# The need for multivocal literature reviews in software engineering: complementing systematic literature reviews with grey literature

Vahid Garousi
Software Engineering Research Group
Department of Computer Engineering
Hacettepe University, Ankara, Turkey
vahid.garousi@hacettepe.edu.tr

Michael Felderer
Quality Engineering Research Group
Institute of Computer Science
University of Innsbruck, Austria
michael.felderer@uibk.ac.at

Mika V. Mäntylä
M-Group, Faculty of Information Technology
and Electrical Engineering
University of Oulu, Oulu, Finland
mika.mantyla@oulu.fi

## ABSTRACT

Systematic Literature Reviews (SLR) may not provide insight into the "state of the practice" in SE, as they do not typically include the "grey" (non-published) literature. A Multivocal Literature Review (MLR) is a form of a SLR which includes grey literature in addition to the published (formal) literature. Only a few MLRs have been published in SE so far. We aim at raising the awareness for MLRs in SE by addressing two research questions (RQs): (1) What types of knowledge are missed when a SLR does not include the multivocal literature in a SE field? and (2) What do we, as a community, gain when we include the multivocal literature and conduct MLRs? To answer these RQs, we sample a few example SLRs and MLRs and identify the missing and the gained knowledge due to excluding or including the grey literature. We find that (1) grey literature can give substantial benefits in certain areas of SE, and that (2) the inclusion of grey literature brings forward certain challenges as evidence in them is often experience and opinion based. Given these conflicting viewpoints, the authors are planning to prepare systematic guidelines for performing MLRs in SE.

## CCS Concepts

General and reference → Document types → Surveys and overviews. General and reference → Cross-computing tools and techniques → Empirical studies.

## Keywords

Multivocal Literature Reviews; MLR; Systematic literature reviews; SLR; grey literature; research methodology; empirical software engineering.

## 1. INTRODUCTION

Systematic Literature Reviews (SLR) and Systematic Mapping (SM) studies have become common in software engineering (SE) to systematically collect evidence and to structure a given research area, respectively. Many such SLRs and SMs appear regularly in various SE venues (journals and conferences) each year [1, 2].

While SLR or a SM studies are valuable, other SE researchers have recently reported that "*the results of a SLR or a SM study could provide an established body of knowledge, focusing only on research contributions*" [3]. Since those studies do not include the "grey" literature (non-published, nor peer-reviewed sources of information), which are constantly produced by SE practitioners in a great scale, those studies do not provide insight into the "state of the practice" in SE. For a practical (practitioner-oriented) field such as SE, synthesizing and combing both the state-of-the art and –practice is very important. Unfortunately, it is a reality that a large majority of software practitioners do not publish in academic forums [4], and this means that the voice of the practitioners would be limited in review studies if we do not consider grey literature in addition to academic literature in those studies.

SLRs which include both the academic (formal) and the grey literature were termed as *Multivocal Literature Reviews (MLR)* in other fields, e.g., education, e.g., [5, 6], in the early 1990's. The main difference between a MLR and a SLR or a SM is the fact that, while SLRs and SMs use as input only academic peer-reviewed articles, MLRs in addition also use sources from the grey literature, e.g., blogs, white papers and web-pages [3]. Furthermore, for fields "*characterized by an abundance of diverse documents and a scarcity of systematic investigations*" [7], multivocal synthesis is highly recommended as an appropriate tool for investigations. Researchers also have reported that: *"another potential use of multivocal literature reviews is in closing the gap between academic research and professional practice*" [8].

We thus believe that, in a practical field such as SE, MLRs should be conducted in addition to SLRs. However, only a few MLRs have been published in SE so far. To address the issue, this paper aims at raising the need (awareness) for (more) MLRs in SE.

The remainder of this paper is structured as follows. A review of the related work is presented in Section 2. We describe the study goal and research methodology in Section 3. Section 4 presents the results (answering the study's two RQs). Section 5 summarizes the findings. Finally, in Section 6, we draw conclusions, and suggest areas for further research.

## 2. BACKGROUND AND RELATED WORK

### 2.1 MLRs in other fields

MLRs have become popular in several other fields, e.g., education, e.g., [5, 6]. For example, a 1991 paper [5] in the area of education research proposed an approach based on exploratory case studies to conduct rigorous MLRs.

While the notions of "MLR" and "multivocal" have been used in the research community, still many sources use the "grey" literature terminology and whether/how to include them in SLRs, e.g., [9-11]. For example, [9] discusses the advantages and challenges of including grey literature in state-of-the-evidence


*EASE'16*, June 01-03, 2016, Limerick, Ireland
 ISBN 978-1-4503-3691-8/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2915970.2916008

reviews, in the context of evidence-based nursing. [10] discusses the challenges and benefits of including for grey literature in SLRs.

Hopewell et al. [12] conducted a review of five studies, in the area of evidence-based medicine, comparing the effect of the inclusion or exclusion of 'grey' literature in meta-analyses of randomized medical trials.

The issue of the grey literature has become such important that there is even an International Journal on the topic of Grey Literature (www.emeraldinsight.com/toc/ijgl/1/4).

## 2.2 MLRs in SE

The 'multivocal' terminology has only been recently started to appear in the SLRs in SE , i.e., since 2013 in [13]. We found only three SLRs in SE which explicitly used the 'multivocal' terminology: [3, 13, 14]. [3] is a 2015 MLR on the financial aspect of managing technical debt. [13] is a 2013 MLR on technical debt. [14] is a 2015 MLR on iOS applications testing.

Many other SLRs have also included the grey literature in their reviews and have not used the 'multivocal' terminology, e.g., [15]. A 2012 MSc thesis entitled "*On the quality of grey literature and its use in information synthesis during systematic literature reviews*" [16] explored the state of including the grey literature in the SE SLRs. Two of the RQs in that study were: (1) What is the extent of usage of grey literature in SE SLRs? and (2) How can we assess the quality of grey literature? The study found that the ratio of grey evidence in the SE SLRs were only about 9%, and the grey literature included concentrated mostly in recent past (~48% between years 2007-2012).

# 3. GOAL, RESEARCH QUESTIONS AND METHODOLOGY

## 3.1 Goal and research questions

The goal of this study is to raise the need for (more) MLRs in SE. Based on the above goal, we raise the following two research questions (RQs):

- **RQ 1-** What types of knowledge (opportunities) are missed when a SLR does not include the multivocal literature in a SE field? To answer this RQ, we will select a few representative SLRs already published, and identify the missing knowledge due to not including the multivocal literature.
- **RQ 2-** What do we, as the SE community, gain when we include the multivocal literature in review studies and conduct MLRs? To answer this RQ, we will select a few representative MLRs already published, and identify the knowledge gain by including the multivocal literature.

We discuss next the research methodology that we developed to answer each of the two RQs.

## 3.2 Research methodology

### *3.2.1 Methodology for RQ1*

To address RQ 1, our methodology was to first select a small subset of SE SLRs, which have not included the grey literature, and then search for grey literature in their focus areas, to find the types of missing information available via the grey literature sources that were not included in those SLRs. Since grey literature in SE is usually in forms of online article, blog post and even video talks, we planned to conduct the searches using the Google and YouTube search engines and major forums where practitioners post questions and discuss technical issues, e.g., Stack Overflow (www.StackOverflow.com). To keep our efforts manageable, from the pool of all SE SLRs, we sampled (chose) three SLRs to answer RQ1, as shown in Table 1. Two of these SLRs [17, 18] are studies that the authors have co-authored in recent years.

**Table 1- The three SLRs sampled (chosen) to answer RQ1**

| Ref. | Year | Topic |
|---|---|---|
| [17] | 2013 | An SM on Graphical User Interface (GUI) testing |
| [18] | 2015 | An SLR on using metrics in agile and lean software development |
| [19] | 2015 | An SLR on definitions, precedents and outcomes of technical debt |

### *3.2.2 Methodology for RQ2*

Similar to RQ1, from the small set of MLRs in SE, we sampled three MLRs to answer RQ2, as shown in Table 2. We also show the number of literature entries (formal versus grey) and the ratio of the grey sources for each MLR.

Again, two of these MLRs are studies that the authors have been involved in and are under peer review as of this writing: (1) a MLR on deciding when and what to automate in testing (ManAutoTest), and (2) a MLR on test maturity and test process improvement (TM/TPI).

**Table 2- The three MLRs sampled to answer RQ2**

| Ref. | Year | Topic | Num. of lit. entries F, G* (% of grey) |
|---|---|---|---|
| Under review | 2015 | A MLR on deciding when and what to automate in testing (ManAutoTest) | 26, 52 (66%) |
| Under review | 2015 | A MLR on test maturity and test process improvement (TM/TPI) | 130, 51 (28%) |
| [13] | 2013 | A MLR on technical debt (TechDebt) | 0, 35 (100%) |

* F: Formal, G: Grey

It is interesting to observe that, for the ManAutoTest MLR, about 66% (more than half) of the study pool were from the grey literature. To answer RQ 2 (assessing what the community would gain when we include multivocal literature), we reviewed, for each MLR, the types of contributions and evidence that were utilized from the grey literature to answer the RQs of each MLR.

# 4. RESULTS

## 4.1 RQ1: Knowledge missed when a SLR does not include multivocal literature

### *4.1.1 SM on GUI testing [17]*

Following the methodology for RQ1, stated in Section 3.2, we conducted Google and YouTube searches for 'GUI testing', 'user interface testing' and 'UI testing' and found a very large number of hits and interest by practitioners on this very hot topic. A search for 'User Interface Testing' on Google, YouTube and Stack Overflow returned 74M, 237K and 6,286 hits, respectively (as of this writing: Dec. 2015).

There are also specific books, e.g., test patterns of a popular UI testing tool named Selenium [20, 21], numerous video talks on the topic in various conferences, e.g., in Google Test Automation Conference (GTAC), e.g., [22-24], various white-papers, e.g., [25], and a large number of commercial GUI testing tools. As we reviewed the SM study of GUI testing [17], none of these sources were included in its pool of sources and thus, we believe that

crucial state-of-the-practice, e.g., test patterns [20], were missing in the set of results presented by the above SM [17].

In this particular area, several important types of knowledge are missed when that SM did not include the multivocal literature, e.g., the high number of test tools and cutting-edge automations approaches available in the industry, the challenges faced and success storied made by practitioners in this area. Missing such information is of such importance that not including those information would have profound impact in even steering the research directions for the research area, i.e., when researchers do not realize the practical real-life challenges of GUI testing, their solutions is not likely going to address real industrial challenges. We believe that this issue is quite generalizable to all SE areas.

### *4.1.2 SLR on metrics in agile and lean [18]*

Similar to the previous SLR, our search string for Agile and lean metrics was: 'agile lean metrics'. We found 10M, 42K, and 129 hits in Google, YouTube and Stack Overflow, respectively. Many books exits on the topic of software metrics in general, e.g., [26] and metrics from agile and lean domain in particular, e.g. velocity, tested features. We also found several new books (published after 2015), e.g., [27, 28] on agile metrics. Also guidance about using metrics is given in various websites, marketing particular agile approaches such as the Scaled Agile Framework [29]. Tool vendors have written entries and provided Webinars on metrics as part of tool marketing and guidance effort, e.g., Rally VersionOne [30]. Additionally, several whitepapers [31] and videos [32] of the topic could be found.

An important aspect not covered in the original SLR [18] was tool support for particular metrics. It might well be that metrics used is influenced by the metrics that tools provide as default as there is a burden in creating custom metrics. Consultants' advices on goodness and suitability of the metrics [33, 34] were also missed. However, the SLR measured the importance of metrics from the primary studies, yet, the higher level advices from consultants could have been an additional benefit. The SLR collected agile and lean metrics only from the primary books presenting the methodologies for Scrum [35], XP [36], Kanban [37] and Lean [38] but this part could have been extended with more systematic search.

### *4.1.3 SLR on technical debt [19]*

Similar to the previous two SLRs, a search for 'Technical Debt' on Google, YouTube and Stack Overflow returned 9M, 216K and 693 hits, respectively. Furthermore, there are also available two specific books on technical debt [39, 40], blog entries, e.g., [41], slides and videos on the topic presented at various industrial conferences, e.g., [42], white papers, e.g., [43], as well as implementations of the concept as in SonarQube [44].

Differing from the other investigated areas, for technical debt the important role of multivocal literature to gain a holistic view on the phenomenon has already been recognized [13]. In a MLR on the dimensions, attributes, precedents and outcomes of technical debt, the authors point out that, according to their previous SLR on technical debt [19], a comprehensive definition and conceptual model are missing in the academic literature. In academic literature, code decay and architectural deterioration are commonly recognized to be major dimensions of technical debt, but other dimensions like knowledge distribution and documentation debt as well as testing debt covered in multivocal literature are missing.

### *4.1.4 Summary and meta-analysis of the three cases*

An important area missed in our SLRs focusing on UI testing, Agile metrics, and Technical debt are available tools and their features. If industry is going to follow academic advice on these areas, it will need tools to do it efficiently and effectively. Thus, an area we suggest that academic SLRs should include to ease industrial adaptation is to look at the current and available tools that industry is using. Additionally, when our advice and tools match with the state-of-the-practice tool, it will ease the technology transfer from academia to industry.

Additionally we would like to make a connection to five levels of 'closeness' between academia and industry as presented by Wohlin [45]: Level 1: Not in Touch, Level 2: Hearsay, Level 3: Sales Pitch, Level 4: Offline, and Level 5: One Team, two of which are depicted in Figure 1. This can be applied for our context in whether including or excluding grey literature from SLRs in SE. When the multivocal literature is not included in SLRs, the synthesis is conducted in a quite 'closed' environment (only the state of the art), and the results would not be very beneficial to practitioners, since the SLR contents will be mostly in the level 1 (not in touch) or at most in level 2 (hearsay), as per Figure 1. When grey literature is included, then the closeness can be characterized as Level 3 (Sales Pitch) or Level 4 (Offline).

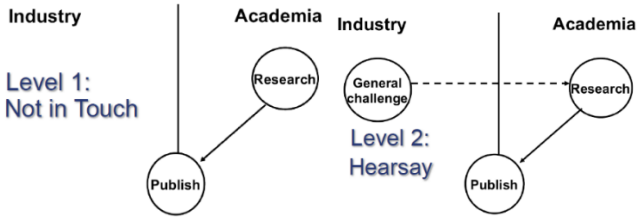


**Figure 1-The first two levels of closeness between academia and industry, proposed by Wohlin [45]**

We should also mention that, for some areas of SE, not including the multivocal literature may not lead to missing too much knowledge, e.g., the fields related to formal methods, since the 'voice of practice' is quite limited in such areas. Except for the matters on adaption of formal methods perhaps. However, for a subject such as deciding when and what to automate in testing, as we found out in our recent MLR, the voice of practice is broad and even perhaps more active than the academic literature. Thus, a SLR in such areas really has to include the multivocal literature.

## 4.2 RQ 2: What the community would gain when we include multivocal literature

### *4.2.1 MLR on ManAutoTest*

The first and the third author recently completed a MLR on deciding when and what to automate in testing (ManAutoTest), for which the online repository is available at [46]. For this MLR, only 26 were academic sources, while 52 sources were from the grey literature. We reviewed for this MLR too, the types of contributions and evidence that were taken from the grey literature to answer the MLR's RQs.

For the ManAutoTest MLR, if we were to exclude the grey sources from the pool, we would simply miss a major pile of experience and knowledge from practicing test engineers on the topic. To put this in quantitative terms, we partitioned the synthesis of a major output of that MLR (factors to be considered for deciding when and what to automate in testing) by the type of source where they were mentioned in: either formal or grey literature, as shown in Figure 2. As we can see, out of the total of

15 factor categories, grey sources contributed a total of 219 occurrences (instances) while academic sources discussed only a total 67 of factor instances.

Furthermore, we can see that, if we were to not include the grey literature, two categories (namely: test oracle and development process) would not have existed. This all denotes a major source of knowledge and experience. In addition, we extracted in the MLR study a large number of qualitative quotes, related and in support of the factors presented in Figure 2, e.g., a presentation by IBM engineers expressed: *"Main Application has lot of interdependency with other Applications which in turn cannot be automated."*, referring to the System Under Test (SUT)-related factors, and "*Once automated, regression tests can be efficient, and effective. Accordingly, ABB decided to focus its attempt to establish automated testing in build regression tests, which are most suited for automation and where the benefits could be attained*": in an industrial experience report by two engineers of the ABB Corporation.

Additionally, the type of evidence found in grey literature were valid viewpoints, ideas of cause-effect relationships that could be scientifically studied as well as explanations why and in what context some things works while others do not. We did not find any hard core empirical evidence. The stated findings were mostly based on claims and experience. However, the source of evidence was difficult to identify as the reporting was low quality. Furthermore, replication of reported results is not possible

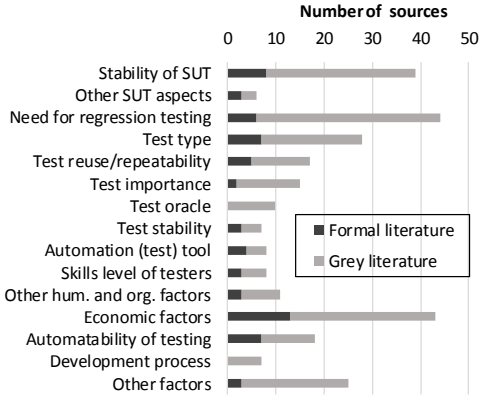


**Figure 2-A major output of the MLR on ManAutoTest**

### *4.2.2 MLR on TM/TPI*

For test maturity and test process improvement (TM/TPI), which is a field of practical and academic relevance, the first and the second author also took academic and grey literature into account to answer the study's RQs in an MLR. Overall, 130 academic sources and 51 grey literature sources were considered.

Analogous to the ManAutoTest MLR, we would have missed information from practice on test maturity and test process improvement, if we were to exclude the grey sources. For instance, one RQ in that MLR addressed the applied test maturity models. Overall, 57 different TM/TPI models were identified among the sources. From these sources 14 were grey literature reporting test maturity models such as TMap, Agile TMM or Test Maturity Index which would have been lost in a regular SLR (by not including the grey literature). Furthermore, Figure 3 shows as a major output of the MLR the number of papers per TM/TPI model using or extending a source model. Without grey literature, the usage of TMap and some other models would not have been considered. Finally, several qualitative statements on drivers, impediments, objectives, and benefits of TM/TPI, which were investigated in separate RQs, would have been simply lost without taking grey literature into account.

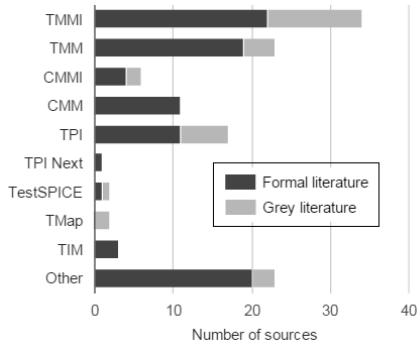


**Figure 3-A major output of the MLR on TM/TPI**

### *4.2.3 MLR on technical debt*

The previous Sections 4.2.1 and 4.2.2 have presented studies where both academic and grey literature sources have been used. For technical debt, we considered the MLR [13] which includes 35 sources and 11 industry interviews in 2013, and compared it to two regular SLRs for technical debt: (SLR 1) from 2012 with 19 sources [19], an earlier version of work for the MLR [13] with the same set of authors; (SLR 2) again a traditional SLR [47] but by different authors published in 2015 and including 94 sources.

From Table 3, we can see that SLR2 provides the most comprehensive technical debt classification from the SE viewpoint. We assume this is due to being the most recent study rather than a study conducted with superior methodology.

**Table 3-Types of technical debt, synthesized in three secondary studies**

| MLR [13]: in 2013, 35 sources, 11 interviews | SLR1 [19]: in 2012, 19 sources | SLR2 [47]: in 2015, 94 sources |
|---|---|---|
| - | - | Requirements |
| Design / Architectural | Design / Architectural | Architectural |
| Design / Architectural | Design / Architectural | Design |
| - | Unimplemented features | - |
| Code | Code decay | Code |
| Testing | Testing | Test |
| | - | Build |
| | Documentation | Documentation |
| Environment | Infrastructural | Infrastructure |
| - | - | Versioning |
| | Known issues / Defects | Defect |

Additionally, we attempted to map the causes of technical debt from all studies. Whereas, in Table 3 we saw the most recent study (SLR2 [47] published in 2015) had the most comprehensive the results, we see in Table 4 that this is not the case for technical debt 'causes'. Overall, there seems to be less consensus on the causes of technical debt. For example, SLR2 is missing causes related to the attitudes or process that may allow technical debt to go unnoticed. On the other hand, the MLR and SLR1 do not consider the technical gap and explicit decisions that cause technical debt at all.

**Table 4-Causes of technical debt, synthesized in three secondary studies**

| MLR [13] | SLR1 [19] | SLR2 [47] |
|---|---|---|
| Prioritization | Project constraints | Technical compromise |
| Pragmatism | Low visibility of debt | Environment |
| Processes | Reckless vs. prudent | Technological gap |
| Attitudes | Deliberate vs. inadvertent | Technical decision |
| Ignorance and oversight | | |

With respect to our RQ2 of finding out the benefits of MLRs, it appears that in the area of technical debt, there is currently little to gain by having MLRs. One cause could be that 57% of the sources of the most recent SLR (SLR2) come from Managing Technical Debt Workshop, IEEE software, Cutter IT Journal, and Agile Conference which are academic publication forums with high industry participation. Thus, the industry involvement in those forums could simply make MLRs obsolete.

### *4.2.4 Summary and synthesis of the three cases*

By synthesis of the above three cases, we discuss the answer to RQ2 in these aspects: (1) growth of the interest by academia versus industry to different SE topics, (2) usefulness of industry viewpoints, and (3) quality of evidence in grey literature

To assess the growth of the interest by academia versus industry for two areas of ManAutoTest and TM/TPI, Figure 4 shows the trends of the annual number of sources (formal versus grey literature) for their corresponding two MLRs. As we can see, in the case of ManAutoTest, in terms of the level of interest, the grey literature has somewhat passed the formal literature.

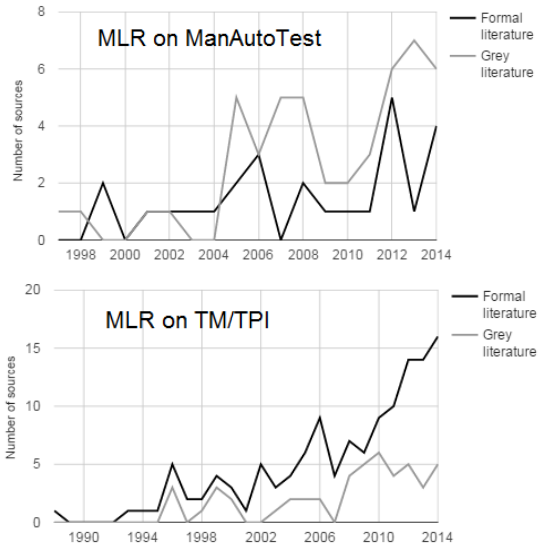


**Figure 4-Trends of the annual number of sources (formal versus grey literature) for two MLRs**

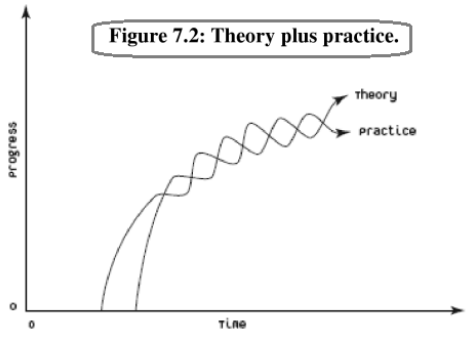


**Figure 5-A figure adopted from the "Software Creativity 2.0" book by Glass and DeMarco [4]**

Also, quite interestingly, the trends in Figure 4 resemble quite nicely to an abstract visualization of the relationship of "*Theory versus practice*" from the "*Software Creativity 2.0*" book by Glass and DeMarco [4], as shown in Figure 5. This seems to denote that sometimes the industry takes over academia in a certain field and then academic 'catches up' and vice versa. Thus, close linkage between the two 'camps' is indeed important and conducting MLRs is a good constructive effort in that direction.

## 5. DISCUSSIONS

### 5.1 Implications, recommendations, and open issues

Although our results are not comprehensive, we believe that our initial effort raises the need (awareness) for (more) MLRs in SE.

While many researchers conduct empirical studies in the form of opinion surveys and semi-structured interviews to gather voice of practitioners, we believe that a vast knowledge base of public is already available online and can be, quite conveniently, analyzed and synthesized by SE researchers without investing major time and effort in conducting opinion surveys, as we conducted for the two MLRs (ManAutoTest and TM/TPI).

Nursing literature suggests a rubric of six element when grey literature should be included in their area, see Table 5. We claim that for any SE area, there would be at least one 'Yes' in this rubric! The importance of context (the 6th element in the rubric) has been extensively been discussed in the SE literature, e.g., [48, 49]. We think that SE is still hampered by the low volume and quality of evidence (elements #4 and 5). Additionally, many of our interventions are complex with complex outcomes (elements #1 and 2) as in addition to the technical challenges we often simultaneously face challenges relating to human factors, economics, and management.

**Table 5- A rubric to aid decision making on whether to include grey literature in state-of-the-evidence reviews (taken from [9])**

| | YES | NO |
|---|---|---|
| Complex intervention | ☐ | ☐ |
| Complex outcome | ☐ | ☐ |
| Lack of consensus about measurement of outcome | ☐ | ☐ |
| Low volume of evidence | ☐ | ☐ |
| Low quality of evidence | ☐ | ☐ |
| Context important to implementing intervention | ☐ | ☐ |

Note: One or more "yes" responses suggest inclusion of grey literature.

Simultaneously, we agree there are problems in working with and including sources from the grey literature. We found that source of evidence in grey literature was often opinion or experience based rather than relying on systematic data collection and analysis as done in scientific papers. Other authors have recognized similar difficulties, e.g., [13]: "*there are apparent issues of reliability and validity associated with these writings due to their diversity*".

Due to potential value and the problems related to grey literature, the authors are planning to prepare a set of systematic guidelines for conducting MLRs in SE based on our own experiences and by adopting guidelines of MLRs in other areas, e.g., education [5] and nursing [9].

### 5.2 Limitations and threats to validity

We discuss below some of the potential threats to the validity of our study and steps we have taken to minimize or mitigate them. The threats are discussed in the context of the four types of threats to validity based on a standard checklist for validity threats presented in [50].

Internal validity: Internal validity is a property of scientific studies which reflects the extent to which a causal conclusion based on a study and the extracted data is warranted [50]. A threat to internal validity in this study lies in the selection bias (i.e., randomness of the SLR studies included in our pool of objects under study). We

did a random sampling for them and, thus, we believe we have minimized the internal validity.

Construct validity: Construct validity is concerned with the extent to which the objects of study truly represents theory behind the study [50]. Our two research questions and the data to address them were carefully selected and discussed among the three researchers to investigate the need for (more) MLRs in SE.

Conclusion validity: Conclusion validity of a study deals with whether correct conclusions are reached through rigorous and repeatable treatment [50]. To guarantee conclusion validity to a reasonable extent, we investigated three cases for each RQ and performed a synthesis for each subsequently.

External validity: External validity is concerned with the extent to which the results of this study can be generalized [50]. To answer the RQs, we selected different topics from SE to support generalization of our discussions to the whole field of SE to a reasonable extent.

# 6. CONCLUSIONS AND FUTURE WORKS

As discussed in this paper, for a practical (practitioner-oriented) field such as SE, synthesizing and combing both the state-of-the art and –practice is very important. A Multivocal Literature Review (MLR) is a systematic approach to do so. However, only a few MLRs have been published in SE so far. To address that issue, this paper raised the need (awareness) for (more) MLRs in SE. Our future work directions include the followings: (1) developing guidelines for conducting MLRs in SE, (2) conducting more MLRs, and (3) assessing the industrial usefulness of MLRs by asking our industry partners to read and evaluate them.

# REFERENCES


[1] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering–a systematic literature review," *Information and software technology,* vol. 51, pp. 7-15, 2009.

[2] K. Petersen, S. Vakkalanka, and L. Kuzniarz, "Guidelines for conducting systematic mapping studies in software engineering: An update," *Information and Software Technology,* vol. 64, pp. 1-18, 2015.

[3] A. Ampatzoglou, A. Ampatzoglou, A. Chatzigeorgiou, and P. Avgeriou, "The financial aspect of managing technical debt: A systematic literature review," *Information and Software Technology,* vol. 64, pp. 52-73, 8// 2015.

[4] R. L. Glass and T. DeMarco, *Software Creativity 2.0*: developer.* Books, 2006.

[5] R. T. Ogawa and B. Malen, "Towards Rigor in Reviews of Multivocal Literatures: Applying the Exploratory Case Study Method," *Review of Educational Research,* vol. 61, pp. 265-286, 1991.

[6] M. Q. Patton, "Towards Utility in Reviews of Multivocal Literatures," *Review of Educational Research,* vol. 61, pp. 287-292, 1991.

[7] W. F. Whyte, *Participatory Action Research* SAGE Publications, 1990.

[8] R. F. Elmore, "Comment on "Towards Rigor in Reviews of Multivocal Literatures: Applying the Exploratory Case Study Method"," *Review of Educational Research,* vol. 61, pp. 293-297, 1991.

[9] K. M. Benzies, S. Premji, K. A. Hayden, and K. Serrett, "State-of-the-Evidence Reviews: Advantages and Challenges of Including Grey Literature," *Worldviews on Evidence-Based Nursing,* vol. 3, pp. 55-61, 2006.

[10] Q. Mahood, D. Van Eerd, and E. Irvin, "Searching for grey literature for systematic reviews: challenges and benefits," *Research Synthesis Methods,* vol. 5, pp. 221-234, 2014.

[11] S. Hopewell, M. Clarke, and S. Mallett, "Grey literature and systematic reviews," in *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, H. R. Rothstein, A. J. Sutton, and M. Borenstein, Eds., ed: John Wiley & Sons, 2006.

[12] H. S, M. S, C. M, and E. M, "Grey literature in meta-analyses of randomized trials of health care interventions," *Cochrane Database Systematic Reviews,* 2007.

[13] E. Tom, A. Aurum, and R. Vidgen, "An exploration of technical debt," *Journal of Systems and Software,* vol. 86, pp. 1498-1516, 2013.

[14] I. Kulesovs, "iOS Applications Testing," vol. 3, pp. 138-150, 2015.

[15] M. Sulayman and E. Mendes, "A Systematic Literature Review of Software Process Improvement in Small and Medium Web Companies," in *Advances in Software Engineering*. vol. 59, D. Ślęzak, T.-h. Kim, A. Kiumi, T. Jiang, J. Verner, and S. Abrahão, Eds., ed: Springer Berlin Heidelberg, 2009, pp. 1-8.

[16] A. Yasin and M. I. Hasnain, "On the Quality of Grey literature and its use in information synthesis during systematic literature reviews, Master Thesis," Blekinge Institute of Technology, Sweden, 2012.

[17] I. Banerjee, B. Nguyen, V. Garousi, and A. Memon, "Graphical User Interface (GUI) Testing: Systematic Mapping and Repository," *Information and Software Technology,* vol. 55, pp. 1679–1694, 2013.

[18] E. Kupiainen, M. V. Mäntylä, and J. Itkonen, "Using metrics in Agile and Lean Software Development – A systematic literature review of industrial studies," *Information and Software Technology,* vol. 62, pp. 143-163, 6// 2015.

[19] E. Tom, A. Aurum, and R. Vidgen, "A Consolidated Understanding of Technical debt," in *European Conference on Information Systems*, 2012, p. 16.

[20] D. Kovalenko, *Selenium Design Patterns and Best Practices*: Packt Publishing Ltd, 2014.

[21] N. Garg, *Test Automation Using Selenium Webdriver with Java: Step by Step Guide*: AdactIn Group Pty, 2014.

[22] D. Dary, "GTAC 2014: Selendroid - Selenium for Android," *https://www.youtube.com/watch?v=gIPxxF4lhGo,* Last accessed: Dec. 2015.

[23] J. Kaasila, "GTAC 2015: Mobile Game Test Automation Using Real Devices," *https://www.youtube.com/watch?v=WFBfRk-GLRo,* Last accessed: Dec. 2015.

[24] G. Zhu and A. Momtaz, "GTAC 2013: Android UI Automation," *https://www.youtube.com/watch?v=O1u8iBLUFL0,* Last accessed: Dec. 2015.

[25] S. Anantharamiah, "Simplify GUI test automation by using a labor-saving design pattern," *http://www.ibm.com/developerworks/rational/library/10/simplify-gui-test-automation-using-a-design-pattern/,* 2010, Last accessed: Dec. 2015.

[26] D. Nicolette, *Software Development Metrics*: Manning Publications, 2015.

[27] D. S. Vacanti, *Actionable Agile Metrics for Predictability: An Introduction*: Leanpub, 2015.

[28] A. Maurya, *Scaling Lean: Mastering the Key Metrics for Startup Growth*: Portfolio, 2016.

[29] Scaled Agile Inc. (2015, Last accessed: Jan. 2016). *Metrics for Scaled Agile framework*. Available: http://scaledagileframework.com/metrics/

[30] CA Technologies. (2015, Last accessed: Jan. 2016). *Agile Reports and Metrics*. Available: https://www.rallydev.com/product-feature/reporting-analytics

[31] S. Pillai. (2015, Last accessed: Jan. 2016). *Agile Project Reporting and Metrics*. Available: https://www.scrumalliance.org/community/articles/2013/july/agile-project-reporting-and-metrics

[32] TechEd North America. (Last accessed: Jan. 2016). *From Vanity to Value, Metrics That Matter: Improving Lean and Agile, Kanban, and Scrum*. Available: https://www.youtube.com/watch?v=Y8ymtdLiBnA

[33] M. Levison. (2009, Last accessed: Jan. 2016). *What is a Good Agile Metric?* Available: http://www.infoq.com/news/2009/11/good-agile-metrics

[34] W. Hayes. (2014, Last accessed: Jan. 2016). *Agile Metrics: Seven Categories*. Available: https://insights.sei.cmu.edu/sei_blog/2014/09/agile-metrics-seven-categories.html

[35] K. Schwaber and J. Sutherland, *The scrum guide*: Scrum.org, 2013.

[36] K. Beck and C. Andres, *Extreme programming explained: embrace change*: Addison-Wesley Professional, 2004.

[37] D. J. Anderson, *Kanban*: Blue Hole Press, 2010.

[38] M. Poppendieck and T. Poppendieck, *Lean software development: An agile toolkit*: Addison-Wesley Professional, 2003.

[39] C. Sterling, *Managing software debt: building for inevitable change*: Addison-Wesley Professional, 2010.

[40] G. Suryanarayana, G. Samarthyam, and T. Sharma, *Refactoring for Software Design Smells: Managing Technical Debt*: Morgan Kaufmann, 2014.

[41] J. Shore. (2004, Last accessed: Jan. 2016). *Design Debt*. Available: http://www.jamesshore.com/Articles/Business/Software%20Profitability%20Newsletter/Design%20Debt.html

[42] N. Zakharenko. (2015, Last accessed: Jan. 2016). *Technical Debt*. Available: https://www.youtube.com/watch?v=JKYktDRoRxw

[43] M. Oak, "How to Minimize Technical Debt?," 2015.

[44] D. Racodon. (2015, Last accessed: Jan. 2016). *Technical Debt*. Available: http://docs.sonarqube.org/display/SONAR/Technical+Debt

[45] C. Wohlin, "Software Engineering Research under the Lamppost," in *Proceedings of the International Joint Conference on Software Technologies*, 2013.

[46] V. Garousi and M. V. Mäntylä, "Online Paper Repository for the SLR on 'When and What to Automate in Software Testing?'," in *http://goo.gl/zwY1sj*, Last accessed: Sept. 2015.

[47] Z. Li, P. Avgeriou, and P. Liang, "A systematic mapping study on technical debt and its management," *Journal of Systems and Software,* vol. 101, pp. 193-220, 3// 2015.

[48] P. Clarke and R. V. O'Connor, "The situational factors that affect the software development process: Towards a comprehensive reference framework," *Inf. Softw. Technol.,* vol. 54, pp. 433-447, 2012.

[49] K. Petersen and C. Wohlin, "Context in industrial software engineering research," presented at the Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement, 2009.

[50] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslén, *Experimentation in Software Engineering: An Introduction*: Kluwer Academic Publishers, 2000.