevance* are 2.75, 3.82, 2.84, and 1.22 respectively. We also summarize the most frequently used criteria for each quality aspect that are listed in Table 6 in a descending order of their counting. Note that many criteria were represented in different manners in SLRs but share same or similar meaning. Hence, these different representations were combined into one criterion in terms of their actual concern.

Reporting. The quality of the *reporting* of a study's rationale, aims, and context is considered significant. The more complete, thorough and unambiguous the reporting is, the more detailed and accurate information can be extracted. Then, researchers could reach more informative findings and conclusions. Most of the SLRs used no more than four criteria in assessing the *reporting* quality (Figure 6). The most asked question in *Reporting* (Table 6) is: *Is there a clear statement (definition) of the aims (goals, purposes, problems, motivations, objectives, questions) of the research?*, though different reviews report varying forms of its representation. They use words such like goals, purposes, problems, motivations, objections and questions instead of aims to convey the same meaning. Next to this, the research context was also concerned by many SLRs.

Rigor. Figure 7 shows 87 of 110 SLRs assessed the research methods employed to establish the validity of data collection tools and the analysis methods. *Rigor* therefore becomes the most important aspect when researchers perform the quality assessment of primary studies. The methodological rigor of a study can directly influence the validity and the quality of its conclusions. With nearly four criteria for *Rigor* per SLR on average, unlike *Reporting*, the criteria are actually dispersed (Table 6). It is noticed, however, some common concerns are shared by almost all the quality assessment of *Rigor*, such as the description, validation, and applicability of the methods, design or measures applied. When implementing the assessment of *Rigor*, the specific needs of the study quality may vary between SLRs.

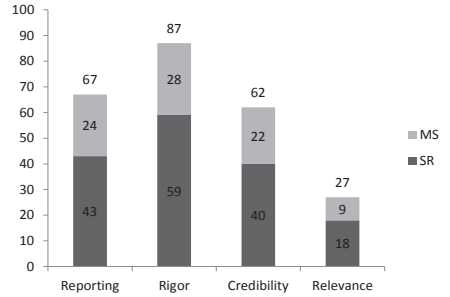

Figure 7: Combinations of SR and MS per aspect

Credibility. In the SLRs, the *credibility* of the study methods for ensuring that the findings are valid and meaningful was considered as important as the quality of *reporting*. Without credible methods and valid results, it could be hard for an SLR to draw a convincing conclusion. On average, many SLRs have about three criteria that addressed the concerns on the *credibility* of primary studies. Among the most used criteria for *Rigor*, the top criteria (shown in Table 6) were suggested by Dybå and Dingsøyr [5], and adopted by many reviewers.

Relevance. Only 25% SLRs assessed the *relevance* of primary studies. It is interesting that the minimum, the low-

er quartile, the median and the upper quartile are all the same number (Figure 6). This indicates most SLRs (Table 6) assessing *rigor* merely applied single criterion. The most frequently asked question: "*Is the study of value for research or practice?*"

We noticed that the criteria of quality assessment of eight SLRs ([S23, S87, S104, S39, S11, S14, S57, S26] reused or adapted directly from Dybå and Dingsøyr's checklist [5] (listed in Table 6). The eleven criteria in their checklist cover the four main aspects of study quality: three (Criteria 1-3) for *Reporting*, five (Criteria 4-8) for *Rigor*, two (Criteria 9-10) for *Credibility*, and the last one for *Relevance*. It forms a typical combination of the assessment criteria from the four aspects of quality concern.

Table 6: Most used criteria for each aspect

	Assessment criterion (question)	#
Reporting	Is there a clear statement (definition) of the aims (goals, purposes, problems, motivations, objectives, questions) of the research?	54
	Is there an adequate description of the context in which the research was carried out?	46
	Is the paper based on research?	25
	Are references maintained accurately?	6
	Does the study answer the research question defined or presents the results in a clear way?	5
	Is the reporting clear and coherent?	5
Rigor	Are the metrics (methods, design, measures) used in the study clearly (fully) defined (description)?	21
	Are the variables/metrics/methods/design used in the study adequately measured and validated (justified)?	19
	Was the data analysis (collected) sufficiently rigorous?	18
	Was there a control group with which to compare treatments?	16
	Are the data collection methods adequately described (defined)?	16
	Was the data collected in a way that addressed the research issue?	16
	Was the research design appropriate to address the aims of the research?	14
	Was the recruitment strategy appropriate to the aims of the research?	11
	Does the study provide description and justification of the data analysis approaches?	11
	Is the methodology (design) used suitable to address the stated research questions?	8
	Is the study design stated clearly?	5
	Are the metrics used in the study the most relevant ones for answering the research questions?	5
Credibility	Is there a clear statement of findings (data) and relate to the aims of research?	37
	Do the researchers discuss any problems (limitations, threats) with the validity (reliability) of their results?	23
	Has the relationship between researcher and participants been considered to an adequate degree?	12
	Is the study replicable?	7
	Has sufficient data been presented to support the findings?	5
	Are the findings credible?	4
Relevance	Is the study of value for research or practice?	15
	Are conclusions, implications for practice and future research, reported suitably for its audience?	6
	Has the approach been validated on a certain scale (either in academia or/ and industry)?	2

Furthermore, some SLRs have lots of criteria listed but only address a part of the four aspects, such as [S37] that have 31 criteria based on eight questions from [5]. Although these criteria assessed the quality of primary studies from many aspects and degrees (e.g., aims, context, design, control group, data collection, data analysis procedures, bias, conclusions), they mainly concentrate on the aspects of *Reporting*, *Rigor* and *Credibility* with less attention to *Relevance*. Another similar case is the study [S79] that includes 19 criteria without considering *Relevance*.

On the opposite side, some SLRs that have five or fewer criteria pay their attention on two or even one aspects only.

According to the guidelines [11], a mapping study (also referred as scoping study) is used to gain a broad review of the primary studies in a specific topic area that aims to identify what evidence is available. Thus, it is not necessary to assess a study's quality for mapping studies. As the results, many more SLRs (75) with quality assessment were found than SMS (35) in this review. Accordingly to our data extraction, it can be observed that the distribution patterns of the aspects that SLR studies and SMS studies mainly focus on are very similar. Figure 8 shows the percentage of SLR studies and SMS studies with the quality concern on each aspect.

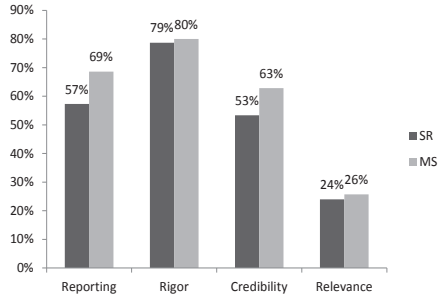

Figure 8: Comparison between SR and MS

In summary, ***Rigor is the most important aspect when reviewers perform the quality assessment of the primary studies, and Reporting and Credibility are also considered by most reviews.*** By considering the review types, ***the main quality aspects that systematic reviews and mapping studies focus on are quite similar*** (see the patterns shown in Figure 8. Like the overall SLR distribution over aspects, the most concerned aspects by both types of secondary studies are the *rigor* of research methods, followed by the quality of *reporting* and *credibility* of the results. Although both *Rigor* and *Relevance* are the emphasis of EBSE [15], ***Relevance received least attention of the researchers when conducting SLRs in SE.***

5.4 Other Issues

For most of the papers adapted the CASP, they did not make full use of the criteria listed in CASP. Along with the increasing number of SLRs in SE, more and more researchers are realizing the importance of quality assessment and include this part in their SLRs. Compared to the increasing number of SLRs, however, the growth rate of the SLRs containing quality assessment is a relatively small. Hence, it is important that SLRs researchers start paying more attention to the quality assessment of the papers included in SLRs. As mentioned in Section 4.1, fewer SMSs have the sections on quality assessment. Albeit quality assessment is not recommended for SMS, this part could be a good choice to increase the credibility of evidence in an SMS.

There also exist some gaps in the current state of art of quality assessment. When performing quality assessment, a relatively large number of researchers did not have explicit purposes and they just included this part as a section of SLR report. In this circumstance, the effort to carry out quality assessment is a kind of waste. Our review has also found

that there are not many SLRs whose quality assessment criteria pay sufficient attention to the *credibility* aspect in their quality assessment criteria.

As recommended in the assessment system of GRADE, factors that could decrease and increase the strength of evidence are both suggested to determine the strength of evidence. However, the current state of the scoring systems are all binary scale (1 or 0) or three-point scale (1, 0.5 or 0). They are all used to increase the confidence that researchers can place in certain target study and no question or criterion is used to deduct the quality score (e.g., -1). This situation might ignore the negative effect of the factors like inconsistency of results that could make to the strength of the evidence.

Since a large number of SLRs in SE is qualitative and observational in nature, the widely used checklists in SE consist of most qualitative criteria, which is quite different to the general guidelines and assessment systems recommended in other disciplines. The systems adapted in medicine and others are both quantitative and qualitative. The reason might be that compared to other disciplines like medicine, software engineering is a relatively young; and there are relatively less number of empirical studies that may result in less number of possibilities to perform quantitive assessment. This could be the reason why few SE researchers have adopted the quality assessment systems from other disciplines except CASP.

6. THREATS TO VALIDITY

This section discusses some of the potential threats to the validity of this study.

Search strategy. In order to exhaustively identify SLRs with quality assessment component and ensure that the process of papers selection was unbiased as far as possible, the approach of quasi-gold standard (QGS) [21] was adopted, which systematically integrates manual and automated search strategies and suggests a relatively rigorous approach for search performance evaluation in terms of sensitivity and precision. As we only searched four online digital libraries, we might have possibly missed some studies even using the QGS strategy. However, the used four digital libraries are not only by some of the biggest publishers in SE but also most frequently used for SLRs in SE. These digital libraries are believed to cover the majority of the high quality publications in SE. To capture as many SLRs as possible, however, we also used the snowballing as the complementary search to reduce the possibility of missing relevant SLRs. In addition, the search strategy was developed by the three research students and reviewed by their supervisor.

Selection criteria. The duplication of papers is a potential threat to frequency counts and to the statistics in this tertiary study. To ensure the removal of these duplicates, each retrieved paper was checked and further read by at least two researchers independently to handle the duplication and limited bias as much as possible. We also found it quite difficult to manage the duplication of empirical studies performed by the same author but which were reported as a part of other papers. We examined and discussed them carefully in order to ascertain whether or not they were the *duplicate* papers.

Data extraction. We extracted the data from the selected SLRs concerning the study type, the assessment purpose and the aspect of quality concern. To further ensure the correctness of the extracted data, the protocol was developed to define the data extraction strategy and format. The review protocol was proposed by the authors, and was then reviewed by their supervisor. We defined a data extraction form to obtain consistent extraction of relevant information and checked whether the data to be extracted would address the research questions. Moreover, the cross-check was necessary among the reviewers, and again we had at least two research students extracting data from each selected SLR independently. The supervisor played the role of expert panel who dealt with any divergences and disagreements between the students during the process.

7. CONCLUSIONS

The quality of the papers included in an SLR indicates the quality of the conclusions of that SLR [11]. This paper reports the first tertiary study aimed at systematically reviewing the state of the practice of quality assessment of the reported SLRs and SMSs on different topics of SE since 2004. Through this tertiary study, we have identified the existing quality assessment systems used by SLR researchers in SE. We have also identified and recommended the most often used guidelines and checklists and provide researchers the references and options when developing their own quality assessment. We have performed an in-depth analysis of the quality assessment criteria used by SLRs researchers to provide researchers with a set of common criteria that is expected to assist in forming appropriate quality assessment criteria for SLRs in SE.

We also thoroughly analyzed that the reported purposes of assessing the quality of the papers included in SLRs in SE. That analysis has provided a list of most commonly reported purposes of quality assessment along with a series of corresponding criteria for each purpose. This information can help researchers to develop the quality criteria according to the purpose of their quality assessment: rather than '*blindly*' performing the quality assessment for no particular reason.

Regarding the aspects of the quality assessment, we enumerated and explained the aspects so that researchers could gain a better understanding of the quality assessment from various aspects. We also enlist the most concerned criteria for each aspect that researchers could choose from for their own checklists. Based on our observation, we suggest that researchers should care more about the aspect of credibility, which directly impacts the validity and generalizability of a study's conclusion. In the meantime, relevance, as one important strength of EBSE [15], was an ignored aspect by many existing SLRs, but it implies the possible influence and value of a study to the research community and software industry. This needs more serious considerations in the quality assessment of any future SLRs.

The next step of this thread of research could be to identify the gaps between the actual practices of quality assessment in EBSE and the existing guidelines for quality assessment.

Based on the results, more concrete improvements can be suggested to further improve the research methodologies for evidence based software engineering.

8. REFERENCES

[1] J. Biolchini, P. G. Mian, A. C. C. Natali, and G. H. Travassos. Systematic review in software engineering. *System Engineering and Computer Science Department COPPE/UFRJ, Technical Report ES*, 679(05):45, 2005.

[2] D. S. Cruzes and T. Dybå. Research synthesis in software engineering: A tertiary study. *Information and Software Technology*, 53(5):440–455, 2011.

[3] F. Q. Da Silva, A. L. Santos, S. Soares, A. C. C. França, C. V. Monteiro, and F. F. Maciel. Six years of systematic literature reviews in software engineering: An updated tertiary study. *Information and Software Technology*, 53(9):899–913, 2011.

[4] F. Q. da Silva, A. L. Santos, S. C. Soares, A. C. C. França, and C. V. Monteiro. A critical appraisal of systematic reviews in software engineering from the perspective of the research questions asked in the reviews. In *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, page 33. ACM, 2010.

[5] T. Dybå and T. Dingsøyr. Strength of evidence in systematic reviews in software engineering. In *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement*, pages 178–187. ACM, 2008.

[6] T. Dyba, T. Dingsoyr, and G. K. Hanssen. Applying systematic reviews to diverse study types: An experience report. In *Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on*, pages 225–234. IEEE, 2007.

[7] T. Dyba, B. A. Kitchenham, and M. Jorgensen. Evidence-based software engineering for practitioners. *Software, IEEE*, 22(1):58–65, 2005.

[8] T. Greenhalgh and R. Taylor. *How to read a paper*. BMJ Publishing Group London, 2002.

[9] P. H. R. U. in Oxford. Critical appraisal skills programme. http://www.casp-uk.net/, 2013.

[10] A. R. Jadad, R. A. Moore, D. Carroll, C. Jenkinson, D. J. M. Reynolds, D. J. Gavaghan, and H. J. McQuay. Assessing the quality of reports of randomized clinical trials: is blinding necessary? *Controlled clinical trials*, 17(1):1–12, 1996.

[11] S. Keele. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, EBSE Technical Report EBSE-2007-01, 2007.

[12] B. Kitchenham. Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33:2004, 2004.

[13] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman. Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, 51(1):7–15, 2009.

[14] B. Kitchenham, R. Pretorius, D. Budgen, O. Pearl Brereton, M. Turner, M. Niazi, and S. Linkman. Systematic literature reviews in software engineering–a tertiary study. *Information and Software Technology*, 52(8):792–805, 2010.

[15] B. A. Kitchenham, T. Dyba, and M. Jorgensen. Evidence-based software engineering. In *Software Engineering, 2004. ICSE 2004. Proceedings. 26th International Conference on*, pages 273–281. IEEE, 2004.

[16] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, and J. Rosenberg. Preliminary guidelines for empirical research in software engineering. *Software Engineering, IEEE Transactions on*, 28(8):721–734, 2002.

[17] A. D. Oxman, G. W. Group, et al. Grading quality of evidence and strength of recommendations. *Bmj*, 328(19):1490–4, 2004.

[18] K. Petersen and N. B. Ali. Identifying strategies for study selection in systematic reviews and maps. *Empirical Software Engineering*, 2011.

[19] M. Riaz. Systematic review of software maintainability prediction and metrics. http://chao.stat.nthu.edu.tw/, 2011.

[20] D. I. Sjoberg, T. Dyba, and M. Jorgensen. The future of empirical methods in software engineering research. In *Future of Software Engineering, 2007. FOSE'07*, pages 358–378. IEEE, 2007.

[21] H. Zhang and M. Ali Babar. On searching relevant studies in software engineering. In *Proceedings of the 14th international conference on evaluation and assessment in software engineering (EASE)*, 2010.

[22] H. Zhang and M. Ali Babar. Systematic reviews in software engineering: An empirical investigation. *Information and Software Technology*, 55(7):1341–1354, 2013.

[23] H. Zhang, M. A. Babar, and P. Tell. Identifying relevant studies in software engineering. *Information and Software Technology*, 53(6):625–637, 2011.

APPENDIX

A. SELECTED SLR LIST

[S1] M. Abdellatief, A. B. M. Sultan, A. A. A. Ghani, and M. A. Jabar. A mapping study to investigate component-based software system metrics. *Journal of systems and software*, 86(3):587–603, 2013.

[S2] W. Afzal and R. Torkar. On the application of genetic programming for software engineering predictive modeling: A systematic review. *Expert Systems with Applications*, 38(9):11984–11997, 2011.

[S3] W. Afzal, R. Torkar, and R. Feldt. A systematic review of search-based testing for non-functional system properties. *Information and Software Technology*, 51(6):957–976, 2009.

[S4] M. O. Ahmad, J. Markkula, and M. Ovio. Kanban in software development: A systematic literature review. In *Software Engineering and Advanced Applications (SEAA), 2013 39th EUROMICRO Conference on*, pages 9–16. IEEE, 2013.

[S5] M. S. Ali, M. Ali Babar, L. Chen, and K.-J. Stol. A systematic review of comparative evidence of aspect-oriented programming. *Information and software Technology*, 52(9):871–887, 2010.

[S6] V. Alves, N. Niu, C. Alves, and G. Valença. Requirements engineering for software product lines: A systematic literature review. *Information and Software Technology*, 52(8):806–820, 2010.

[S7] N.-D. Anh, D. S. Cruzes, and R. Conradi. Dispersion, coordination and performance in global software teams: A systematic review. In *Proceedings of the ACM-IEEE international symposium on Empirical software engineering and measurement*, pages 129–138. ACM, 2012.

[S8] M. A. P. Araújo, V. F. Monteiro, and G. H. Travassos. Towards a model to support in silico studies of software evolution. In *Empirical Software Engineering and Measurement (ESEM),*

2012 ACM-IEEE International Symposium on, pages 281–289. IEEE, 2012.

[S9] T. B. C. Arias, P. van der Spek, and P. Avgeriou. A practice-driven systematic review of dependency analysis solutions. *Empirical Software Engineering*, 16(5):544–586, 2011.

[S10] M. T. Baldassarre, D. Caivano, B. Kitchenham, and G. Visaggio. Systematic review of statistical process control: An experience report. In *11th Evaluation and Assessment in Software Engineering Conference, BCS UK*, pages 94–102, 2007.

[S11] A. Bandi, B. J. Williams, and E. B. Allen. Empirical evidence of code decay: A systematic mapping study. In *Reverse Engineering (WCRE), 2013 20th Working Conference on*, pages 341–350. IEEE, 2013.

[S12] M. Bano, S. Imtiaz, N. Ikram, M. Niazi, and M. Usman. Causes of requirement change-a systematic literature review. In *Evaluation & Assessment in Software Engineering (EASE 2012), 16th International Conference on*, pages 22–31. IET, 2012.

[S13] M. Bano and D. Zowghi. Users' involvement in requirements engineering and system success. In *Empirical Requirements Engineering (EmpiRE), 2013 IEEE Third International Workshop on*, pages 24–31. IEEE, 2013.

[S14] D. Basten, O. Pankratz, and D. Joosten. Assessing the assessors-an overview and evaluation of it project success reports. In *Proceedings of the 21st European Conference on Information Systems*, 2013.

[S15] S. Bayona, J. A. Calvo-Manzano, and T. San Feliu. Critical success factors in software process improvement: A systematic review. In *Software Process Improvement and Capability Determination*, pages 1–12. Springer, 2012.

[S16] S. Bayona, J. A. Calvo-Manzano, and T. San Feliu. Review of critical success factors related to people in software process improvement. In *Systems, Software and Services Process Improvement*, pages 179–189. Springer, 2013.

[S17] S. Beecham, N. Baddoo, T. Hall, H. Robinson, and H. Sharp. Motivation in software engineering: A systematic literature review. *Information and Software Technology*, 50(9):860–878, 2008.

[S18] V. Boucharas, M. van Steenbergen, S. Jansen, and S. Brinkkemper. The contribution of enterprise architecture to the achievement of organizational goals: a review of the evidence. In *Trends in Enterprise Architecture Research*, pages 1–15. Springer, 2010.

[S19] H. P. Breivold and I. Crnkovic. A systematic review on architecting for software evolvability. In *Software Engineering Conference (ASWEC), 2010 21st Australian*, pages 13–22. IEEE, 2010.

[S20] H. P. Breivold, I. Crnkovic, and M. Larsson. A systematic review of software architecture evolution research. *Information and Software Technology*, 54(1):16–40, 2012.

[S21] C. Catal. On the application of genetic algorithms for test case prioritization: a systematic literature review. In *Proceedings of the 2nd international workshop on Evidential assessment of software technologies*, pages 9–14. ACM, 2012.

[S22] P. Cedillo, A. Fernandez, E. Insfran, and S. Abrahão. Quality of web mashups: A systematic mapping study. In *Current Trends in Web Engineering*, pages 66–78. Springer, 2013.

[S23] L. Chen and M. Ali Babar. A systematic review of evaluation of variability management approaches in software product lines. *Information and Software Technology*, 53(4):344–362, 2011.

[S24] P. A. da Mota Silveira Neto, I. d. Carmo Machado, J. D. McGregor, E. S. De Almeida, and S. R. de Lemos Meira. A systematic mapping study of software product lines testing. *Information and Software Technology*, 53(5):407–423, 2011.

[S25] E. A. N. da Silva and D. Lucrédio. Software engineering for the cloud: a research roadmap. In *Software Engineering (SBES), 2012 26th Brazilian Symposium on*, pages 71–80. IEEE, 2012.

[S26] F. Q. da Silva, A. C. C. França, M. Suassuna, L. M. de Sousa Mariz, I. Rossiley, R. C. de Miranda, T. B. Gouveia, C. V. Monteiro, E. Lucena, E. S. Cardozo, et al. Team building criteria in software projects: A mix-method replicated study. *Information and Software Technology*, 55(7):1316–1340, 2013.

[S27] J. R. F. da Silva, F. A. P. da Silva, L. M. do Nascimento, D. A. Martins, and V. C. Garcia. The dynamic aspects of product derivation in dspl: A systematic literature review. In *Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on*, pages 466–473. IEEE, 2013.

[S28] V. R. L. de Mendonca, C. L. Rodrigues, F. A. A. d. Soares, and A. M. R. Vincenzi. Static analysis techniques and tools: A systematic mapping study. In *ICSEA 2013, The Eighth International Conference on Software Engineering Advances*, pages 72–78, 2013.

[S29] J. Díaz, J. Pérez, P. P. Alarcón, and J. Garbajosa. Agile product line engineering?a systematic literature review. *Software: Practice and experience*, 41(8):921–941, 2011.

[S30] O. Dieste and N. Juristo. Systematic review and aggregation of empirical studies on elicitation techniques. *Software Engineering, IEEE Transactions on*, 37(2):283–304, 2011.

[S31] T. Dybå and T. Dingsøyr. Empirical studies of agile software development: A systematic review. *Information and software technology*, 50(9):833–859, 2008.

[S32] H. Edison, N. Bin Ali, and R. Torkar. Towards innovation measurement in the software industry. *Journal of Systems and Software*, 86(5):1390–1407, 2013.

[S33] M. El-Attar and J. Miller. Constructing high quality use case models: a systematic review of current practices. *Requirements Engineering*, 17(3):187–201, 2012.

[S34] E. Engström, M. Skoglund, and P. Runeson. Empirical evaluations of regression test selection techniques: a systematic review. In *Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement*, pages 22–31. ACM, 2008.

[S35] K. R. Felizardo, S. G. MacDonell, E. Mendes, and J. C. Maldonado. A systematic mapping on the use of visual data mining to support the conduct of systematic literature reviews. *Journal of Software*, 7(2):450–461, 2012.

[S36] A. Fernandez, E. Insfran, and S. Abrahão. Usability evaluation methods for the web: A systematic mapping study. *Information and Software Technology*, 53(8):789–817, 2011.

[S37] A. M. Fernández-Sáez, M. Genero, and M. R. Chaudron. Empirical studies concerning the maintenance of uml diagrams and their use in the maintenance of code: A systematic mapping study. *Information and Software Technology*, 55(7):1119–1142, 2013.

[S38] L. S. González, F. G. Rubio, F. R. González, and M. P. Velthuis. Measurement in business processes: a systematic review. *Business Process Management Journal*, 16(1):114–134, 2010.

[S39] W. Gunathilake and T. Neligwa. A quality assessment framework for kms software: Reflections on conducting a systematic literature review. *KIM2013*, 4:13, 2013.

[S40] C. Gutierrez, J. Garbajosa, J. Diaz, and A. Yague. Providing a consensus definition for the term" smart product". In *Engineering of Computer Based Systems (ECBS), 2013 20th IEEE International Conference and Workshops on the*, pages 203–211. IEEE, 2013.

[S41] Ø. Hauge, C. Ayala, and R. Conradi. Adoption of open source software in software-intensive organizations–a systematic literature review. *Information and Software Technology*, 52(11):1133–1154, 2010.

[S42] S. Heckman and L. Williams. A systematic literature review of actionable alert identification techniques for automated static code analysis. *Information and Software Technology*, 53(4):363–387, 2011.

[S43] G. Holl, P. Grünbacher, and R. Rabiser. A systematic review and an expert survey on capabilities supporting multi product lines. *Information and Software Technology*, 54(8):828–852, 2012.

[S44] E. Hossain, M. A. Babar, and H.-y. Paik. Using scrum in global software development: a systematic literature review. In *Global Software Engineering, 2009. ICGSE 2009. Fourth IEEE International Conference on*, pages 175–184. IEEE, 2009.

[S45] P. Jamshidi, A. Ahmad, and C. Pahl. Cloud migration research: a systematic review. *Cloud computing, IEEE Transactions on*, 1(2):142–157, 2013.

[S46] S. U. Khan, M. Niazi, and R. Ahmad. Critical success factors for offshore software development outsourcing vendors: A systematic literature review. In *Global Software Engineering, 2009. ICGSE 2009. Fourth IEEE International Conference On*, pages 207–216. IEEE, 2009.

[S47] S. U. Khan, M. Niazi, and R. Ahmad. Barriers in the selection of offshore software development outsourcing vendors: An exploratory study using a systematic literature review. *Information and Software Technology*, 53(7):693–706, 2011.

[S48] B. Kitchenham, P. Brereton, M. Turner, M. Niazi, S. Linkman, R. Pretorius, and D. Budgen. The impact of limited search procedures for systematic literature reviews?a participant-observer case study. In *Empirical Software Engineering and Measurement, 2009. ESEM 2009. 3rd*

International Symposium on, pages 336–345. IEEE, 2009.

[S49] B. A. Kitchenham, E. Mendes, and G. H. Travassos. Cross versus within-company cost estimation studies: A systematic review. *Software Engineering, IEEE Transactions on*, 33(5):316–329, 2007.

[S50] H. Koziolek. Sustainability evaluation of software architectures: a systematic review. In *Proceedings of the joint ACM SIGSOFT conference–QoSA and ACM SIGSOFT symposium–ISARCS on Quality of software architectures–QoSA and architecting critical systems–ISARCS*, pages 3–12. ACM, 2011.

[S51] M. Lavallée and P. N. Robillard. The impacts of software process improvement on developers: a systematic review. In *Proceedings of the 2012 International Conference on Software Engineering*, pages 113–122. IEEE Press, 2012.

[S52] J. Li, H. Zhang, L. Zhu, R. Jeffery, Q. Wang, and M. Li. Preliminary results of a systematic review on requirements evolution. In *Evaluation & Assessment in Software Engineering (EASE 2012), 16th International Conference on*. IET, 2012.

[S53] Z. Li, P. Liang, and P. Avgeriou. Application of knowledge-based approaches in software architecture: A systematic mapping study. *Information and Software Technology*, 55(5):777–794, 2013.

[S54] Z. Li, H. Zhang, L. O?Brien, R. Cai, and S. Flint. On evaluating commercial cloud services: A systematic review. *Journal of Systems and Software*, 86(9):2371–2393, 2013.

[S55] D. Liu, Q. Wang, and J. Xiao. The role of software process simulation modeling in software risk management: A systematic review. In *Empirical Software Engineering and Measurement, 2009. ESEM 2009. 3rd International Symposium on*, pages 302–311. IEEE, 2009.

[S56] S. Mahdavi-Hezavehi, M. Galster, and P. Avgeriou. Variability in quality attributes of service-based software systems: A systematic literature review. *Information and Software Technology*, 55(2):320–343, 2013.

[S57] L. Major, T. Kyriacou, and O. P. Brereton. Systematic literature review: teaching novices programming using robots. *IET software*, 6(6):502–513, 2012.

[S58] T. Martínez-Ruiz, J. Münch, F. García, and M. Piattini. Requirements and constructors for tailoring software processes: a systematic literature review. *Software Quality Journal*, 20(1):229–260, 2012.

[S59] E. Mendes. A systematic review of web engineering research. In *Empirical Software Engineering, 2005. 2005 International Symposium on*, pages 10–pp. IEEE, 2005.

[S60] M. Misbhauddin and M. Alshayeb. Uml model refactoring: a systematic literature review. *Empirical Software Engineering*, pages 1–46, 2013.

[S61] S. Montagud, S. Abrahão, and E. Insfran. A systematic review of quality attributes and measures for software product lines. *Software Quality Journal*, 20(3-4):425–486, 2012.

[S62] S. Nair, J. L. de la Vara, M. Sabetzadeh, and L. Briand. Classification, structuring, and assessment of evidence for safety–a systematic literature review. In *Software Testing, Verification and Validation (ICST), 2013 IEEE Sixth International Conference on*, pages 94–103. IEEE, 2013.

[S63] M. Niazi, N. Ikram, M. Bano, S. Imtiaz, and S. U. Khan. Establishing trust in offshore software outsourcing relationships: an exploratory study using a systematic literature review. *IET software*, 7(5):283–293, 2013.

[S64] J. Nicolás and A. Toval. On the generation of requirements specifications from software engineering models: A systematic literature review. *Information and Software Technology*, 51(9):1291–1307, 2009.

[S65] S. Nidhra and M. Yanamadala. Knowledge transfer challenges and mitigation strategies in global software development. *International Journal of Information Management*, 11(4):333–355, 2012.

[S66] K. Oliveira, J. Pimentel, E. Santos, D. Dermeval, G. Guedes, C. Souza, M. Soares, J. Castro, F. Alencar, and C. Silva. 25 years of requirements engineering in brazil: a systematic mapping. In *Proc. of the 16th Requirements Engineering Workshop. Montevideo, Uruguay. Abr*, 2013.

[S67] L. B. R. Oliveira, M. Guessi, D. Feitosa, C. Manteuffel, M. Galster, F. Oquendo, and E. Y. Nakagawa. An investigation on quality models and quality attributes for embedded systems. In *ICSEA 2013, The Eighth International Conference on Software Engineering Advances*, pages 523–528, 2013.

[S68] S. Ouhbi, A. Idri, J. L. Fernández-Alemán, and A. Toval. Requirements engineering education: a systematic mapping study. *Requirements Engineering*, pages 1–20, 2013.

[S69] C. Pacheco and I. Garcia. A systematic literature review of stakeholder identification methods in requirements elicitation. *Journal of Systems and Software*, 85(9):2171–2181, 2012.

[S70] T. Patikirikorala, A. Colman, J. Han, and L. Wang. A systematic survey on the design of self-adaptive software systems using control engineering approaches. In *Software Engineering for Adaptive and Self-Managing Systems (SEAMS), 2012 ICSE Workshop on*, pages 33–42. IEEE, 2012.

[S71] T. K. Paul and M. F. Lau. Redefinition of fault classes in logic expressions. In *Quality Software (QSIC), 2012 12th International Conference on*, pages 144–153. IEEE, 2012.

[S72] B. Pérez, Á. L. Rubio, M. Zapata, et al. A systematic review of code generation proposals from state machine specifications. *Information and Software Technology*, 54(10):1045–1066, 2012.

[S73] F. J. Pino, F. García, and M. Piattini. Software process improvement in small and medium software enterprises: a systematic review. *Software Quality Journal*, 16(2):237–261, 2008.

[S74] A. M. Pitangueira, R. S. P. Maciel, M. de Oliveira Barros, and A. S. Andrade. A systematic review of software requirements selection and prioritization using sbse approaches. In *Search Based Software Engineering*, pages 188–208. Springer, 2013.

[S75] N. Qureshi, M. Usman, and N. Ikram. Evidence in software architecture, a systematic literature review. In *Proceedings of the 17th International Conference on Evaluation and Assessment in Software Engineering*, pages 97–106. ACM, 2013.

[S76] D. Radjenović, M. Heričko, R. Torkar, and A. Živkovič. Software fault prediction metrics: A systematic literature review. *Information and Software Technology*, 55(8):1397–1418, 2013.

[S77] D. Rattan, R. Bhatia, and M. Singh. Software clone detection: A systematic review. *Information and Software Technology*, 55(7):1165–1199, 2013.

[S78] D. S. Reis, R. O. Prates, et al. Applicability of the semiotic inspection method: a systematic literature review. In *Proceedings of the 10th Brazilian Symposium on on Human Factors in Computing Systems and the 5th Latin American Conference on Human-Computer Interaction*, pages 177–186. Brazilian Computer Society, 2011.

[S79] M. Riaz. Maintainability prediction of relational database-driven applications: a systematic review. In *Evaluation & Assessment in Software Engineering (EASE 2012), 16th International Conference on*. IET, 2012.

[S80] M. Riaz, E. Mendes, and E. Tempero. A systematic review of software maintainability prediction and metrics. In *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*, pages 367–377. IEEE Computer Society, 2009.

[S81] K. Riņkevičs and R. Torkar. Equality in cumulative voting: A systematic review with an improvement proposal. *Information and Software Technology*, 55(2):267–287, 2013.

[S82] S. K. Saha, M. Selvi, G. Buyukcan, and M. Mohymen. A systematic review on creativity techniques for requirements engineering. In *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*, pages 34–39. IEEE, 2012.

[S83] N. Salleh, E. Mendes, and J. Grundy. Empirical studies of pair programming for cs/se teaching in higher education: A systematic literature review. *Software Engineering, IEEE Transactions on*, 37(4):509–525, 2011.

[S84] I. Santiago, Á. Jiménez, J. M. Vara, V. De Castro, V. A. Bollati, and E. Marcos. Model-driven engineering as a new landscape for traceability management: A systematic literature review. *Information and Software Technology*, 54(12):1340–1356, 2012.

[S85] A. C. Santos, M. E. Delamaro, and F. L. Nunes. The relationship between requirements engineering and virtual reality systems: A systematic literature review. In *Virtual and Augmented Reality (SVR), 2013 XV Symposium on*, pages 53–62. IEEE, 2013.

[S86] R. d. Santos Rocha and M. Fantinato. The use of software product lines for business process management: A systematic literature review. *Information and Software Technology*, 55(8):1355–1373, 2013.

[S87] P. Sfetsos and I. Stamelos. Empirical studies on quality in agile practices: A systematic literature review. In *Quality of Information and Communications Technology (QUATIC), 2010 Seventh International Conference on the*, pages 44–53. IEEE, 2010.

[S88] A. Shahrokni and R. Feldt. A systematic review of software robustness. *Information and Software Technology*, 55(1):1–17, 2013.

[S89] S. P. Shashank, P. Chakka, and D. V. Kumar. A systematic literature survey of integration testing in component-based software engineering. In *Computer and Communication Technology (ICCCT), 2010 International Conference on*, pages 562–568. IEEE, 2010.

[S90] M. Shen, W. Yang, G. Rong, and D. Shao. Applying agile methods to embedded software development: A systematic review. In *Software Engineering for Embedded Systems (SEES), 2012 2nd International Workshop on*, pages 30–36. IEEE, 2012.

[S91] B. M. Shuaibu, N. M. Norwawi, M. H. Selamat, and A. Al-Alwani. Systematic review of web application security development model. *Artificial Intelligence Review*, pages 1–18, 2013.

[S92] M. Staples and M. Niazi. Systematic review of organizational motivations for adopting cmm-based spi. *Information and software technology*, 50(7):605–620, 2008.

[S93] K.-J. Stol and M. A. Babar. Reporting empirical research in open source software: the state of practice. In *Open Source Ecosystems: Diverse Communities Interacting*, pages 156–169. Springer, 2009.

[S94] M. Sulayman and E. Mendes. A systematic literature review of software process improvement in small and medium web companies. In *Advances in software engineering*, pages 1–8. Springer, 2009.

[S95] M. Svahnberg, T. Gorschek, R. Feldt, R. Torkar, S. B. Saleem, and M. U. Shafique. A systematic review on strategic release planning models. *Information and Software Technology*, 52(3):237–248, 2010.

[S96] M. Svahnberg, T. Gorschek, T. T. L. Nguyen, and M. Nguyen. Uni-repm: a framework for requirements engineering process assessment. *Requirements Engineering*, pages 1–28, 2013.

[S97] R. B. Svensson, M. Host, and B. Regnell. Managing quality requirements: A systematic review. In *Software Engineering and Advanced Applications (SEAA), 2010 36th EUROMICRO Conference on*, pages 261–268. IEEE, 2010.

[S98] T. Tahir and A. Jafar. A systematic review on software measurement programs. In *Frontiers of Information Technology (FIT), 2011*, pages 39–44. IEEE, 2011.

[S99] F. Tamy Ishii, G. C. L Leal, E. V. Cardoza Galdamez, R. Balancieri, T. F. Calvi Tait, and E. H. Moriya Huzita. Knowledge management in distributed software development: a systematic review. In *XVIII Congreso Argentino de Ciencias de la Computación*, 2013.

[S100] A. Y. Teka, N. Condori-Fernandez, and B. Sapkota. A systematic literature review on service description methods. In *Requirements Engineering: Foundation for Software Quality*, pages 239–255. Springer, 2012.

[S101] M. Turner, R. Kaur, and P. Brereton. A lightweight systematic literature review of studies about the use of pair programming to teach introductory programming. In *The 20th Annual Psychology of Programming Interest Group Conference, Lancaster University, UK*, pages 205–219, 2008.

[S102] M. Turner, B. Kitchenham, P. Brereton, S. Charters, and D. Budgen. Does the technology acceptance model predict actual use? a systematic literature review. *Information and Software Technology*, 52(5):463–479, 2010.

[S103] J. Vanhanen and M. V. MÄNTYLÄ. A systematic mapping study of empirical studies on the use of pair programming in industry. *International Journal of Software Engineering and Knowledge Engineering*, 23(09):1221–1267, 2013.

[S104] R. L. Vivian, E. H. M. Huzita, G. C. L. Leal, and A. P. C. Steinmacher. Context-awareness on software artifacts in distributed software development: a systematic review. In *Collaboration and Technology*, pages 30–44. Springer, 2011.

[S105] G. S. Walia and J. C. Carver. A systematic literature review to identify and classify software requirement errors. *Information and Software Technology*, 51(7):1087–1109, 2009.

[S106] J. Wen, S. Li, Z. Lin, Y. Hu, and C. Huang. Systematic literature review of machine learning based software development effort estimation models. *Information and Software Technology*, 54(1):41–59, 2012.

[S107] D. Weyns and T. Ahmad. Claims and evidence for architecture-based self-adaptation: a systematic literature review. In *Software Architecture*, pages 249–265. Springer, 2013.

[S108] D. Weyns, M. U. Iftikhar, S. Malek, and J. Andersson. Claims and supporting evidence for self-adaptive systems: A literature study. In *Software Engineering for Adaptive and Self-Managing Systems (SEAMS), 2012 ICSE Workshop on*, pages 89–98. IEEE, 2012.

[S109] B. J. Williams and J. C. Carver. Characterizing software architecture changes: A systematic review. *Information and Software Technology*, 52(1):31–51, 2010.

[S110] H. Zhang, B. Kitchenham, and D. Pfahl. Reflections on 10 years of software process simulation modeling: a systematic review. In *Making globally distributed software development a success story*, pages 345–356. Springer, 2008.

B. OTHER SLR LIST

[SP1] A. Ampatzoglou, S. Charalampidou, and I. Stamelos. Research state of the art on gof design patterns: A mapping study. *Journal of Systems and Software*, 86(7):1945–1964, 2013.

[SP2] A. Ampatzoglou and I. Stamelos. Software engineering research for computer games: A systematic review. *Information and Software Technology*, 52(9):888–901, 2010.

[SP3] N. M. N. Daud and W. W. Kadir. Systematic mapping study of quality attributes measurement in service oriented architecture. In *Information Science and Digital Content Technology (ICIDT), 2012 8th International Conference on*, volume 3, pages 626–631. IEEE, 2012.

[SP4] U. Eklund and J. Bosch. Archetypical approaches of fast software development and slow embedded projects. In *Software Engineering and Advanced Applications (SEAA), 2013 39th EUROMICRO Conference on*, pages 276–283. IEEE, 2013.

[SP5] E. Engström, P. Runeson, and M. Skoglund. A systematic review on regression test selection techniques. *Information and Software Technology*, 52(1):14–30, 2010.

[SP6] T. Hall, S. Beecham, D. Bowes, D. Gray, and S. Counsell. A systematic literature review on fault prediction performance in software engineering. *Software Engineering, IEEE Transactions on*, 38(6):1276–1304, 2012.

[SP7] G. K. Hanssen, F. O. Bjørnson, and H. Westerheim. Tailoring and introduction of the rational unified process. In *Software Process Improvement*, pages 7–18. Springer, 2007.

[SP8] L. B. Lisboa, V. C. Garcia, D. Lucrédio, E. S. de Almeida, S. R. de Lemos Meira, and R. P. de Mattos Fortes. A systematic review of domain analysis tools. *Information and Software Technology*, 52(1):1–13, 2010.

[SP9] A. M. Magdaleno, C. M. L. Werner, and R. M. d. Araujo. Reconciling software development models: a quasi-systematic review. *Journal of Systems and Software*, 85(2):351–369, 2012.

[SP10] J. Maras, L. Lednicki, and I. Crnkovic. 15 years of cbse symposium: impact on the research community. In *Proceedings of the 15th ACM SIGSOFT symposium on Component Based Software Engineering*, pages 61–70. ACM, 2012.

[SP11] B. Mohabbati, M. Asadi, D. Gašević, M. Hatala, and H. A. Müller. Combining service-orientation and software product line engineering: A systematic mapping study. *Information and Software Technology*, 55(11):1845–1859, 2013.

[SP12] M. J. Monasor, A. Vizcaino, M. Piattini, and I. Caballero. Preparing students and engineers for global software development: a systematic review. In *Global Software Engineering (ICGSE), 2010 5th IEEE International Conference on*, pages 177–186. IEEE, 2010.

[SP13] J. Noll, S. Beecham, and I. Richardson. Global software development and collaboration: barriers and solutions. *ACM Inroads*, 1(3):66–78, 2010.

[SP14] J. Pernstål, R. Feldt, and T. Gorschek. The lean gap: A review of lean approaches to large-scale software systems development. *Journal of Systems and Software*, 86(11):2797–2821, 2013.

[SP15] M. Sulayman and E. Mendes. An extended systematic review of software process improvement in small and medium web companies. In *Evaluation & Assessment in Software Engineering (EASE 2011), 15th Annual Conference on*, pages 134–143. IET, 2011.

[SP16] M. Unterkalmsteiner, T. Gorschek, A. M. Islam, C. K. Cheng, R. B. Permadi, and R. Feldt. Evaluation and measurement of software process improvement?a systematic literature review. *Software Engineering, IEEE Transactions on*, 38(2):398–424, 2012.

[SP17] D. Wahyudin, R. Ramler, and S. Biffl. A framework for defect prediction in specific software project contexts. In *Software Engineering Techniques*, pages 261–274. Springer, 2011.