Potential Metrics for Agile and Lean: Systematic Literature Review and Survey

Kalyan Chakravarthy Maddila

Faculty of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona Sweden
This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Master of Science in Software Engineering. The thesis is equivalent to 20 weeks of full time studies.

Contact Information:
Author: Kalyan Chakravarthy Maddila
E-mail: kcma12@student.bth.se

University advisor:
Dr. Krzysztof Wnuk
Department of Software Engineering

Faculty of Computing
Blekinge Institute of Technology
SE-371 79 Karlskrona, Sweden

Internet: www.bth.se
Phone: +46 455 38 50 00
Fax: +46 455 38 50 57
ABSTRACT

Context: Despite continuously increasing importance of Agile and Lean in software development, the number of studies that investigate on use of metrics relevant to Agile or Lean are limited and yet few studies implements was unclear. Unclear is which are the prominent metrics that are useful in industries, and their purpose of usage.

Objectives: Main goal of this study is to find the metrics useful in Agile and Lean practicing industries; that are evaluated in industries by systematically identifying all the metrics from empirical evidence found in Literature as well as verifying which of them are prominently being used in industries. In addition, the purpose of using these metrics in industries are reported, and causes for dissatisfaction on use of some of the identified metrics among surveyed companies are investigated and reported.

Methods: Two research methodologies are used; Systematic Literature Review (SLR) and Industrial Survey. SLR is performed using snowballing as search approach to select primary studies. SLR is used to identify all the metrics that are useful for Agile and Lean software development. Rigor and relevance analysis is performed to assess the quality of the resulted primary studies. Industrial survey was conducted in order to verify and extend the empirical evidence exists in Literature regarding metrics by finding which of them are more prominently being used. Moreover causes for dissatisfaction over outcome of metrics use for process improvements were observed by performing comparative analysis between unsatisfied respondents results and satisfied respondents results.

Results: In total 20 metrics were identified from the studies having high rigor and high relevance. Moreover 11 out of these 20 metrics were identified to be prominently being used in industries using survey and other 9 metrics are found useful for Agile or Lean methods but need more awareness. Evidence from both SLR and survey shows that most of these identified or potential metrics are used for time associated purposes which are predictability, tracking, forecasting or planning, and very little evidence found for metrics that are being used directly for quality purpose. It was observed that some of the surveyed respondents who answered not satisfied with the metrics being used are not aware of the potential benefits these metrics can offer in Agile or Lean settings.

Conclusion: Evidence from both SLR and survey shows that the identified 20 metrics are very important and useful for Agile or Lean methods. 11 out of these 20 metrics are prominently being used by industries and evidence shows for other 9 metrics are also useful for Agile but needs more awareness for industries to realize their potential benefits in large scale. Also, more evidence is found for metrics that are used for time related purposes which are being dominant and important in industries than quality focused metrics. Therefore, it is important for industries not only to know which metrics are appropriate for Agile or Lean but also to have a deep understating of metrics behaviors. This will help to realize the level predictability these metric’s offer in order to make right assumptions or planning.

Keywords: Agile software development, Lean software development, metrics
ACKNOWLEDGEMENT

I would like to extend my gratitude to my supervisor Krzysztof Wnuk for his invaluable support and feedback from the idea initiation till end. This thesis would not have been possible without Wnuk endless guidance throughout the research. His perpetual energy and enthusiasm in research extremely motivated me in both personal and professional level. I should mention that it is a great pleasure for me to have Wnuk as my supervisor since his guidance, valuable suggestions, constructive criticism, and incredible patience helped me to successfully implement new ideas and finally complete the thesis in a smooth way.

I would like to thank Jürgen Börstler for his firm planning and deadlines for master thesis. In addition, he was always accessible and willing to help students by necessary updates regarding the course through Its-learning as well as personally which helped me to work productively.

I would like to thank my mom, dad and my sambo who supported me throughout my studies. I would like to express my deepest gratitude to my friend E. Rakesh who is the reason for all my positive energy in life. He is always with me in my heart and memory, although he is not in this world anymore.

Finally, I sincerely appreciate all the participants of the web questionnaire from all over the world.
TABLE OF CONTENTS

ABSTRACT .......................................................................................................................... I
ACKNOWLEDGEMENT ........................................................................................................... II
TABLE OF CONTENTS ........................................................................................................... III
LIST OF FIGURES .................................................................................................................. III
LIST OF TABLES ................................................................................................................... V

1 INTRODUCTION ................................................................................................................ 1
1.1 RELATED WORK.............................................................................................................. 2
1.2 AIMS AND OBJECTIVES ............................................................................................. 3
1.3 RESEARCH QUESTIONS ............................................................................................... 3

2 RESEARCH METHODOLOGY ............................................................................................ 5
2.1 SYSTEMATIC LITERATURE REVIEW ............................................................................ 5
2.2 SNOWBALLING PROCEDURE......................................................................................... 6
  2.2.1 Extended Design and deriving the Start set of papers:............................................ 7
  2.2.2 Performing forward and backward snowballing in iterations:.............................. 9
2.3 DATA EXTRACTION AND SYNTHESIS ....................................................................... 10
2.4 QUALITY ASSESSMENT THROUGH RIGOR AND RELEVANCE ............................... 10
2.5 INDUSTRIAL SURVEY ................................................................................................ 11
  2.5.1 Preparation of the questionnaire: ......................................................................... 12
  2.5.2 Data collection ..................................................................................................... 13

3 RESULTS AND ANALYSIS OF SYSTEMATIC LITERATURE REVIEW .......................... 14
3.1 RESULTS ....................................................................................................................... 14
  3.1.1 Start set ................................................................................................................. 14
  3.1.2 Iteration 1 ............................................................................................................. 14
  3.1.3 Iteration 2 ............................................................................................................. 15
3.2 DATA EXTRACTION AND ANALYSIS ....................................................................... 17
  3.2.1 Measures extracted .............................................................................................. 17
  3.2.2 Categorization based on study types .................................................................. 18
  3.2.3 Quality assessment based on rigor and relevance .............................................. 18
  3.2.4 Metrics resulted from empirical and non-empirical studies ............................... 19
  3.2.5 For which purpose the identified metrics are used ........................................... 21

4 RESULTS AND ANALYSIS OF INDUSTRIAL SURVEY ............................................ 24
4.1 RESULTS ....................................................................................................................... 24
  4.1.1 Respondent’s roles .............................................................................................. 24
  4.1.2 Size of the organizations .................................................................................. 25
  4.1.3 Development methodologies (Agile or Lean) .................................................... 26
4.2 RESULTED METRICS FROM SURVEY .................................................................. 26
  4.2.1 Metrics that are being used in Agile and Lean software industries ................. 26
  4.2.2 Respondents categories ..................................................................................... 28
4.3 COMPARATIVE ANALYSIS BETWEEN SATISFIED AND UNSATISFIED OVER EACH METRIC ... 29
  4.3.1 Inappropriate metrics in Agile/Lean ................................................................. 30
  4.3.2 Potential metrics and prominently being used metrics analysis ..................... 31
4.4 FOR WHICH PURPOSE THE IDENTIFIED METRICS ARE USED ............................. 34
4.5 METRICS ARE CATEGORIZED WITH THEIR FOCUS TO MEASURE TIME (PREDICTABILITY, FORECASTING, PLANNING OR TRACKING) AND QUALITY ASPECTS .................................... 35
4.6 EXPERIENCES SHARED BY SATISFIED AND UNSATISFIED RESPONDENTS .......... 36
  4.6.1 Causes Mentioned by unsatisfied respondents ............................................. 36
  4.6.2 Experiences of satisfied respondents on metrics useful .................................. 38
5 FURTHER COMPARATIVE ANALYSIS BETWEEN SLR AND SURVEY 42
5.1 POTENTIAL METRICS RESULTED FROM LITERATURE AND PROMINENTLY BEING USED IN INDUSTRIES ...................................................................................................................................... 42
5.2 FOR WHICH PURPOSE THE IDENTIFIED METRICS ARE USED ............................................ 43

6 CONCLUSION ................................................................................................................................. 45
6.1 DISCUSSION ............................................................................................................................... 45
6.2 VALIDITY THREATS .................................................................................................................. 46
6.2.1 Construct validity ................................................................................................................. 46
6.2.2 Internal validity .................................................................................................................... 46
6.2.3 External validity ................................................................................................................... 47
6.2.4 Reliability ............................................................................................................................ 47
6.3 FUTURE WORK .......................................................................................................................... 47

7 REFERENCES .................................................................................................................................... 49

8 APPENDIX A: RIGOR AND RELEVANCE SCORES ........................................................................... 52

9 APPENDIX B: QUESTIONNAIRE ..................................................................................................... 53
LIST OF FIGURES

Figure 1: Start set procedure ........................................................................................................7
Figure 2: Systematic Literature Review using snowballing approach ...........................................9
Figure 3: Snowballing Iteration-I ..................................................................................................15
Figure 4: Snowballing Iteration-2 .................................................................................................16
Figure 5: Classification of primary studies ....................................................................................18
Figure 6: Rigor and Relevance Analysis ........................................................................................19
Figure 7: Roles of Respondents .....................................................................................................24
Figure 8: Size of organizations ........................................................................................................25
Figure 9: Agile/Lean methods ..........................................................................................................26
Figure 10: Metrics Resulted from Survey ........................................................................................27
Figure 11: Satisfied and Unsatisfied Respondents on metrics use ..................................................28
Figure 12: Satisfied and unsatisfied responses over each metric .....................................................30
Figure 13: Comparisons on metrics being used by satisfied and unsatisfied respondents ........33
Figure 14: purpose of using metrics ...............................................................................................34
Figure 15: Frequency of satisfied responses over each metric with respect to process aspects ....36
LIST OF TABLES

Table 1: Research Questions ................................................................. 4
Table 2: The summary of the identified metrics ........................................... 18
Table 3: Metrics categorized according to Empirical and non-Empirical studies .......... 20
Table 4: Metrics resulted from empirical and non-empirical studies further categorized to Time and Quality ................................................................. 21
Table 5: Resulted Metrics categorized according to the frequency of responses .......... 27
Table 6: Inappropriate metrics resulted ........................................................... 31
Table 7: Prominently being used metrics from satisfied respondents in large scale .......... 32
Table 8: Potential Metrics needs more awareness to industries ............................. 32
Table 9: Metrics that are mostly being used by unsatisfied respondents ..................... 33
Table 10: Number of metrics overlapped ...................................................... 42
Table 11: Comparison of potential metrics identified from literature and survey .............. 43
Table 12: Purpose of metrics being used ...................................................... 44
1 INTRODUCTION

Nowadays, several companies discovered the potential advantages of Agile methodologies in terms of timely delivery quality software and started the transformation from traditional methods to Agile methods [19, 20]. This transition have significantly changed the software development from past 10 years [20]. Methods such as Scrum [22] and Extreme Programming [23], have a common basis in the Agile manifesto and its principles [21].

On the other hand, Traditional metrics are not sufficient in Agile software development [27]. Because most of the traditional Agile metrics fails to improve agility which certainly in conflict with Agile and Lean principles due to focus on chasing obstacles, instead of grabbing opportunities [26, 27]. For example, Anderson [27], provides a good argument for the failure of most commonly suggested Agile metrics which are called as traditional metrics. This is demonstrated by showing the metrics inability to measure Agile as well as increase agility by violating Reinertsen criteria for a good Agile metric [26], because of the focus on high cost [27].

Several researchers started investigated process metrics or measurements that aids to continuous flow and, more actionable metrics that promises transparency and predictability [4, 8, 28]. Since process metrics shows a strong indication for continuous flow [4, 14, 28]. For example, Peterson [4], increased the efficiency and effectiveness of the maintenance process by combining a palette of indicators. But these set of metrics evaluated in one only on company and in particular to maintenance flow in Lean. On the other hand, quality associated are also being used [1, 13]. It is still unclear which metrics are used for time related or flow as well as which are directed to quality.

Lean software development is considered to be Agile, but still there is subtle difference due to its focus on end to end workflow [28]. Researchers has begun investigating the benefits Lean brings, especially in terms of improvements that are observed due to the significance of metrics found in Lean software development [6, 10, 30].

Most of the metrics in Lean are concerned on work or process flow whereas metrics in Agile are concerned on both rapid delivery and high quality. Lean and Agile complement each other and overlap in several principles and practices [28], resulting in mixing the metrics from both i.e. metrics from Lean migrating to Agile [14] and metrics which are mainly used for Agile were also being used in Lean [24]. Even though, studies gave a clear understanding differences of Lean and Agile, since Lean focusing on the whole, the sub-optimization of local processes can be avoided [28]. But there is no study spotted the trends of metrics with respect to usage in Agile and Lean. Moreover from industrial perspective, which of them are used in industries and the purpose or process aspects of driving these metrics seems to be important for deep understanding.

Agile methodologies provide a different approach to requirements engineering than traditional approaches [32]. Requirements in Agile methodologies are iteratively discovered and refined while changes are discussed and introduced, often late in the projects [33]. Moreover, Agile requirements engineering is characterized by highly interactive collaboration between customer and developer, which is a well-recognized feature of most Agile requirements engineering processes [32, 33]. This helps in making better project decisions that adds value to both customer and vendor [32]. The most common scenario seems to be minimizing requirements engineering and management activities in Agile, especially metrics in Agile from requirements point of view. Before
Agile has arrived, there are many significant metrics focused to requirements activities [34].

Recently researchers started to explore the metrics in Agile and Lean from requirements point of view [5, 35]. Even though requirements in Agile is integral part of the development process, studies shows that metrics are focused to process improvements certainly associates to requirements. In an industrial case study [5], set of measures were evaluated and significant improvements of development flow is shown possible by assessing requirements flow. In a qualitative study, Wnuk et al. [35] suggested requirements associated metrics to Agile practitioners by investigating the inefficiency of currently used metrics. Not only studies shows that metrics that enforce process improvements are somehow associated to requirements, but also studies such as [5, 35] seems to indicate that process focused metrics could have strong association with requirements that enforces Agile and Lean process improvements.

1.1 Related work

In a case study [14], besides exploring implementation of Lean concepts that aids to continuous improvement and flow on an Agile software development, this study also discusses some of the metrics that can be used for measuring and tracking project progress. The metrics suggested are related to time focus, which is story rate per iteration and other metric used as an indicator which is cumulative flow diagram to plot the story rate [14].

Peterson and Wohlin [5], evaluates a set of measures to assess the requirements flow with the goals of increasing throughput and creating transparency, and found useful in the progress of whole development flow.

Wnuk K. et al [35], in his exploratory qualitative study suggested metrics from the requirements point of view. Moreover these metrics are categorized into time and quality aspects.

Several studies regarding metrics are published from past 10 years in Agile and Lean [36]. However, only a handful of systematic reviews was published that focus on process improvements and yet these reviews are either written from academic point of view [7, 9] or focusing on only Lean [9]. Author suggested that when right metrics are used in Agile it certainly leads to a user satisfaction [7]. Feyh M. [9], conducted a systematic mapping on measures and indicators in software development, leaving behind Agile. Spotting on Lean leaving behind Agile gives the metrics that are more aligned to Lean principles, due to the subtle difference from Agile. In a systematic literature review, some of the metrics that enforces was captured but all process metrics couldn’t be captured since the study focuses only on indicators or visualization techniques [17].

However, there are very few studies which focus on metrics useful for Agile or Lean settings but no study has updated with the information on which of those metrics in Agile or Lean are well being used by the industries and which of them are not much being used in industries but have the potential. This study fills this gap by investigating all the useful metrics with empirical evidence found in literature and verify with the industries.
1.2 Aims and Objectives

Process metrics are defined in this thesis as metrics that are an invaluable tool for industries wanting to monitor, evaluate, and improve their operational performance [53].

Main goal of this study is to find the prominent metrics that are being used in Agile and Lean software industries, and their purpose of usage. Moreover causes for dissatisfaction among some of the surveyed industrial respondents over metrics being used are investigated. The key research objectives for achievement of this goal include:

- First, find out useful metrics with empirical evidence available in Literature through systematic literature review.
- Second, find out which of the identified metrics that are prominently being used in industries through survey.
- Third, find out the purpose of using the identified prominent metrics through literature and survey.
- Finally, causes for dissatisfaction among some of the surveyed respondents over metrics being used were investigated by comparing with results of other respondents who are satisfied with the same metrics being used.

(Note: this task is derived when found during the survey there are one category of respondents who are not satisfied with the outcome of the metrics being used even though there are other category of respondents who answers they are satisfied with the use of same set of metrics (see section 4.3). This is important to make aware of the industries if the problem lies in not using the appropriate metrics for Agile or if the appropriate metrics not properly used)

1.3 Research Questions

The research questions investigated in this work are as follows:

<table>
<thead>
<tr>
<th>Research Questions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1. Which of the identified metrics are prominently being used in industries?</td>
<td>Two category of metrics will be resulted here: One, prominently being used metrics means metrics that are found with empirical evidence in literature and being used in large scale among surveyed industries (frequency of responses) Other, metrics which are found with empirical evidence in literature but not much being used much among the surveyed industries Main intention of this question is not only to find out metrics that are much being used but also to highlight the metrics that have potential in Agile and Lean settings but are not much aware by the industries in the real world.</td>
</tr>
<tr>
<td>RQ1.1. Which metrics are found with empirical evidence in literature?</td>
<td></td>
</tr>
<tr>
<td>RQ2. For what purpose the identified metrics are used in industries?</td>
<td>To find out for which purpose(s) these metrics are used in industries through literature and also verified through survey.</td>
</tr>
<tr>
<td>RQ3. What are the causes for dissatisfaction over metrics being used among some of the surveyed industries?</td>
<td>On other words this question is derived to know: whether the use of inappropriate metrics by these industries or lack of understanding of potential benefits these metrics (resulted from RQ1) can offer in Agile and Lean settings leads to dissatisfaction or there might be a possibility it can be both. This needs to be known to make aware of the industries if not using the appropriate metrics to Agile/Lean settings or else if potential metrics are not properly used. Note: here potential metrics refers to potential in Agile/Lean settings and evidence is gathered from literature and survey.</td>
</tr>
</tbody>
</table>

Table 1: Research Questions
2 RESEARCH METHODOLOGY

A mixed approach, qualitative and quantitative, is used in this study. The systematic literature review and industrial survey are the main activities to answer the research questions of this study.

2.1 Systematic Literature review

It is important to note that planning for systematic literature review is independent of the search approach [57]. In this research, the basic steps for planning and motivation of systematic literature review, snowballing is used as search approach instead of database search. In this section firstly why database was not chosen is explained, followed by advantages for using snowballing procedure, and extended design for start set.

Why not Database Search?

The reason database search is not chosen as search approach is due the fact that it is very difficult to formulate good search strings, since all too often the terminology used is not standardized and if search string is extended to capture in a broader perspective then a large number of irrelevant papers will be found in the search [42, 55, 56]. Moreover creates substantial manual work that also is error-prone [42]. Other challenges with the database searches mentioned in [42], such as selection of databases, different interfaces for the databases, different ways of constructing search strings, different search limitations in the databases and identification of synonyms of terms used which certainly leads to the risk of missing important literature [42, 57]. Wohlin illustrates the difficulty with inconsistency in terminology using an example which resulted when using database search few relevant papers was not caught with the search string formulated using the guidelines, but these papers were caught during the snowballing procedure [42]. In this study, terminology inconsistency was found since metrics that enforces process improvements in Agile was not specified as process metrics in most of the literature. As mentioned earlier when database search is used in this situation may results in broader perspective of capturing studies is required where search string must be extended resulting in large number of irrelevant papers and creating substantial manual work as mentioned earlier. Moreover, second goal of this research is to find the relation between process metrics and requirements which is not straight forward in the literature, where using formulating keywords (database search approach) will have a high risk of obtaining irrelevant papers.

Thus, other alternative approach which is snowballing procedure is considered, following the guidelines by Wohlin [42].

Search approach: Snowballing as a search approach

Several authors stress the importance of systematic approaches for building knowledge from literature, including evidence-based software engineering by Kitchenham et al. [37], information systems research, e.g. by Webster and Watson [38] and the concept of synthesis of research results by Hayes [41], Pickard et al. [39] or Miller [40], Wohlin [42].

Wohlin extended and detailed the steps for using snowballing as a search approach for systematic literature studies by complementing previous guidelines for systematic review in software engineering [42]. The snowballing guidelines are illustrated and evaluated by replicating a published reliability study of systematic literature review [42].
However the evaluation provides the evidence that using snowballing as the main approach is as good as the database search in terms of efficiency. The main advantage of using snowballing is its focus on papers actually referenced or papers citing papers included, where there is more possibility that the noise is less than using a database approach. Also, it was proved to be particularly useful for extending a systematic literature study, since new studies almost certainly must cite at least one paper among the previously relevant studies or the systematic study already conducted in the area [42].

**Why Google Scholar was not chosen for start set:**

Wohlin suggested google scholar to find the start set of papers, which is an initial point for the snowballing process [42]. It was mentioned that to avoid publisher bias, google scholar was used. But on the other hand Google scholar lacks in including the scope of its coverage (it finds too much information) and, lacks in providing certainty in terms of scholarly value and currency of some of the records [43]. May be for an experts in the particular study would be easy to pick up the good start set of papers, but it doesn’t seems applicable for everyone as from the observation of Google scholar disadvantages. Moreover it the tentative start set of papers resulted is mentioned far from perfect in the experimental study, since the resulted start set in the replication study had same author in common of all papers. However, since Wohlin used research question from the original study, no action was taken [42]. In this study in order to mitigate this risk care full observation from studies from author’s point of view is also taken into observation.

However the experimentation by Wohlin, showed promising results that snowballing can be used as good as multiple database search in terms of efficiency. This motivated the author in this study to choose snowballing as search approach for systematic literature review with a change in selection of database and extending the design of start set, which is further explained in section 2.2.

**Why Engineering Village is used as a database to find the start set of papers:**

In this study, ‘Engineering village’ database is used for selecting the tentative start set of papers. Knisely and Knisely suggested ‘Engineering village’ as an excellent place to start an engineering search than any other databases available [43]. This study was done by conducting comparisons of databases and search engines for engineering information with reference to the synthesis of published results provided by librarians on this topic [43].

When coming to the comparison with google scholar since it was used for finding tentative start set, it was seen that google scholar has disadvantages when including the scope of its coverage (it finds too much of information), uncertainty about the scholarly value and currency of some of the records, and the sorting of records according to how relevant and popular they are (not how current) [43]. In this study google scholar is not chosen, since the particular study is very much narrowed down i.e., from software -> Agile software -> metrics -> process -> requirements. This indicate that this study needs consistency and scope of finding relevant articles should be more where google scholar is not preferable.

### 2.2 Snowballing procedure

The snowballing method involves two steps: (1) deriving the start set of the papers and (2) performing forward and backward snowballing in iterations.
2.2.1 Extended Design and deriving the Start set of papers:

According to Wohlin, a good start set should have the following characteristics [42]:

- Targeting diversity of papers which should cover different communities, years, publishers and authors. To achieve this, it is important to have these covered in the start set.
- Number of papers or size of the start set depends on the size of the area being studied. For example, more specific focus requires fewer papers than a broad area.
- If too many papers resulted due to the general search string formulation, then papers having highly cited and more relevant references may be an alternative to obtain perfect start papers.
- To mitigate the risk of missing relevant papers using slightly different terminology, synonyms can be preferred with keywords derived from research questions.

![Figure 1: Start set procedure](image)

**Step 1: Search String and Database selection:**

During the start set identification, the preliminary search keywords were derived from the research questions, moreover keywords are taken into account based on the known articles from related work and synonyms. For example, methods such as Scrum [22] and Extreme Programming [23], have a common basis in the Agile manifesto and its principles [21], so they are included in search string. The following search string was used to identify the start set was the following:

(Requirements engineering OR requirement complexity OR requirements specification OR requirements management process) AND (metric* OR measur* OR indicator) AND (Agile OR Scrum OR XP OR Lean OR Kanban) AND (process OR management or development) AND (Software)

Identifying a good start set is very similar to the challenges in identifying search strings in database searches. As mentioned earlier in section 2.1, this research faces the difficulty with inconsistent terminology so focus here is to derive a good start set for snowballing instead of focusing on formulating search string unlike database search. Thus, a general search string is derived from the research questions, synonyms and related work. Next step is the selection of database, here Engineering village is selected.
instead of using Google scholar which is generally suggested for identifying start set [42]. Reasons for selection of Engineering village and why Google scholar is not appropriate for this research is detailed in section 2.1.

The formulated search string was executed in the Engineering Village database as outlined above. The query resulted in 818 potential papers.

**Step 2: Tentative Start set**

Here keeping the characteristics of good start set in mind, precaution was taken by excluding duplicate authors. This is because author citing his/her papers in the specific area is obvious and is obtained through snowballing. Tentative set of papers were identified from the resulted 818 papers based on the following criteria:

A. Exclusion Criteria:
   - Non-peer reviewed,
   - Duplicate authors,
   - Duplicate studies,
   - Out of scope and

B. Inclusion Criteria:
   - Time Frame: 2004 to 2014
   - Language: English
   - Title- Is it tentatively a paper to include?
      If yes:
      Abstract and Conclusion are reviewed

At the end, tentative papers which was reviewed by the author are cross reviewed by the supervisor having significant experience in the area to mitigate the risk of avoiding relevant papers.

**Step 3: Start set**

Go through Full text:

If there are studies that can’t be decided by abstract and conclusion, then the author go through the full paper and decision is made.

Looking at the number of citations and references:

In the detailed phase, one more filtration was made by looking at the number of citations and references of each candidate. This is because some of the papers were found cited among the identified tentative papers. In this case it is sufficient to take the paper into consideration that has more relevant references and highly cited since the snowballing is driven using citations and references. Important issue here is targeting diversity so after reviewing the full text, paper that has more relevant citations/references is included when compared to papers which had less citations/references. This is also because due to the motivation of snowballing that more relevant citations and references in a paper gives the possibility of more coverage of relevant studies [42]. If the citations or references are relevant or not, is justified by reading the title and abstract. After this filtration, the final candidates were included followed by final review of candidates by external reviewer.
2.2.2 Performing forward and backward snowballing in iterations.

![Diagram of Systematic Literature Review using snowballing approach](image)

**Figure 2: Systematic Literature Review using snowballing approach**

Backward and forward snowballing is applied to the resulted tentative start set of papers from step (1). This resulted in two iterations. First, backward snowballing is performed which is looking at the reference of the papers and applying following inclusion and exclusion criteria. Next forward snowballing is applied by looking at the citations of the papers. Google Scholar was used as a source to identify the citations. The inclusion criteria remained stable between the forward and backward snowballing. Here, the reason for including papers is chosen from 2004, since this is the year it was started to realize that there is much need for better metrics for Agile [27].

**A. Inclusion Criteria:**

1. **1st step: Basic Inclusion Criteria**
   - Time frame: 2004 to 2014
   - Language: English
   - Type of publication - peer reviewed

2. **2nd step: Detailed Inclusion Criteria**
   - Title – is it tentatively a paper to include?
   - Publication venue - Is it published in a place where relevant papers may be published?
   - Authors - Do we know that the authors have published relevant paper in the area studied before?
   - How they were used when referring to them, place and context of the reference

**B. Exclusion Criteria:**

The articles that don’t satisfy the above mentioned requirements and are not the formal research of requirements associated metrics in Agile or Lean would not be considered as primary studies in this work.
2.3 Data extraction and Synthesis:

MS Excel sheets were used to extract data from primary studies. Below data extraction properties are outlined and mapped with research question. After extracting, the data was synthesized by following the steps recommended by Cruzes et al. [44]. The synthesis process started by going through all the codes, then descriptive codes were organized and grouped based on their similarities.

General Information
- Paper ID
- Title
- Author(s)
- Publication year
- Source

Information extracted for RQ 1.1
- Metrics name
- Agile method
  - Agile (Scrum, XP etc)
  - Lean (Kanban)
- Metrics relation with Requirements
  - Quality
  - Flow
  - Estimations
  - Tracking or monitoring

Information extracted for RQ 2
- Time related purpose
  - Tracking/monitoring
  - Forecast
  - Predictability
- Quality focused
  - Direct focus to quality

Classification of studies for RQ 1.1
- Type of study
  - Experiment, survey, case study or interviews
- Research methods
  - Case study, experiment, survey, interview etc.

The studies are categorized in two dimensions, namely research methodology (case study, survey, framework and tool proposal) defined by Runeson [45] and type of study (evaluative, proposal, solution, validation) suggested by Wierlinga et al. [46]. Moreover quality of the primary studies are further investigated in the next section which adds empirical value to answer RQ 1.1.

2.4 Quality Assessment through rigor and relevance
Quality assessment is conducted for the primary studies obtained from snowballing. Apply rigor and relevance assessment for the studies to assess the trustworthiness of the identified metrics. The assessment is made by following the checklist for rigor and relevance proposed by Ivarsson et al. [47]. The idea behind using this model it provides the evidence to synthesize the results, since this study is more inclined for industry practitioners to get decision support on selection of appropriate metrics. See Appendix A for values given for primary studies based on the criteria below.

Checklist for Rigor and relevance followed by Ivarsson et al. [47]:

- **Rigor**
  - **Context (C):** If a study is described to the extent it can be comparable to other settings in general, in particular it explains subject’s type (graduate, professionals or researchers), development experience, development methodology, duration of the observation. If all these factors are highlighted in the study then C is assigned to 1, if any of these factors is missing then C is assigned to 0.5 and if none of the factors are mentioned then C is assigned to 0.
  - **Design (D):** When the research design in a study is transparent, structured and easy for the reader to understand the design in general and in specific mentions the outcome variables, measurement criteria, treatments, number of subjects, and sampling, then D is assigned to 1. If any of those factors is missing, then D is assigned to 0.5 and if no design is provided at all then, D is assigned to 0.
  - **Validity Threats (V):** If internal, external, conclusion and construct validity are evaluated in the study, then V is assigned to 1. If any of those mentioned threats are ignored in the study then V assigned to 0.5 and if no validity threats are mentioned then, V is assigned to 0.

- **Relevance**
  - **Users or Subjects (U):** if the subjects used in the study are industry then, U is assigned to 1. If masters (Msc.) or graduates are mentioned then, U is assigned to 0.5 and if information about subjects is missing then, U is assigned to 0.
  - **Scale (S):** when an industrial size application is used in the study then, S is assigned to 1. Otherwise, S is assigned to 0.
  - **Research Method (RM):** The chosen RM examines the real world contexts with relevance for practitioners i.e. case study, action research, experiment in a real situation and surveys/interviews. If the study belongs to any of them then, RM is assigned to 1. If the study doesn’t belong to any of them, then RM is assigned to 0.5
  - **Context (C):** if the study matches to the real world or industries settings then, C is assigned to 1. Otherwise C is assigned to 0.

**2.5 Industrial Survey**

The overall objective of the web questionnaire is to find empirical evidence to research questions. For this it is important to reach large population of practitioners. A web-based questionnaire was chosen due to the advantages over paper based questionnaire or interviews. Automation of data collection and analysis provides flexibility and convenience to both researchers and participants, also enables to reach a large number of respondents from geographically diverse locations [48].

The goal of the survey is to accumulate as much information as possible not only in relation to collected metrics resulted from systematic literature review but also to find any other process metrics that are being used by practitioners. Thus an inclusive
approach is followed, which provoked the use of convenience sampling [48, 49]. To analyze the survey results, descriptive analysis is conducted, followed by comparative analysis between respondents [49, 50, 51]. The respondents are divided into two categories: one category of respondents who answered satisfied with the metrics being used and other category was those who answered not satisfied with the metrics being used.

2.5.1 Preparation of the questionnaire:

Based on the results of the systematic literature review, the questionnaire was iteratively developed and designed. The questionnaire was discussed with the supervisor and agreements were made based on how well the questions reflected the research questions and the research goals. The questionnaire went through several amendments until it was finalized and published.

The questionnaire contained 7 questions of different formats, i.e., single and multiple choice questions. Respondents can choose pre-defined answer from the options or can supply their own answer. Open-coding method is used for analyzing those answers [50]. Author in this study don’t have a preconceived theory in mind for data analysis. But as far as satisfied and unsatisfied respondents were concerned, metrics are evaluated and compared. So, author read the answers and categorize the data according to the similar meaning by open coding.

The Questionnaire was divided into two parts: 4 questions related to metrics was designed to collect information about which metrics the respondents use, purpose of using these metrics, whether respondents were satisfied by using the metrics or not and moreover questions were designed in order to collect their experiences on using these metrics (See Questionnaire in Appendix B). Remaining 3 questions relates to demographic questions, which are designed to collect demographic information about the participants such as the role (e.g. Scrum master), type of methods used which is Agile, Lean or both and size of the organization (See Questionnaire in Appendix B). Mapping of questions in the questionnaire and research questions were depicted in a table attached in Appendix B.

The following paragraphs described how the demographics part of the questionnaire were formulated.

Roles of respondents: The targeted respondents may work on different roles in Agile software development companies. It is important to note that roles and responsibilities in Agile or Lean varies when compared to traditional software development [58]. Moreover, roles that are responsible for improving process as a whole such as keeping track of team’s as well as work progress, requirements (working with user stories, acceptance tests) are considered as target audience for this questionnaire. This is because these responsibilities holding roles obviously work with metrics and measurements [58]. A list of possible roles in Agile development was synthesized as follows [58]:

Only coach in XP is included since he/she is responsible for the process as a whole who keeps track of the project’s process [58]. Where as in Scrum all the four roles involved are included in Questionnaire. Scrum master and manager were involved in team’s progress and provides directions to keep the work according to plan [58]. Product owner and Scrum team involved in developing user stories and conducts acceptance/unit tests [58]. In Lean only project leader or Kanban team leader is responsible for time estimations and the team progress [58]. Rest or other roles such as technical coordinator, programmer, chief architect etc. are not considered. Since their responsibilities were not involved in working on metrics or measurements that enforce process improvements [58].
Size of the Organizations: Size of the organization is important for analyzing the results. The classification that is followed in this questionnaire has no names instead ranges of number of employees in organizations were used which is inspired from [59]. This is because there is no agreement on classification of organizations size in the literature. The categories are: less than 10 employees, 11-20 employees, 21-50 employees, 51-100 employees, 101-200 employees, 201-500 employees, and over 501 employees.

Development Method (Agile, Lean or both): It is possible that organization using Agile methods can migrate to Lean or Kanban, which means metrics used in Agile can be applied in Lean and vise-versa. This is important to make sure only Agile or Lean practicing industrial respondents answers are collected, since this study is only limited to Agile and Lean context.

2.5.2 Data collection

The survey was conducted using the website called Survey Monkey which is a web-survey support service. Invitations to participants were sent to potential audience via emails or social network websites for ex. www.linkedin.com, google professional groups and yahoo professional groups, www.xing.com. These professional social networks increase the number of potential respondents. The strategy used to target potential audience by the keywords agile and leans as main and requirements or metrics or measures for combinations are used. Author used the possible combinations of those keywords to find as many relevant groups as possible and then posted the questionnaire.

In addition, emails were also sent to individuals based on their designed titles e.g. project manager, scrum master, agile mentor, only to name a few. The names and email addresses were acquired from the reference in literature, blogs and professional social networks. Here, master students and undergraduates as well as less experience employees working in agile/lean were avoided because their experience could have been insufficient in order to answer the questionnaire.

All information or answers from respondents were automatically stored by SurveyMonkey.com. The questionnaire was published online on 8th August 2014 and was switched after 60 days on 8th October 2014. In total, about 900 individual invitations were sent to people either on professional social networks or by emails. The number of collected responses is 89, and 33 of them were deleted because the respondents did not fill the entire questionnaire. In total, 56 responses are analyzed.
3 RESULTS AND ANALYSIS OF SYSTEMATIC LITERATURE REVIEW

3.1 Results
In total 4 papers (marked C1, C2, C3 and C4: see in phase 2) were selected for the start set from 818 papers resulted from Engineering Village database.

3.1.1 Start set

Phase 1:
In total, 18 candidates from 820 candidates were identified for inclusion for tentative start set (next phase), after applying inclusion and exclusion criteria.

Phase 2:
Finally 4 candidates for inclusion were identified from 18 tentative start set based on number of relevant citations and references. First going through the title, look at the relevant study and then abstract. Finally full-text of all the 4 candidates is read before pileup to snowballing.
The 4 candidates are now denoted as C1, C2, C3 and C4 to indicate that they are candidates for inclusion. The 4 candidates are:


3.1.2 Iteration 1

3.1.2.1 Backward Snowballing:
During backward snowballing, 67 references and 74 citations were evaluated. 35 references were removed based on the publication year, 3 were not in English, 5 were removed based on the publication type, 2 were not pair-reviewed, and 11 were removed based on the title screening, 5 removed after abstract or full paper screening. Finally two papers denoted (C5 and C6) are included.

3.1.2.2 Forward Snowballing:

During forward snowballing, 30 citations were removed based on the title, 4 based on the publication language, 16 based on the abstract and 11 turned out to be duplicates. Finally, three papers denoted (C7, C8 and C9) were included.


In total, first iteration resulted in including 5 papers, denoted C5, C6, C7, C8 and C9. The results from the first iteration of forward and backward snowballing are depicted in Figure 3.

3.1.3 Iteration 2

3.1.3.1 Backward Snowballing:

During backward snowballing, 170 references were examined in this iteration. 36 references were removed based on the title, 73 were duplicates, 64 were removed based on the year, 36 were removed based on the publication year and 2 references were removed after abstract or full read. Finally, 5 candidates (denoted C10, C11, C12, C13 and C14) are included.


3.1.3.2 Forward Snowballing:

During forward snowballing, since papers C7, C8 and C9 had no citations in Google scholar, only 74 citations were analyzed in this iteration. 20 citations turned out to be duplicates, 31 were removed based on reading the abstract and 19 based on reading the title. Finally 3 candidates (denoted C15, C16, C17) are included from citations of C5 and one (denoted C18) from citations of C6.


In total second iteration resulted in including 9 papers, denoted C10 to C18. The results from the second iteration of forward and backward snowballing are depicted in Figure 4.
3.2 Data Extraction and Analysis

3.2.1 Measures extracted

29 measures and metrics are identified from the 18 identified papers, denoted as M1, M2 and so on, see Table 2. We focus on metrics associated with requirements as thus ignoring other metrics suggested among these papers. The identified metrics are categorized according to the project type (Agile and Lean) they are already applied or used.

10 measures are described for Agile projects, 12 measures are described for Lean projects, and 5 measures are described for both Agile and Lean. Only 2 measures described for both Agile and Lean. M29 is mentioned in 6 studies and M2 is mentioned in 5 studies. M4 is repeated by three studies whereas M3, M5, M9, M13, M14, M22 and M24 are mentioned by two studies. Remaining measures are just mentioned by one study.

<table>
<thead>
<tr>
<th>ID</th>
<th>Metric name</th>
<th>Project type</th>
<th>Mentioned in Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Cosmic</td>
<td>Agile</td>
<td>C1</td>
</tr>
<tr>
<td>M2</td>
<td>Project Velocity/ measure of throughput</td>
<td>Agile &amp; Lean</td>
<td>C2, C3, C7, C9, C12</td>
</tr>
<tr>
<td>M3</td>
<td>Function points (requirements inventory)</td>
<td>Agile &amp; Lean</td>
<td>C8, C10, C16</td>
</tr>
<tr>
<td>M4</td>
<td>Lead time/Time in State (TIS)</td>
<td>Agile</td>
<td>C8, C15</td>
</tr>
<tr>
<td>M5</td>
<td>Queue/ measure of Queue</td>
<td>Agile &amp; Lean</td>
<td>C8</td>
</tr>
<tr>
<td>M6</td>
<td>Work In Progress (WIP)/ Number of work items</td>
<td>Agile</td>
<td>C8, C9</td>
</tr>
<tr>
<td>M7</td>
<td>Requirements Ambiguity</td>
<td>Agile</td>
<td>C7</td>
</tr>
<tr>
<td>M8</td>
<td>Requirements completeness</td>
<td>Agile</td>
<td>C7</td>
</tr>
<tr>
<td>M9</td>
<td>Aspectual Density per Sprint</td>
<td>Agile</td>
<td>C7</td>
</tr>
<tr>
<td>M10</td>
<td>Requirements maturity index</td>
<td>Agile</td>
<td>C7</td>
</tr>
<tr>
<td>M11</td>
<td>Problem per user month (PUM)</td>
<td>Agile</td>
<td>C7</td>
</tr>
<tr>
<td>M12</td>
<td>RTF curve</td>
<td>Agile</td>
<td>C7</td>
</tr>
<tr>
<td>M13</td>
<td>Story point</td>
<td>Agile/Scrum</td>
<td>C9, C12</td>
</tr>
<tr>
<td>M14</td>
<td>Product backlog/ Product backlog rating</td>
<td>Agile/Scrum</td>
<td>C17, C18</td>
</tr>
<tr>
<td>M15</td>
<td>Story rate per iteration</td>
<td>Agile &amp; Lean</td>
<td>C14</td>
</tr>
<tr>
<td>M16</td>
<td>Maintenance inflow</td>
<td>Lean</td>
<td>C4</td>
</tr>
<tr>
<td>M17</td>
<td>workload</td>
<td>Lean</td>
<td>C4</td>
</tr>
<tr>
<td>M18</td>
<td>Combined analysis</td>
<td>Lean</td>
<td>C4</td>
</tr>
<tr>
<td>M19</td>
<td>Linear regression</td>
<td>Lean</td>
<td>C5</td>
</tr>
<tr>
<td>M20</td>
<td>Estimation error</td>
<td>Lean</td>
<td>C5</td>
</tr>
<tr>
<td>M21</td>
<td>cost</td>
<td>Lean</td>
<td>C5</td>
</tr>
<tr>
<td>M22</td>
<td>Value stream mapping</td>
<td>Lean</td>
<td>C10, C13</td>
</tr>
<tr>
<td>M23</td>
<td>Defect state over time</td>
<td>Lean</td>
<td>C11</td>
</tr>
<tr>
<td>M24</td>
<td>variance</td>
<td>Lean</td>
<td>C9, C11</td>
</tr>
<tr>
<td>M25</td>
<td>Fault Slippage</td>
<td>Lean</td>
<td>C9</td>
</tr>
<tr>
<td>M26</td>
<td>Cycle time</td>
<td>Lean</td>
<td>C9</td>
</tr>
<tr>
<td>M27</td>
<td>Rework rate</td>
<td>Lean</td>
<td>C9</td>
</tr>
</tbody>
</table>
Table 2. The summary of the identified metrics.

<table>
<thead>
<tr>
<th>M28</th>
<th>Cumulative flow diagrams (CFD)</th>
<th>Agile &amp; Lean</th>
<th>C4, C5, C9, C11, C14, C17</th>
</tr>
</thead>
<tbody>
<tr>
<td>M29</td>
<td>Burn down charts</td>
<td>Agile &amp; Lean</td>
<td>C9, C17</td>
</tr>
</tbody>
</table>

3.2.2 Categorization based on study types

The resulted studies are categorized in two dimensions, namely research methodology (case study, survey, framework and tool proposal) defined by Runeson [45] and type of study (evaluative, proposal, solution, validation) suggested by Wierlinga et al. [46], see Figure 5.

Out of 18 studies, 11 papers (C4, C5, C6, C8, C10, C11, C13, C14, C15, C16, C18) were classified as evaluation research and, 2 papers were proposing solutions; C1 that suggested COSMIC method for ensuring the quality of the user stories and C3 that suggested throughout metrics for Agile based on value). One paper (C2) proposed a framework for measuring the impact of RE processes on the organization. Three papers (C7, C9 and C17) described surveys (literature reviews or industrial surveys but one was a combined survey and evaluation research) and one paper (C12) did not describe the research method.

![Figure 5: Classification of primary studies](image)

Figure 5: Classification of primary studies

3.2.3 Quality assessment based on rigor and relevance

We classified 11 studies as having the highest rigor and relevance (C4, C5, C6, C8, C10, C11, C13, C14, C15, C16, C17), see top-right in Figure 6. Moreover, we classified 5 studies (C1, C2, C3, C7, C12) as having low rigor and low relevance, see bottom left in Figure 6. Finally, only one study (C9) has high rigor and low relevance see bottom right in Figure 6. One study has low rigor but high relevance (C18), see top-left in Figure
6. Table for rigor and relevance values distribution over the primary studies were attached in Appendix A.

![Rigor and Relevance Analysis](image)

**Figure 6: Rigor and Relevance Analysis**

### 3.2.4 Metrics resulted from empirical and non-empirical studies (RQ 1.1)

Total 11 studies (C4,C5,C6,C8,C10,C11,C13,C14,C15,C16,C17) are resulted from classification of studies (see Fig. 6.) were also resulted with high rigor and relevance. This provides strong empirical evidence for the metrics provided from these 11 studies. Therefore, 18 metrics were identified (M2, M3, M4, M5, M6, M14, M15, M16, M17, M18, M19, M20, M21, M22, M23, M24, M28, M29) from the 11 empirical studies. Moreover all the systematic literature review related studies identified from the resulted primary studies are reviewed to check if the metrics mentioned in these studies are classified from empirical background. 2 metrics; story point (M13) and cycle time (M26) are identified from empirical background, since the classification in the systematic literature study (C9) in which these metrics are mentioned were found to be empirical studies [9]. There are no other metrics found empirical from the systematic literature studies identified from the primary studies. Therefore in total 20 metrics are identified as prominent metrics which are used in Agile and Lean industries.

Among evaluation research type papers, in paper (C11) suggested measuring defect state over time (M23), variance (M24) and visualizing the flow using the cumulative flow diagrams (M28) [11]. The application of these measurements in a case study showed that during and after the transition to Agile practices, the defect inflow is higher, bottlenecks are visible, and defect resolution cycles are longer when compared to before the transition [11].

Kai Petersen [4], proposed and evaluated four Lean indicators concerned at improving maintenance process by detecting and eliminating waste. Among which three indicators are identified, which associate with requests from customers or customer needs i.e., measurement inflow (M16), workload (M17) and combined analysis (M18) [4]. For visualizing and analyzing the requirements flow and maintenance inflow of requests or waste for each phase, cumulative flow diagram (M28) is used [4, 5]. These Lean indicators were evaluated in an industrial case study at Ericsson AB and the study
demonstrated that these indicators were able to identify the presence or absence of inefficiencies and ineffectiveness in the maintenance process [4]. Petersen [5], also suggested estimation error (M20) as a way to quantify the continuous the requirements flow, linear regression (M19) as a way to quantify the bottlenecks in the requirements flow and to save costs author proposed cost measure (M21)

Other authors, also working with Ericsson, focused on evaluating monitoring and tracking workflows (C13) [13]. Next, several studies (C6, C8, C16) evaluated lead-time (M4), Queue (M5), Work In Progress (M6) while paper (C10) focused on evaluating waste measures by using value stream mapping (M22) at Ericsson [6, 8, 16]. In addition, study (C6) evaluates the function points (M3) [6]. Furthermore, one paper (C14) focused on evaluating how story rate per iteration (M15) can be used to track the development progress in a web-based application [14] and one paper (C15) evaluated how monitoring throughput (M2) and queue (M5) can be used for discovering process bottlenecks at Ericsson [15]. Finally, rating the product backlog (M14) and associated product complexity factors were evaluated as a way to measure the quality of the test process at a company called Software People (C18) [18].

Furthermore, metrics that are resulted from empirical studies and remaining which are mentioned in the non-empirical studies are structured in Table 3 according to their usage with respect to Agile and Lean methods. Total 7 metrics (M2, M5, M13, M15, M26, M28, and M29) common metrics were identified which are applied in both Agile as well as Lean methodologies settings.

<table>
<thead>
<tr>
<th>Agile</th>
<th>Lean</th>
<th>Common (Agile/Lean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2, M3, M4, M5, M6, M13, M14, M15, M26, M28, M29</td>
<td>M2, M5, M13, M15, M16, M17, M18, M19, M20, M21, M22, M23, M24, M26, M28, M29</td>
<td>M2, M5, M13, M15, M26, M28, M29</td>
</tr>
<tr>
<td>Evaluated metrics from Empirical studies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1, M7, M8, M9, M10, M11, M12</td>
<td>M25, M27</td>
<td>-</td>
</tr>
<tr>
<td>Metrics identified from non-empirical studies</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Metrics categorized according to Empirical and non-Empirical studies
3.2.5 For which purpose the identified metrics are used (RQ2)

<table>
<thead>
<tr>
<th>External attributes</th>
<th>Evaluated Metrics from empirical studies</th>
<th>Other Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality (inventory)</td>
<td>M3, M14</td>
<td>M1, M7, M7, M8, M9, M10, M12, M25</td>
</tr>
<tr>
<td>Time (Predictability, tracking, forecasting and planning)</td>
<td>M2, M4, M5, M6, M13, M15, M16, M17, M18, M19, M20, M21, M22, M23, M24, M26, M28, M29</td>
<td>M11, M27</td>
</tr>
</tbody>
</table>

Table 4: Metrics resulted from empirical and non-empirical studies further categorized to Time and Quality

**Quality focused metrics:**

Metrics were extracted which are focused to quality in Agile and Lean projects as whole. A broad perspective on quality and consider the process quality, the requirements quality and other associated quality aspects such as inventory purpose of using the metrics are included into this category. Total 9 metrics are classifies into this category, among which only 2 were empirical, see Table 4.

Quality of requirements in Agile can be improved by measuring the documentation using the Cosmic method (M1) [1]. Similarly, Function points (M3) can be used to improve the quality requirements effort estimation [6]. Other requirements quality aspects include measuring requirements ambiguity (M7), completeness (M8), aspectual density per sprit (M9), maturity index (M10) and RTF curve (M12). M7 and M8 are focused indirectly by counting the number of misinterpreted requirements per total number requirements and the number of correctly validated requirements per total number of requirements respectively.

Fault slippage (M25) measured as the number of work items that can overload the requirements process and hinder processing requirements was suggested in [8, 9]. By detecting, e.g. the number of requirements waiting to be implemented, or faults to be fixed early detection of potential defect can be performed, which in turn improves the requirements process quality.

Only two quality related metrics (M3, M14) were mentioned in evaluation studies conducted with high rigor and relevance at Ericsson or Though Works see also Section 3.2.2. M1 was proposed in a solution paper while the remaining metrics (M7, M8, M9, M10, M12 and M25) originate from literature studies. It can thus be suggested that quality related requirements metrics have not yet been extensively evaluated in industry with Ericsson being the leader in these evaluations.

**Time focused metrics:**

Time related metrics are associated with requirements flow, velocity/throughput or other temporal aspect of working with requirements. From the flow perspective, throughput is defined as the number of elements processed during a time period [15]. Moreover purpose/process aspects of these metrics used in industries for multiple use but most common or usually mentioned in literature are predictability, tracking, forecasting and planning [4, 5, 6, 14, 15, 17]. Thus these purposes observed are related to time aspects since they enforces process improvements through timely manner [10,
It should be noted that if a metric that describes the number of elements to be process or processed during a time frame then this metric is considered as time focused.

20 metrics were classified into this category and 16 of them were actually evaluated in empirical studies, three metrics from literature surveys and one in a mixed type studies, see Table 3. It can thus be suggested that time-related measures are more popular among the identified papers and more often evaluated or applied in empirical contexts. This is logical as these metrics are more closely related to the fundamentals of both Lean and Agile that advocate focusing on speed and frequent delivery.

Velocity (M2) is identified as the efficient metric for change indicator component for XP projects by classifying the former Agile components to measure their impact on requirements engineering processes [2]. Project velocity is designed to improve the workflow during the iterations, which in turn helps customer and development teams to make better estimations on implementations of requirements and applies in every kind of Agile project [2, 7].

Velocity (M2) is suggested as a way to track the amount of product backlog or number of work items that a team completes in a single sprint [9, 12]. Lead Time or Time In State (TIS) (M4) is one of the most commonly recommended metric in Lean software and is identified as inventory based Lean metrics because it apparently reduces the risk that the requirements becomes outdated and provides predictability[8, 9, 16]. Value Stream Mapping (M22) uses Lead Time (M4) in identifying lead time or waste caused by requirements, summarized as motion wastes [10, 13].

On the other hand, queue metric (M5) is also a leading indicator which enables preventive actions by indicating the future lead time [8], or discovering and monitoring potential bottlenecks with support of throughput [15]. Despite being rather difference concepts, queue and throughput complement each other for bottleneck identification. Moreover, Work In Progress (M6), defined as the number of work items is help preventing from large queues, especially in the requirements phase and can be used as a base metric for measuring throughput via the number of work items completed [8, 9] or velocity (M2) [9, 12]. Finally, the number of completed story points (M13) per a single sprint or the number of story points per iteration (M15) [7, 9, 12] can be used to measure throughput. Story rate per Iteration (M15) helps to measure productivity over the project duration. M15 was identified during an investigation of continuous improvement through Lean principles.

Petersen et al. [5], evaluated at Ericsson three metrics associated with requests from customer needs in Lean software development are: measurement inflow (M16), Workload (M17) and Combined analysis (M18). Also, for visualizing and analyzing the maintenance inflow of requests or waste for each phase, cumulative flow diagrams (M28) are used and the results demonstrates that these indicators were able to identify the presence or absence of inefficiencies in the maintenance process [4]. M28 is also used to visualize the flow of requirements [5], and for representation of requirements elicitation and validation [17], which was even recommended as a better alternative to burn down charts used in Agile software development [14].

Petersen also suggested estimation error (M20) as a way to quantify the continuous the requirements flow, linear regression (M19) as a way to quantify the bottlenecks in the requirements flow and to save costs author proposed cost measure (M21) [5]. The results from industrial evaluations of M19, M20 and M21 are promising [5]. From the systematic mapping study [9], variance (M24), fault slippage (M25), cycle time (M26) and rework rate (27) are identified as time related metrics which revolve around or
associates to requirements. In a qualitative survey, respondents mentioned that they used burn down charts (M29) to keep the team aware about how accurate the estimates were and how well the teams performed [17].
4 RESULTS AND ANALYSIS OF INDUSTRIAL SURVEY

4.1 RESULTS

4.1.1 Respondent’s roles

One of the question in questionnaire was formulated in order to find the roles or designations of the respondents. In Agile development, less agreements on roles are presented and most of the seniors or experts in Agile manage multiple roles [58, 59]. For example; consultant, coach and mentor can be titled for a senior as Agile coach. Thus a list of possible roles from studies presents the roles and responsibilities, according to the definitions and principles [54, 58, 59]. Therefore the question is designed as open-ended question where respondents can choose the answer from the options given or specify role in the text box provided. (See question 6 in Appendix B)

The most common roles from who answered are Scrum masters (15.7% or 11 responses) and project managers (15.7% or 11 responses). Others and Agile mentors roles have the same number of responses (13.7% or 9 responses). Answers provided in additional textbox include roles such as business analyst or area manager in Agile, Lean Agile coach, configuration manager, PMO (project management officer), Agile consultant and trainer. Author merged the roles mentioned in others to the respective categories of roles listed. Third largest category was product owner (11.8% or 8 respondents). Scrum team developers and Agile team developers had the same number of responses (9.8 % or 7 each), followed by requirements engineer and Agile team developers who also had the same number of responses (7.8% or 5 each). The remaining responses are working as Kanban team leader (3%), Independent tester (2%) and software tester (2%).

Figure 7: Roles of Respondents
We can see that the first and second place categories are relatively high in number since Scrum masters, product managers, Agile mentors and almost everyone mentioned in other category are generally few in an organization. Moreover the majority roles of responses (Scrum master, Agile mentor, Agile and Kanban team leaders, product and project managers) who answered this questionnaire are seniors and called experts in Agile methods [59]. This would seem to indicate that the senior or experts in this domain represented a substantial part of the respondents. This gives the questionnaire extra credibility and allows for the generalization of results.

### 4.1.2 Size of the organizations

Respondents were asked for size of their organization. This question allows selecting only one from the predefined answers. Since the respondent allowed choosing only one option, the results of the questions are computed based on the number of respondents received for each category (See question 7 in Appendix B). The computation is done over the number of respondents and not over the number of responses.

![Figure 8: Size of organizations](image)

The answers depicted in figure 8, we can see nearly half of the respondents (47.1% or 30 out of 56 respondents) work for large companies (>501 employees). 13.7% (7) of respondents come from the organizations holding 51-100 employees. 9.8% of respondents come from organizations having 101-200 and 201-500 employees (5 each). 6 respondents worked for companies with less than 10 employees, one respondent worked for a company between 11-20 employees and 2 respondents worked for a company between 21 and 50 employees.

Here it is important to note that having 47.1% of the respondents coming from very large organizations and 33.3% of the remaining half coming from 51-500 employees can give the results of the questionnaire more credits as these organizations would most likely have well-defined software development processes, and for being certified such as CMMI. This gives the questionnaire extra credibility and allows for generalization of results.
4.1.3 Development methodologies (Agile or Lean)

In this question respondents were asked in which method the metric(s) you mentioned usually use or apply (See question 3 in Appendix B). This is to ensure that all the respondents who answered were practicing Agile or Lean development practices. Respondents are allowed to choose one answer among the predefined list.

![Figure 9: Agile/Lean methods](Image)

The results of the question are presented based on the total number of respondents received for each category.

Highest number of respondents (59% or 33 respondents) answered metrics they are using applied in Agile. Lean methods (9 % or 5 respondents) receives the least number of respondents where metrics are applied. 32% (18 respondents) of respondents answered that they used the metrics both Agile and Lean settings. Here it is important to note that the purpose of asking this question to respondents is to make sure that all the respondents considered are only practicing Agile or Lean development practices. Thus, all the 56 respondents are considered for further analysis since all of them were practicing either Agile or Lean or both development methods.

4.2 RESULTED METRICS FROM SURVEY

4.2.1 Metrics that are being used in Agile and Lean software industries

This question is an open-ended question where respondents are allowed to select more than one answer in the predefined list because there might be more chance that organizations use multiple metrics. Respondents also are allowed to add metrics if they are not present on the list in the text box provided with name ‘others’. (See question 1 in Appendix B)
Division of survey resulted metrics into categories

**Category 1:** Metrics which obtain the frequency > 30 responses

**Category 2:** Metrics which obtain the frequency of responses 20 to 29 falls into this category. Since the variation in response percentage is very close i.e., 6 to 8%, these metrics are clustered to one category.

**Category 3:** Metrics which obtains the frequency of responses varying from 10 to 19 responses falls into this category (3 to 5%).

**Category 4:** Metrics which gets <10 responses fall into this category. Since, we can see in figure 14. Most of the metrics are clustered within this boundary.

<table>
<thead>
<tr>
<th>Category</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>work in progress (WIP) (M6), story point (M13), burndown chart (M29) and product backlog (M14)</td>
</tr>
<tr>
<td>2</td>
<td>Velocity (M2), Cost (M21), cycle time (M26), lead time (M4) and CFD (M28)</td>
</tr>
<tr>
<td>3</td>
<td>Function points (M3), Queue (M5), Linear regression (M19), rework rate (M27), story rate per iteration (M15), workload (M17), estimation error (M20), VSM (M22), defect state over time (M23)</td>
</tr>
<tr>
<td>4</td>
<td>Cosmic (M1), Req. ambiguity (M7), req. completeness (M8), aspsectual density per sprint (M9), Requirements maturity index (M10), problem per user month (M11), rtf curve (M12), maintenance flow (M16), combined analysis (M18), variance (M24), fault slip through (M25), others</td>
</tr>
</tbody>
</table>

Table 5: Resulted Metrics categorized according to the frequency of responses
Total 5 additional metrics were answered by respondents in the additional text box provided. In which 4 of them were identified as already mentioned in the list, these metrics are time to market, velocity, story points and lead time. One metric identified as not relevant to process metrics which is ROI - rate of interest. Only one new metric found which is called unexpected issues per sprint.

4.2.2 Respondents categories

One of the goal of this thesis is to find the prominent metrics using the frequency of satisfied responses. Moreover to find the causes for dissatisfaction over metrics use over process improvements. For these goals to achieve it is important to distinguish satisfied respondents from unsatisfied respondents. Question asked was whether respondents were satisfied with the outcome of metrics being used (see question 5 in Appendix B). Moreover metrics were analyzed accordingly in section 4.3 and section 4.4.

Respondents are allowed to choose only one option, if they are satisfied or not. If respondents opts ‘no’, they reason ‘why’ is also asked by providing a text box.

Since the respondent allowed choosing only one option, the results of the questions are computed based on the number of respondents received for each category.

37 respondents were satisfied with the metrics and the remaining 19 respondents was not satisfied with the metrics (See Figure 11). Only 8 respondents mentioned the reason why they are not satisfied with the current metrics provided in other category. These 8 respondent answers were analyzed in the section 4.6.1, which is appropriate to discuss the reasons since that section investigates on causes for dissatisfaction of respondents (see section 4.6).

It is important to note that, more number of respondents are satisfied which are nearly double to that of unsatisfied respondents, see Figure 11. This seems to indicate that the satisfied respondent’s results represent a substantial part of overall respondents. So, comparison analysis between satisfied and unsatisfied respondent results is further investigated to detect any anomalies that could be found. This gives the results extra credibility and allows for the generalization of results.
4.3 Comparative analysis between satisfied and unsatisfied over each metric

<table>
<thead>
<tr>
<th>Category_ID’s</th>
<th>Metrics</th>
<th>Satisfied vs. Unsatisfied Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>product backlog</td>
<td><img src="image1.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>work in progress (WIP)</td>
<td><img src="image2.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>burndown chart</td>
<td><img src="image3.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>story point</td>
<td><img src="image4.png" alt="Graph" /></td>
</tr>
<tr>
<td>2</td>
<td>Velocity</td>
<td><img src="image5.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>CFD</td>
<td><img src="image6.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>lead time</td>
<td><img src="image7.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>cycle time</td>
<td><img src="image8.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>Cost</td>
<td><img src="image9.png" alt="Graph" /></td>
</tr>
<tr>
<td>3</td>
<td>VSM</td>
<td><img src="image10.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>defect state over time</td>
<td><img src="image11.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>Estimation error</td>
<td><img src="image12.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>story rate per iteration</td>
<td><img src="image13.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>Queue</td>
<td><img src="image14.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>Function point</td>
<td><img src="image15.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>workload</td>
<td><img src="image16.png" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>Linear regression</td>
<td><img src="image17.png" alt="Graph" /></td>
</tr>
</tbody>
</table>
In specific, how many respondents are satisfied and unsatisfied over each metric is analyzed (See Figure 12). This enables the transparency to observe the frequency of responses by satisfied and unsatisfied respondents over each metric.

As depicted in figure 12 we can observe the frequency of satisfied and unsatisfied respondents over each metric. We can see that all the metrics in the categories 1, 2 and 3 shows satisfied responses domination over metrics when compared with unsatisfied responses. This is not enough to say these are prominent metrics, so further analysis were done in the section 4.3.2. On the other side there are set of metrics observed in category 4 where unsatisfied responses was dominated over satisfied responses on metrics,

4.3.1 Inappropriate metrics in Agile/Lean

In category 4 (see figure 12), in total 7 metrics were found dominated by unsatisfied responses. These metrics are: Cosmic (M1), Req. ambiguity (M7), req. completeness (M8), aspectual density per sprint (M9), Requirements maturity index (M10), problem per user month (M11) and rtf curve (M12). This shows a clear indication that these are not efficient metrics which doesn’t seem to show process improvements in Agile or Lean. This could be one of the cause for overall dissatisfaction of metrics by respondents.

Even though in literature, these metrics claimed to be good in Agile for example; cosmic (M1) claimed to improve the quality of Agile user stories and rework rate to track the creation of process output [1], but in real industrial usage evidence shows these are not useful in Agile/Lean. Moreover from the analysis of metrics identified from non-empirical studies found through Literature analysis all of these 7 metrics comes under this category (see table 3). Thus evidence from both literature and survey analysis concludes that these 7 metrics was found to be inappropriate in Agile/Lean settings and are not included for further analysis.
Furthermore only 4 metrics (maintenance flow (M16), combined analysis (M18), variance (M24), fault slip through (M25)) are left in category 4 whose responses are less than 10 responses. Even though responses for these metrics is less but with respect to satisfied vs. unsatisfied responses, says that more number of respondents were satisfied with these metrics except for maintenance flow metric (M16). Maintenance flow metric (M16) had equal number of satisfied (3) and unsatisfied responses (3), so this metric is left in for further analysis since there might be the chance this metric may not be used properly by unsatisfied respondents (See in Figure 12).

In conclusion, evidence shows that 7 metrics (listed below) was found to be not useful or inappropriate in Agile or Lean projects.

| Inappropriate metrics in Agile or Lean | Cosmic (M1), Req. ambiguity (M7), req. completeness (M8), aspectual density per sprint (M9), Requirements maturity index (M10), problem per user month (M11) and rtf curve (M12) |

Table 6: Inappropriate metrics resulted

4.3.2 Potential metrics and prominently being used metrics analysis (RQ1)

The identified metrics were in relation to the respondents’ satisfaction with using them. The analysis is done here separately for satisfied and unsatisfied respondents:

a. Prominent metrics from satisfied respondents.

Prominent metrics were considered when responses over the metric is equal to or above the mean value (17) and metrics that were highlighted by satisfied respondents in section 4.6.2.

b. Prominently used metrics by unsatisfied respondents.

Prominently used metrics were considered by when responses over the metric is equal to or above the mean value (5) and metrics that were highlighted by unsatisfied respondents in section 4.6.1.

Here it is important to note that prominent term is not used since the responses in this category are unsatisfied with metrics, so metrics resulted can’t be prominent instead prominently used metrics can found. This observation is important to know whether prominent metrics resulted from satisfied respondents overlapped with prominently used metrics by unsatisfied respondents. This will leads to further investigation of why respondents are unsatisfied with the outcome of metrics use even though prominent metrics are used, detailed in section 4.6.

c. Comparisons of prominent metrics from satisfied respondents and prominently used metrics from unsatisfied respondents.

a. Prominent metrics from satisfied respondents

Prominent metrics computation:

Prominent metrics = metrics obtaining at least mean value (17) of satisfied responses or above (see figure 12) + highlighted by satisfied respondents (in section 4.6.2)

In total, 11 prominently being used metrics found from satisfied respondents. Metrics that reached the mean value (16.5–17) of total satisfied responses (35) are work
in progress (WIP) (M6), story point (M13), burndown chart (M29), product backlog (M14), velocity (M2) and CFD (M28). All these metrics are also highlighted by respondents that how they are useful (see section 4.6.2).

Moreover 5 other metrics are highlighted by satisfied respondents that these metrics very useful. Thus these metrics are considered as prominent metrics even though number of responses is slightly less than mean value (see figure 12). These metrics are: Cost (M21), cycle time (M26), lead time (M4), Queue (M5) and VSM (M22). For more details for usefulness of these metrics shared by satisfied respondents is explained in the section 4.6.2.

There are other 9 metrics from category 3 and 4 (see figure 12) didn’t reach the mean value but still dominated by the satisfied responses (see table 8 below). This seems to indicate that these metrics are useful but need more awareness to the industries.

<table>
<thead>
<tr>
<th>Total satisfied Respondents</th>
<th>Prominently being used metrics by many of the surveyed respondents(11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>work in progress (WIP) (M6), story point (M13), burndown chart (M29) and product backlog (M14), velocity (M2), Cost (M21), cycle time (M26), lead time (M4) and CFD (M28), Queue (M5) and VSM (M22).</td>
</tr>
</tbody>
</table>

Table 7: Prominently being used metrics from satisfied respondents in large scale

<table>
<thead>
<tr>
<th>Total satisfied Respondents</th>
<th>Metrics that are not much being used among surveyed respondents (9)</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>Function Points (M3) Story rate per iteration (M15), Maintenance flow (M16), Workload (M17), Combined analysis (M18), Linear regression (M19), Estimation error (M20), Defect state over time (M23), Variance (M24)</td>
</tr>
</tbody>
</table>

Table 8: potential Metrics needs more awareness to industries

b. Metrics that are mostly being used by unsatisfied respondents

In total, 19 respondents out of 56 respondents were not satisfied with the metrics (see Figure 11) they are using. 11 metrics were found mostly being used (see figure 12) and highlighted by unsatisfied respondents (see section 4.6.1). This is considered when the metrics which has mean value of unsatisfied responses (5) or above. The metrics resulted mostly being used among surveyed unsatisfied respondents are: Velocity (M2), function points (M3), wip (M6), product backlog (M14), workload (M17), estimation error (M20), cost (M21), defect state over time (M23), story point (M13), rework rate (M27) and burndown (M29).

<table>
<thead>
<tr>
<th>Number of total unsatisfied Respondents</th>
<th>Prominent metrics (11)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>Velocity (M2), function points (M3), wip (M6), product backlog (M14), workload (M17), estimation error (M20), cost (M21),</td>
</tr>
</tbody>
</table>
Table 9: metrics that are mostly being used by unsatisfied respondents

c. Comparisons of metrics being used by satisfied and unsatisfied respondents

When compared with the prominently being used metrics by satisfied respondents and being used by unsatisfied respondents (see Tables 7 and 9), 6 metrics were found overlapped and other 5 metrics that are prominently used by unsatisfied metrics which are certainly not much used by satisfied respondents (see in Figure 13). Question can be raised here that 6 metrics that are found prominent if these metrics were also used by unsatisfied respondents in a large number, then why these respondents are unsatisfied with the outcome of metrics use. This seems to indicate that respondents fail to use the metrics to the extent they should and could, since evidence shows that these metrics are prominently used or very useful for Agile and Lean settings (see table 7). Further investigation is done to comparing the experiences shared by satisfied and unsatisfied respondents over the use of these metrics in section 4.6.

Figure 13: Comparisons on metrics being used by satisfied and unsatisfied respondents
4.4 For which purpose the identified metrics are used (RQ5)

In this questionnaire, respondents were asked the purpose of using the metrics they mentioned being used in their organization (see question 4 in Appendix B). The options given were time, quality and both. In this respective question it was clearly mentioned that metrics that belongs to time related metrics that targets or enables predictability, forecasting, planning or even metrics that’s tracking. Quality related metrics are the metrics which targets quality directly or inventory or can be even testing. For more information about quality and time focused metrics see literature section 3.2.5.

But it can be possible that in software industry there can be multiple metrics used for both purposes; here it means quality as well as time, so another option ‘both quality and time’ is provided. So, respondents were asked to choose only one option in the predefined list, in addition a text box named ‘others’ is provided. This allows respondents to specify if something is missing in the predefined options. (see question 4 in Appendix B)

Since the respondent allowed choosing only one option, the results of the questions are computed based on the number of respondents received for each category and not over the number of responses.

It was observed (in Figure 14) that more number of respondents answered time which is 29 (52%), followed by quality which is 19 (34%) and finally number of respondents answered for both process aspects is 8 (14%). In other category, 9 answers were provided; where 5 of them mentioned “reporting, capacity management (resource utilization), risks, predictability, forecasting and planning” which falls in time related process aspect, 1 of them mentioned quality and remaining 3 responses were ignored due to irrelevance information. However, the author moved 8 of them into their respective categories.

Here it is important to note that more number of respondents is focused to time related metrics which certainly means metrics that enables good predictions, forecasting,
tracking or planning are more important rather than metrics that are directed to quality. This indicates that the intention of driving the process metrics in Agile or Lean industries is more for predictability (time) process aspect compared to driving the metrics for quality.

4.5 Metrics are categorized with their focus to measure time (predictability, forecasting, planning or tracking) and quality aspects

Satisfied responses analysis:

<table>
<thead>
<tr>
<th>Category_ID’s Metrics</th>
<th>&quot;Time (T) vs. &quot; Quality (Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>work in progress (WIP)</td>
<td>30</td>
</tr>
<tr>
<td>product backlog</td>
<td>24</td>
</tr>
<tr>
<td>story point</td>
<td>18</td>
</tr>
</tbody>
</table>

2

<table>
<thead>
<tr>
<th>Metrics</th>
<th>&quot;Time (T) vs. &quot; Quality (Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velocity</td>
<td>17</td>
</tr>
<tr>
<td>lead time</td>
<td>12</td>
</tr>
<tr>
<td>cycle time</td>
<td>10</td>
</tr>
<tr>
<td>Cost</td>
<td>7</td>
</tr>
</tbody>
</table>

3

<table>
<thead>
<tr>
<th>Metrics</th>
<th>&quot;Time (T) vs. &quot; Quality (Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSM</td>
<td>10</td>
</tr>
<tr>
<td>defect state over time</td>
<td>8</td>
</tr>
<tr>
<td>Queue</td>
<td>8</td>
</tr>
<tr>
<td>Estimation error</td>
<td>6</td>
</tr>
<tr>
<td>story rate per iteration</td>
<td>6</td>
</tr>
<tr>
<td>Function point</td>
<td>5</td>
</tr>
<tr>
<td>workload</td>
<td>3</td>
</tr>
<tr>
<td>Linear regression</td>
<td>2</td>
</tr>
</tbody>
</table>
It was observed that all metrics except cost (M21) in category 1 and 2, more responses received for time associated purpose i.e., for predictability, forecasting or planning and for cost (M21) more responses received for quality (see figure 17).

In category 3, three metrics; vsm (M22), defect state over time (M23) and queue (M5) received more responses for time than quality. Metrics such as estimation error (M20), story rate per iteration (M15) and function point (M23), received equal number of responses who are using for time and quality. Only two metrics in category 3, which are workload (M17) and linear regression (M19), are more used for quality compared to responses for time (see figure 17).

In category 4, except variance (M24) rest of all metrics were used for the quality related process aspect than time. Since variance got equal number of responses who used for time and quality (see figure 17).

In Conclusion, overall 9 metrics were more used for time related process aspects, 6 metrics (M21, M17, M19, M25, M18, M16) were more used for quality purpose and other 4 metrics (M20, 15, M3, M24) has the same number of responses for time and quality.

4.6 Experiences shared by satisfied and unsatisfied respondents

4.6.1 Causes Mentioned by unsatisfied respondents

In this section reasons for dissatisfaction over use of metrics mentioned among some of the surveyed respondents will be discussed. In the section 4.2.2, number of satisfied and dissatisfied respondents were analyzed but respondents who answered ‘no’ which means respondents who are not satisfied was also asked the reason why? This reasons will be first discussed here. (see question 5 in Appendix B)

Only 8 respondents out 19 unsatisfied respondents mentioned the reasons why they are not satisfied with the metrics provided in other category.

Respondents who are not satisfied with the usage of above metrics, mentioned the following reasons;
- More precision needed
- Metrics don’t give a correct picture sometimes
- Projections don’t add up sometimes
One of the respondents (project manager) explained more in detail the reason is that it’s difficult comparing the data between the projects and threshold cannot be the same. He also mentioned that even though they have sprint metrics like velocity, the challenging part is that they need to harmonize the project level scope, cost and schedule variance, completion rate to stabilize the system.

Another respondent (project manager) explained the reason that metrics provides some value but they are not definitive especially function points. He stressed, Function points seemed to be reasonable approach to measure requirements up front projects especially when they come from government in a contract bid. Other metrics, Burndown chart helps to see the true amount of work left and Work in progress helps them to see how much work team members have and react if they take too much at once.

Few of them mention that it is due to sub optimization, connected to incentives. For example, one of the respondents (Agile mentor) mentioned that resource utilization rate has negative impact on Lean thinking which leads to sub optimization.

They also claim that processes are not in a state that metrics can be applied in meaningful level.

One of the respondent stressed that management does not know Lean thinking and related metrics, for example, he say that’s throughput time does not seems relevant for Lean.

In the questionnaire, respondents were also asked for their experiences of these metrics. All the unsatisfied respondents’ answers were categorized here. In total 10 out of 19 respondents shared their experiences of using the metrics in specific. Interestingly it was found except one respondent, remaining all highlighted the metrics which are important and useful for them.

Metrics that are highlighted by these respondents:

- For multi-project management, workload (M17) and cost (M21) are the metrics that are tracked throughout the process.
- Burndown chart (M29) helps to see the true amount of work left, one of them stressed that Cumulative flow diagram (M28) for projection of completing projects is the most important.
- Work in progress (M6) helps them to see how much work team members have and react, if they take too much at once.
- Story point (M13) is the simplest and best matrix to track.
- Other metrics are velocity (M2), cost (M21) and product backlog (M14) are very useful.
- Resource utilization rate has negative impact on Lean thinking.

Author observed that from the experiences written by unsatisfied respondents (see below bullet points) and reasons why they are not satisfied with metrics (see above bullet points), respondents shared their reasons in general but not specific to what problem in metrics that makes them disappointed with the metrics use. For example one of the respondent mentioned (see above bullet points) metrics don’t give up a correct picture sometimes instead of specifying which metrics exactly doesn’t add up. On the other hand, they have shared the experiences of metrics in specific which are useful and important to them instead of sharing their experiences on metrics they are not satisfied. This would seem to indicate that respondents are not aware of exactly which metrics are not performing well and where the problem lies. It can also be seen that all the respondents know what they want from metrics. Thus the problem seem to be either in using the process metrics to the extent in should and could or lack of awareness of which metrics to be used for a specific situation. Further interpretation is explained in section
4.6.3 with the data collected by comparing and analyzing with the experiences from the satisfied respondents (reported in section 4.6.2) over the these metrics use.

4.6.2 Experiences of satisfied respondents on metrics useful

Observations made from the experiences of the respondents using the concerned metrics, in addition comparisons between metrics was also mentioned where relevant.

One of the question was asked to respondents to know the importance and real experiences of metrics that are used by practitioners in the industries. (see question 2 in Appendix B)

In order to avoid irrelevant or false data, convenient sampling is followed as mentioned in section 2.5 by allowing respondent’s option to answer this question, since their interest to share their experiences regarding metrics that can certainly enables to share the right information.

In total 37 respondents answered this question where 5 respondents were terminated as they provide no valid information in relevance to the question.

**Observations on useful metrics that are highlighted by satisfied respondents:**

**Visualization measures or indicators:** burndown charts (M29) and cumulative diagrams (M28), which is more useful?

**Burndown charts (M29):**
Many survey respondents mentioned that burndown charts is one of the very useful measure that keeps everyone engaged and stressed that it’s very useful for their work. It is very important for visualization to give insights or estimations of true amount of work left and progress of team. Early identification of schedule delays can be easily detected by burndown charts where appropriate actions can be taken before it’s not too late.

**Cumulative Flow Diagrams (M28):**
Some of the Scrum masters mentioned that a Cumulative flow diagram is the very important for projection of completing projects. Few product owners mentioned that this metric or indicator is very useful for checking the lead time of maintenance issues.

Interestingly one of the Agile team leader mentioned that burn up chart is much more useful than burn down charts because it shows moving goal posts, while the latter obscures them. This is a new metric found apart from the identified list and its effectiveness compared burn down is very promising as mentioned by Agile team leader.

It is important to note that, not only respondents mentioning burndown charts are more useful when compared to cumulative diagrams, it is also observed from frequency of responses analysis between these two metrics in Figure 12 that more responses (26) were using burndown charts when compared to cumulative flow diagrams (17).

**Velocity (M2) or story points per sprint (M13):** Everything about velocity is observed, in specific various behaviors of velocity is observed when story points are used with different units (per sprints or per person)

Velocity represents a team’s rate of progress over a period of time by converting story points to calendar days or by measuring story points per iteration.
Many respondents mentioned that Velocity which is story points per sprint is the main metric for the projects which gives insights in the progress. This helps to analyze effectiveness of the teams by giving insights of team’s progress and for estimations on release planning. Few respondents mentioned that velocity helps the team make sprint commitments and highlighted that it is the simplest and best matrix to track.

Another important thing which is highlighted and mentioned by an Agile team leader about story point metric is that they typically use story points as a way to calculate velocity as an absolute value in both for points per week (sprint) and to calculate points per person day. Also stressed that Points per person day is the most effective way to ensure accuracy especially in the summer and during holidays when half of the team is out. Here is we can highlight that fact that story points is not standardized and it is used not only for story points but also to helps to analyze the performance or progress of individual persons.

Many Agile team leaders/mentors and Scrum masters mentioned that Project velocity is very important as measures of relative value which drives priority, which is very useful. In other terms, Story points are often used for prioritization.

**Cycle time (M26), Work in progress (M6) and Throughput (M2'): strong relation was observed between these metrics**

Many respondents mentioned that Work in progress helps them to see how much work team members have and react if they take too much at once. One of the senior Agile team developer mentioned that Work in progress and cycle time are good to get us finishing things sooner. Some of the Agile mentor’s stressed WIP and cycle time plays an important role in kanban environments and shared their experience that WIP, cycle time and throughput provides high transparency into the team progress and helps them to suggest the particular team interventions where needed for improving the overall performance. One of the product owner also mentioned that throughput is becoming more important to forecasting than velocity. It is very important to note here that throughput is subtly different than velocity. Few others mentioned that Throughput is becoming more important to forecasting than velocity.

Here It’s important to note that senior industry practitioners in Agile or Scrum or kanban reclaims the fact of the Little law’s by using these three metrics. They highlighted that these metrics are much more efficient than traditional Agile metrics like velocity, story points.

Little’s law,  
Cycle time = WIP /throughput

One of the Agile mentor stressed that after they have moved to more actionable flow metrics like cycle time, wip and throughput from traditional metrics like velocity, story points etc which they were using for several years for software solutions in healthcare domain. Finally they could able to achieve remarkable reduction in cycle time and a very significant improvement in operational efficiency.

**Lead time (M4)/lead time distribution(s), Vsm (M22): strong relation**

Few Agile mentors mentioned Lead time distribution is very useful in pull decisions and forecasting models. One of them also stressed that these metrics (Lead time, vsm) along with cycle time are used for the purpose of releasing.

**Function points (M3), product backlog (M14) and variance (M24):**

Respondents mentioned that function points seemed to be reasonable approach to measure requirements up front projects especially when they come from government in a contract bid

They stressed that it is used with historical project productivity data for project portfolio management.
Other metrics product backlog and variance was just mentioned why are they used for example: Product backlog as requirements pool, Variances for estimation exactness.

**New metrics that are highlighted:**

Interestingly a new metric was found which is unexpected issues per sprint and mentioned by one of the Scrum master. He shares his experience that velocity and unexpected issues per sprint are very important to their team as they are dealing with too many unexpected issues and bugs.

One of the project manager mentioned that they use ROI (return on investment - sometimes measured in relative points sometimes in dollars), TTM (time to market - measured similar to lead time, from concept or request to delivery to users or customers). But ROI is ignored since it doesn’t seem to be relevant to process metrics.

**Observations:**

Metrics highlighted by satisfied respondents that they are very useful
In total 10 metrics are observed which are mostly highlighted by the satisfied respondents who are industrial experts since most of them have senior roles. These 10 metrics are: burndown chart (M29), cumulative flow diagram (M28), velocity (M2)/story points per iteration (M15), lead time (M4), queue (M5), throughput (M2’), work in progress (M6), cycle time (M26), function points (M3) and product backlog (M14).

**Some relationships among metrics observed:**

Moreover some significant relationships among these metrics and behaviors of some metrics in Agile/Lean reported in this section observed are:

- Lead time (M4) distribution not only useful in pull decisions and forecasting models but also when useful for right predictions i.e. purpose of releasing is more accurate when used with cycle time.

- Function points (M3) are used as a reasonable approach to measure requirements up front projects especially when they come from government. Moreover when it used with historical project productivity data it gives more value for project portfolio management.

- Industrial experts highlighted that WIP (M6), cycle time (M26) and throughput (M2) are powerful metrics in Agile or Lean industries when used together provides high transparency into the team progress and helps them to suggest the particular team interventions where needed for improving the overall performance. Moreover there is subtle difference observed between velocity and throughput, where mentioned that throughput is more actionable and promising in Agile/Lean than velocity.

- It was mentioned by few experts that story point (M13) typically used as a way to calculate velocity (M2) as an absolute value in both ways: for points per week (sprint) and to calculate points per person day. Also stressed that Points per person day is the most effective way to ensure accuracy to ensure accuracy especially in the summer and during holidays when half of the team is out.

- When visualization metrics or metrics as indicators which are burndown chart and cumulative flow diagram, from industrial experts experiences and from analysis (see figure 12) that burndown chart (M29) is more useful than cumulative flow diagram (M28). The reason it was mentioned by respondents’ experiences regarding burndown is that it is very important for visualization to give insights or estimations.
of true amount of work left and progress of team. Early identification of schedule delays can be easily detected by burndown charts where appropriate actions can be taken before it’s not too late.

4.6.3 Causes for dissatisfaction on metrics usage over process improvements (RQ6)

(Does the dis-satisfaction of respondents with the outcome of the metrics is due to:

Fail to use the metrics properly or due to use of inefficient metrics which fails to increase agility)

It is clearly observed in figure 12 that inappropriate or irrelevant metrics for Agile or Lean being used by respondents who responded not satisfied with the metrics being used. This indicates that lack of awareness of appropriate metrics that are useful for Agile and Lean settings is one of the reason for not being satisfied with the metrics being used by respondents.

Another observation made in section 4.3.2 was that even though there are set of potential metrics being used by unsatisfied respondents, they answered they are not satisfied with these metrics. However, here it is important to note that using a metric alone doesn’t gives required value or right picture which is observed from experiences shared by respondents who are satisfied (see section 4.6.2), since from the observation in section 4.6.1 shows that all the respondents who are not satisfied with metrics doesn’t mentioned any combination of metrics even though these metrics have the ability to increase agility when compared with satisfied respondents using the same metrics in set of combinations ( see in section 4.6.2 ) Here combination or set of metrics using together means having important relationships among them (see in section 4.6.2) where behavior of one metric influence the behaviors of other metrics to increase the operational performance. This seems to indicate it is important to be aware which metrics are relevant for Agile or Lean and have deep understanding of these metrics behavior in industrial environments practicing Agile or Lean practices. This will help to realize the level of predictability these metric’s promise in order to make right assumptions or planning. May be this combinations may vary from situation to situation in Agile and this has to be investigated and left for future work.
5 FURTHER COMPARATIVE ANALYSIS BETWEEN SLR AND SURVEY

5.1 Potential metrics resulted from Literature and prominently being used in industries

<table>
<thead>
<tr>
<th>Metrics resulted</th>
<th>Literature</th>
<th>Survey</th>
<th>Overlapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metrics resulted</td>
<td>29</td>
<td>30</td>
<td>29</td>
</tr>
</tbody>
</table>

Table 10: Number of metrics overlapped

From Literature review, 20 potentially useful metrics found out of 29 resulted metrics in literature (See Table 2). Here potentially useful metrics refers to the metrics identified from the empirical evidence through rigor and relevance analysis (see section 3.2.3).

Out of these 20 potential metrics, 11 metrics are found prominently being used in large scale among surveyed industries and other 9 metrics need more awareness in industries (see section 4.3). Here prominence is identified based on analysis on frequency of responses over metrics and metrics that are repeatedly highlighted by satisfied respondents (see section 4.3.2).

The 11 prominently being used metrics are: M2, M4, M5, M6, M13, M14, M21, M22, M27, M28 and M29 (See table 9 below). Other 9 metrics are: M3, M15, M16, M17, M18, M19, M20, M23 and M24 (see table 11 below) which are resulted from empirical studies are found to be useful and shows the potential benefits in Agile/Lean settings but still more awareness needed for practitioners to make them useful in large scale, since the evidence from survey analysis shows that these metrics has low number of satisfied responses when compared to prominent metrics responses (see section 4.3).

Table 11 below shows the comparison of potential metrics found in literature and through industrial survey. In table below (✓) shows the presence of that metric and (X) represents the absence of that metric.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Survey</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velocity/throughput (M2)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Lead time (M4)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Queue (M5)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>WIP (M6)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Product Backlog (M14)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cost (M21)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>VSM (M22)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Story point (M13)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cycle time (M26)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Cumulative Flow Diagram (M28)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Burndown Chart (M29)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Function Points (M3)</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
5.2 For which purpose the identified metrics are used (RQ5)

Here time aspects of the process means metrics enables predictability, forecasting, tracking and planning to enforce process improvements (see section 3.2.5). 10 metrics out of 11 prominently being used metrics (M2, M4, M5, M6, M13, M14, M22, M27, M28 and M29) identified from both literature and industrial survey were used to measure the time aspects of the process. Only 1 metric (M21) found to be targeted directly to measure quality aspect of the process (See Table 10 below). We can observe (in Table 12 below) that other 9 metrics has distinct results from survey and literature in the purpose of use. Since the number of responses are less for these metrics, literature evidence is considered. However with the obtained results it can be understood that more evidence is found for metrics used for time aspects and less evidence was found for metrics used for quality.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Survey</th>
<th>Literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>Quality</td>
<td>Time</td>
</tr>
<tr>
<td>Velocity/throughput (M2)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Lead time (M4)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Queue (M5)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>WIP (M6)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Product Backlog (M14)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Cost (M21)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>VSM (M22)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Story point (M13)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Cycle time (M26)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Cumulative Flow Diagram (M28)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Burndown Chart (M29)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Function Points (M3)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Story rate per iteration (M15)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Maintenance flow (M16)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Workload (M17)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Combined analysis (M18)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Linear regression (M19)</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Purpose of metrics being used</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------------</td>
<td></td>
</tr>
<tr>
<td>Estimation error (M20)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defect state over time</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>(M23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variance (M24)</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 12: Purpose of metrics being used
6 CONCLUSION

6.1 DISCUSSION

In this research study author explored various metrics that are potentially useful for Agile or Lean settings in current state of art and currently being used by industries. Moreover author studied the reasons why some of the respondents among survey were not satisfied with the outcome of the metrics being used.

This study was performed in three sequential steps: first all metrics that are reported and empirically evaluated are found through systematic literature review using snowballing. In the next step, the literature findings were verified with the industries by conducting industrial survey. Moreover, the causes for dissatisfaction of some of the surveyed respondents over metrics being used are investigated by comparing and analyzing the results with satisfied respondents.

Answers to Research Questions:

RQ1. Which of the identified metrics are prominently being used in industries?

In total 20 metrics found potentially useful for Agile or Lean development practicing industries which are collected through systematic literature review and industrial survey (see section 5.1). Out of these 20 metrics, 11 metrics (see section 4.3.2) were found prominently being used in large scale among surveyed industries and, for other 9 metrics evidence from both survey (see section 4.3.2) and literature (see section 3.2.4) shows that these metrics provides potential benefits in Agile/Lean settings but still more awareness needed for practitioners to understand the potential usefulness of these metrics and use in large scale.

RQ1.1. Which metrics are found with empirical evidence in literature?

Metrics that are relevant and useful for Agile or Lean software development are found from literature through Systematic Literature Review. After applying inclusion/exclusion criteria 18 research articles were selected where resulted 29 metrics from these articles. After applying study quality assessment through rigor and relevance analysis 20 metrics were resulted from 12 empirical research articles where 11 of these articles has high rigor and high relevance score.

RQ2. For what purpose the identified metrics are used in industries?

Evidence from both SLR (see section 3.2.5) and survey (see section 4.5) shows that metrics that are used for time related purpose which enables predictability, tracking, forecasting or planning are more dominant than metrics that are directed to quality or inventory purpose by industries. On the other hand, this also indicates not only industries are focusing on time associated metrics in Agile or Lean compared to quality focused metrics but also shows that time associated metrics are very important for an evolutionary environment like Agile or Lean.

RQ3. What are the causes for dissatisfaction over metrics being used among some of the surveyed industries?

In this thesis causes were investigated on why some of the surveyed respondents were not satisfied with the metrics being used by comparing with the data collected with the other category of respondents who are satisfied with the metrics. First observation was very straight that metrics which are not relevant or inappropriate for Agile or Lean were used by the respondents who are not satisfied (see figure 12). This shows that some of the industries are not much aware of which metrics are applicable for Agile or Lean development practices. Another important observation which lead to further investigation was that even though potential metrics or metrics that are relevant to Agile or Lean were used, respondents answered that they are not satisfied with these metrics (see section 4.3.2). Thus it was observed from the experiences shared by unsatisfied
respondents that similar set or combination of metrics were used by satisfied respondents which is observed missing by the respondents answered not satisfied with these metrics (see section 4.6). Thus it can said that for a specific measurement, using right combination of metrics in Agile and Lean plays a major role to get the right picture or for continuous improvements.

This indicates that it is important to be aware of potential metrics applicable for Agile or Lean and have deep understanding of these metrics behavior when used in industrial environments practicing Agile or Lean methods. This will help to realize the level of predictability these metric’s promise in order to make right assumptions or planning.

6.2 Validity Threats

Validity threats are discussed in this section and is important to judge the reliability of the results. The main threats for SLR in this study is related to search and researcher bias. The main threats for Industrial survey will be discussed which are in relation with the research design and data collection phases. Based on the classification proposed by Wohlin et al [52], four perspectives of validity is discussed.

6.2.1 Construct validity

In SLR, construct refers to the presence of potential factors or information relevant to metrics whether or not this study able to capture in relevance with aims and objectives. This was minimized by detailing the sections and structuring according to research questions. The threat for using snowballing as a search approach for systematic literature review is obtaining a good start set. Using the characteristics of obtaining a good start set suggested by Wohlin [42], start set design is extended with the change in database selection which mitigates the risk of obtaining too much of irrelevant studies. Moreover same authors are excluded when selection of start set of papers which mitigates the risk of resulting same authors in the start set (see section 2.1).

Construction threat due to bad questionnaire design leads to irrelevant or inappropriate answers to the research questions. This is a major concern in any survey, since the questions need to be constructed or designed that all participants will be able to understand and give the information as accurate as possible to achieve research goals. In order to reduce the potential of this problem, by conducting detailed discussion sessions and revised by the supervisor. Moreover, a pilot study is conducted and published in some of the professional social networks (www.linkedin.com, www.xing.com etc.). The feedback from supervisor and the pilot study analysis was taken into consideration when improving the questionnaire.

6.2.2 Internal validity

Internal validity seems to be a major challenge for literature review part of this study, since the study selection is done by an individual researcher. This is avoided by being as inclusive as possible and verifying with the supervisor wherever in doubt. Another threat can be when deciding whether or not which articles to include or exclude, because finding the metrics related to process improvements and in association with requirements in a study is challenging which is not very straightforward in most of the literature which can leads to the risk of overlooking or skipping useful articles. This risk is mitigated by the involvement of supervisor reviewing on each and every iterations of obtaining start set of papers, and in also every iterations of snowballing procedure, since his experience in this domain adds more value to minimize researcher bias. The internal validity is further enhanced by strictly following snowballing procedure guidelines [42] and guidelines for quality assessment criteria [47].
The selection bias threat in industrial survey is minimized, since the respondents who answered are from different fields in the particular study (Scrum, Agile, Kanban, Lean, XP etc.) and major of the respondents are from high level companies (>500 employees). Moreover these respondents who answered can be satisfied or unsatisfied with the outcome of metrics use, where results can be biased. This is mitigated by separating satisfied respondents and unsatisfied respondents results, and analyzed accordingly.

6.2.3 External validity

External validity refers to the ability to generalize the results to different situations, settings, and groups. The majority of studies resulted in primary studies are case studies (11 out of 18) with high rigor and relevance, see Appendix A. Which gives the results more ability to generalize and relevant to an industrial context. The starts set for snowballing was composed from a database search further increases its external validity.

For the industrial survey, to answer the questionnaire which is related usage of process metrics and regarding requirements respondents need to major or senior roles to obtain more appropriate answers otherwise there is a huge threat that results can be biased. This is mitigated by targeting senior roles and resulting in major part of respondents. The major part of respondents resulted are Scrum masters, Agile mentors, product owners or managers, team leaders and requirements engineers working in Agile or Lean software industries (See Section 4.4.1). This gives results more credibility for generalizing.

6.2.4 Reliability

In order to enhance the reliability of this study, various practices was followed.

• First, In order to avoid the irrelevant studies and maintain consistency in finding the start set of papers for snowballing, ‘engineering village’ database was used which is suggested for an excellent search for engineering instead of google scholar which lacks capturing the scope (it finds too much of information) [43]. Then, to overcome any chance for missing research work and better coverage, manual search was performed on gray literature, for example, google scholar. This action was just performed to cross verify with the primary studies resulted. To further substantiate the search process, repetition of authors in selecting start set of studies is avoided because authors the respective expertise field includes refers their relevant papers which could be found through snowballing.

• Since the keywords “Agile”, “metrics” and “requirements” are very generic, a lot of metrics resulted in relevance as well as non-relevance with requirements, for example requirements terminology is used in the literature but not in relevance with metrics. This is also mitigated by proposing all the identified metrics in survey and cross verified to give the strength to the findings.

• Second, the quality of the identified studies is a major concern in this study since it focuses on both industry practice and literature. Quality is ensured in three stages: 1) by focusing on peer-reviewed articles are only included in selection of primary studies. 2) By classifying the identified studies are classified into research types and 3) by performing quality assessment criteria analysis.

6.3 FUTURE WORK
Future work is planned to focus on several avenues directly related to the results of this study. Firstly, we plan to conduct a wider survey with more respondents that allows statistical analysis and therefore brings significant to our results. Moreover, we plan to further explore the role of requirements-associated Agile and Lean metrics in providing satisfaction of the metrics activities. Another interesting perspective of future work is to conduct in-depth interviews with practitioners’ that allow or detailed insights on which metrics provide the best satisfaction and when. Moreover, we plan to further explore the Agile and Lean projects characteristics where these metrics were or are applied, through industrial case study. Finally, we plan to create a recommendation framework for selecting optimal Agile and Lean metrics that cover requirements associated processes and help to optimize the amount of requirements work in relation to the desired process efficiency.
REFERENCES


## APPENDIX A: RIGOR AND RELEVANCE SCORES

<table>
<thead>
<tr>
<th>Study_ID</th>
<th>Context</th>
<th>Design</th>
<th>Validity</th>
<th>Rigor Sum</th>
<th>User/Subject</th>
<th>Scale</th>
<th>Research Methodology</th>
<th>Context</th>
<th>Relevance Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.5</td>
<td>1</td>
<td>0</td>
<td>1.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>C2</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>C3</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C9</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>2.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C11</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C13</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C14</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>2.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C16</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C17</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>2.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>C18</td>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>1.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>
9 **APPENDIX B: QUESTIONNAIRE**

The metrics and measurements that enforce to Agile and Lean process improvements are increasing. Therefore increasing the challenge in selection of appropriate metrics or measurements that contributes to the process flow. So, we systematically extracted all the metrics and measurements referring to requirements in agile and lean context. Finally, we found two categories of measurements: metrics that needs empirical evidence and metrics that needs to extend the knowledge of empirical evidence, which are already evaluated many years back but failed to update the utilization or efficiency. Thus, one goal of this Questionnaire is to evaluate and extend the empirical knowledge of the found list of metrics to determine the useful metrics and measurements for Agile/Lean. Second, to investigate if there is any relation associated with requirements and metrics exists.

The survey should take about 5-7 minutes of your time. Please answer how YOU perceive the situation to be today, NOT how it should or could be.

**DISCLAIMER:** all the information will be kept confidential and no traceability to individuals will ever be possible.

**Definition:**

Process metrics are defined as metrics that are an invaluable tool for industries wanting to monitor, evaluate and improve their operational performance.

1. Which metrics do you use for process improvements in Agile/Lean? Please specify in 'other’ if you don’t find the metrics you use in the following list. (Choose one or several)
2. From the metrics you used please explain your experiences and highlight which metrics you are talking about:

3. On which of the following method have you usually use the metrics you mentioned? (choose one)
   a) Agile
   b) Lean
   c) Both

4. For which of the following purpose do you use the metric? (choose one) (Note: Metrics that used for Visualization/tracking/monitoring/forecasting comes under time related purpose and metrics that are only intended for quality comes under Quality focused)
   a) Time related
   b) Quality focused
   Other (please specify) ……………

5. Are you satisfied with the outcome of the metrics you are using over process improvements? (Choose one)
   a) Yes
   b) No
   If No, Please specify the reason ……………

6. What is your main role in your organization? (Choose one or several)
   a) Scrum Master
   b) Product Owner
   c) Project manager
   d) Agile Team Leader
   e) Agile mentor
   f) Agile team developer
   g) Scrum team developer
   h) Kanban team developer
   i) Independent tester
   j) Requirements engineer
k) Software tester or test designer
Others ..................

7. What is the size of your organization? (Choose one)
   a) Less than 10 employees
   b) 11-20 employees
   c) 21-50 employees
   d) 51-100 employees
   e) 101-200 employees
   f) 201-500 employees
   g) Over 501 employees