Software Process Improvement
Measurement and Evaluation Framework
(SPI-MEF)
A Framework for Evaluating the Outcome of SPI Initiatives

A.K.M. Moinul Islam and Michael Unterkalmsteiner
This thesis is submitted to the School of Engineering 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 40 weeks of full time studies.

Contact Information:
Author(s):
A.K.M. Moinul Islam
Address: Folkparksvägen 19:20, SE-37240 Ronneby, Sweden
E-mail: moinul@gmail.com

Author(s):
Michael Unterkalmsteiner
Address: Folkparksvägen 12:25, SE-37240 Ronneby, Sweden
E-mail: miciu@gmx.de

University advisor(s):
Dr. Tony Gorschek
School of Engineering
Blekinge Institute of Technology
Box 520
SE – 372 25 Ronneby
Sweden

EMSE Co-supervisors:
Prof. Dieter Rombach (Technical University of Kaiserslautern)
Prof. Giancarlo Succi (Free University of Bolzano)

Internet : www.bth.se/tek
Phone : +46 457 38 50 00
Fax : +46 457 271 25
ABSTRACT

During the last decades, the dependency on software has increased. Many of today’s modern devices embed software to control their functions. The increasing dependency has also taken part in shaping the software development process to produce better quality software. Many researchers and practitioners have spent large investments to improve the software development process. A research area within software engineering that addresses the assessment and improvement issues in development processes is called Software Process Improvement (SPI). One of the essential aspects in software process improvement is measuring the outcome of the implemented changes. The measurement and evaluation of software process improvement provides the means for the organization to articulate the achievement level of their goals.

Although the importance of measuring and evaluating the outcome of software process improvement is paramount, there exist no common guidelines or systematic methods of measuring and evaluating the improvement. This condition evokes difficulties for practitioners to implement software process improvement measurement programs. This issue has raised the challenge to develop and implement an effective framework for measuring and evaluating the outcome of software process improvement initiatives.

This thesis presents a measurement and evaluation framework for software process improvement. SPI-MEF provides guidelines in the form of systematic steps to evaluate the outcome of software process improvement. The framework is based on key concepts which were elaborated in previous work.

In this thesis, a validation of SPI-MEF is also conducted by involving representatives from academia and industry. The validation is aimed to judge the frameworks’ usability, applicability and usefulness. Finally, a refinement of the framework is carried out based on the input from the validation.

Keywords: Software process improvement, Measurement and evaluation framework, Outcome of software process improvement, Concepts in measuring and evaluating software process improvement.
ACKNOWLEDGEMENT

This thesis would not have been possible without the sincere help and contributions from several people. We would like to use this opportunity for expressing our sincere gratitude to them.

First of all, we would like to extend our gratitude to our thesis supervisor Dr. Tony Gorschek for providing us valuable guidance and advice throughout the thesis. His patient guidance made our path to achieve the goals much more pleasant. We would also like to thank our EMSE co-supervisors Prof. Dieter Rombach from the Technical University of Kaiserslautern and Prof. Giancarlo Succi from the Free University of Bolzano for their support and advice. In addition, we are also thankful to the course supervisor of "Master Thesis in Software Engineering", Dr. Robert Feldt, for providing valuable advice in preparing the thesis proposal and writing the thesis report, and for his valuable suggestions at the commencement of the course which helped us to set our goals for the thesis appropriately.

One of the major parts of this thesis was the validation of the proposed framework. The validation would not have been possible without the active participation of the researchers in the Software Engineering Research Lab (SERL) at the Blekinge Institute of Technology (BTH), the researcher from the University of Kaiserslautern and several practitioners from industry. We would like to express our gratefulness to all involved researchers who provided their valuable time for the interview despite their busy schedules. We also like to thank all the industry practitioners who participated in the validation of the framework.

Of course, we are also thankful to the Blekinge Institute of Technology for giving us the opportunity to attend the Master Programme in Software Engineering. In particular, we would like to extend our gratitude to all the course supervisors we met during our studies as well as the staff at the International Office and in the administration. Moreover, we would also like to thank the library staff who helped us to get specific papers and books when we were unable to retrieve them on our own.

Last but not least, we are deeply grateful to our families and friends for being always beside us and inspiring us in all the challenges we met.
# CONTENTS

ABSTRACT ........................................................................................................................................... I

ACKNOWLEDGEMENT ........................................................................................................................ II

CONTENTS .......................................................................................................................................... III

TABLE OF FIGURES .......................................................................................................................... VI

LIST OF TABLES ................................................................................................................................ VII

1 INTRODUCTION ............................................................................................................................. 1
   1.1 AIMS AND OBJECTIVES ........................................................................................................ 1
   1.2 RESEARCH QUESTIONS ....................................................................................................... 2
   1.3 EXPECTED OUTCOMES ....................................................................................................... 2
   1.4 STRUCTURE OF THE THESIS ............................................................................................ 3
   1.5 TERMINOLOGY ..................................................................................................................... 4

2 BACKGROUND ................................................................................................................................ 5
   2.1 SOFTWARE PROCESS IMPROVEMENT ................................................................................... 5
      2.1.1 Software process concepts ............................................................................................. 5
      2.1.2 Software process improvement initiatives ...................................................................... 5
   2.2 INTRODUCTION TO SOFTWARE MEASUREMENT ............................................................... 6
      2.2.1 Software measurement .................................................................................................. 6
      2.2.2 Categorization of software measurement ..................................................................... 7
      2.2.3 Challenges in software measurement ............................................................................ 8
   2.3 MEASUREMENT FRAMEWORKS ............................................................................................ 8
      2.3.1 Goal-Question-Metric (GQM) ...................................................................................... 9
      2.3.2 Quality Function Deployment (QFD) .............................................................................. 10
      2.3.3 Practical Software Measurement (PSM) ....................................................................... 11
      2.3.4 Balanced Scorecard (BSC) ........................................................................................... 13
   2.4 EVALUATION OF SOFTWARE PROCESS IMPROVEMENT .................................................. 14
      2.4.1 Formal assessment .......................................................................................................... 14
      2.4.2 Actual benefit evaluation ............................................................................................... 14
      2.4.3 Confounding factors ....................................................................................................... 15
   2.5 PROPOSED EVALUATION MODEL ......................................................................................... 16
      2.5.1 Measurement levels ....................................................................................................... 16
      2.5.2 Evaluation viewpoints ................................................................................................... 17
      2.5.3 Evaluation area ............................................................................................................... 18
      2.5.4 The measurements ........................................................................................................ 18
      2.5.5 Evaluation methods ....................................................................................................... 18
      2.5.6 Time to evaluate ............................................................................................................. 19
      2.5.7 Holistic evaluation ......................................................................................................... 19

3 RELATED WORK ........................................................................................................................... 20

4 RESEARCH METHODOLOGY ........................................................................................................ 23
   4.1 BRIEF DESCRIPTION OF RESEARCH METHODOLOGIES ............................................. 23
      4.1.1 Conceptual analysis ....................................................................................................... 23
      4.1.2 Expert opinion ............................................................................................................... 24
         4.1.2.1 Interview .................................................................................................................. 24
         4.1.2.2 Self-administered questionnaire .............................................................................. 25
   4.2 FLOW CHART FOR THE RESEARCH METHODOLOGY ....................................................... 26
   4.3 MAPPING OF RESEARCH QUESTIONS AND RESEARCH METHODS .................................... 28
5 PROPOSED FRAMEWORK – SPI-MEF

5.1 INITIALIZATION STAGE

5.1.1 Gather SPI initiative information

5.1.2 Gap analysis of evaluation quality

5.2 DESIGN STAGE

5.2.1 Evaluation scoping

5.2.2 Determination of measures

5.2.2.1 Primary and complementary measures

5.2.2.2 Derivation of measures using GQM

5.2.2.3 Cross-examination

5.2.3 Selection of evaluation methods

5.2.3.1 Metrics baselining

5.2.3.2 Evaluation method

5.2.3.3 Confounding factors

5.2.4 Evaluation scheduling

5.3 IMPLEMENTATION STAGE

5.3.1 Establishment of a measurement program

5.3.2 Conduct evaluation

5.3.3 Holistic view creation

5.3.3.1 Holistic view components

5.3.3.2 Aggregation of multiple evaluation results

5.3.3.3 Presentation of the holistic view

5.3.3.4 Combined holistic view

5.4 EVOLUTION STAGE

5.4.1 Evaluation feedback analysis

5.5 SUMMARY

6 SCENARIO APPLYING SPI-MEF

6.1 BACKGROUND

6.1.1 Initialization – Gather SPI initiative information

6.1.2 Initialization – Gap analysis of evaluation quality

6.2 PHASE I

6.2.1 Design – Evaluation scoping

6.2.2 Design – Determination of measures

6.2.3 Design – Selection of evaluation methods

6.2.4 Design – Evaluation scheduling

6.2.5 Implementation – Establishment of measurement program

6.2.6 Implementation – Conduct evaluation

6.2.7 Implementation – Holistic view creation

6.2.8 Evolution – Evaluation feedback analysis

6.3 PHASE II

6.3.1 Design – Evaluation scoping

6.3.2 Design – Determination of measures

6.3.3 Design – Selection of evaluation methods

6.3.4 Design – Evaluation scheduling

6.3.5 Implementation – Establishment of measurement program

6.3.6 Implementation – Conduct evaluation

6.3.7 Implementation – Holistic view creation

6.3.8 Evaluation – Evaluation feedback analysis

6.4 PHASE III

6.4.1 Design stage

6.4.2 Implementation stage

6.4.3 Evolution stage

7 STATIC VALIDATION

7.1 VALIDATION SCOPING

7.2 VALIDATION DESIGN

7.2.1 Planning

7.2.2 Design

7.2.2.1 Interview design

7.2.2.2 Questionnaire design
TABLE OF FIGURES

Figure 1: Structure of the Thesis ....................................................................................... 3
Figure 2: The GQM Paradigm (inspired by [60]) ................................................................. 9
Figure 3: The four phases of GQM method (inspired by [61]) .............................................. 10
Figure 4: A simplified “House of Quality” structure (inspired by [66] and [67]) .............. 11
Figure 5: The PSM measurement process model (inspired by [69]) .................................... 12
Figure 6: The Balanced Scorecard (inspired by [70]) ......................................................... 13
Figure 7: Variables in experiments according to the definition in [77] .............................. 15
Figure 8: Evaluation model for the evaluation of SPI initiatives (inspired by [28]) .......... 16
Figure 9: Flow chart for the research methodology .......................................................... 26
Figure 10: SPI-MEF overview .......................................................................................... 29
Figure 11: SPI-MEF initialization stage ................................................................................ 31
Figure 12: Evaluation Opportunity Matrix Template ....................................................... 33
Figure 13: SPI-MEF design stage ..................................................................................... 35
Figure 14: Evaluation area (inspired by [28]) .................................................................. 36
Figure 15: Primary and complementary measures (inspired by [28]) ......................... 38
Figure 16: The famous iron triangle metaphor (inspired by [28] and [102]) ................. 39
Figure 17: SPI-MEF interfaces with GQM ........................................................................ 40
Figure 18: Confounding factors in the evaluation of SPI initiatives (adapted from [28]) ... 47
Figure 19: Illustration of lag factor and degradation factor ............................................. 51
Figure 20: SPI-MEF implementation stage ........................................................................ 53
Figure 21: Evaluation Result Template ............................................................................. 54
Figure 22: Interpretation of different evaluation results .................................................... 55
Figure 23: Impact rating scale .......................................................................................... 56
Figure 24: Aggregation of multiple evaluation results ...................................................... 58
Figure 25: Example of the holistic view presentation for the process level .................... 59
Figure 26: Interpretation of different holistic views presented in radar charts ............... 60
Figure 27: Example of the combined holistic view ............................................................ 61
Figure 28: SPI-MEF evolution stage .................................................................................. 63
Figure 29: Software Process Improvement Measurement and Evaluation Framework (SPI-MEF) ........................................................................................................................... 66
Figure 30: SPI-MEF stages applied in scenario phases ..................................................... 69
Figure 31: Evaluation Opportunity Matrix for one solution ............................................. 71
Figure 32: GQM tree for phase I ...................................................................................... 77
Figure 33: Evaluation scheduling for phase I ................................................................... 81
Figure 34: Process level evaluation results for phase I .................................................... 83
Figure 35: Project level evaluation results for phase I ...................................................... 84
Figure 36: GQM tree for phase II (product only) .............................................................. 91
Figure 37: Product’s evaluation scheduling for phase II .................................................. 94
Figure 38: Product level evaluation results for phase II .................................................. 95
Figure 39: Holistic view presentation for the process, project and product level ............ 97
Figure 40: Combined holistic view presentation .............................................................. 100
Figure 41: Static validation steps ..................................................................................... 101
Figure 42: Scope for static validation .............................................................................. 103
Figure 43: Approach for validation through expert opinion .......................................... 105
LIST OF TABLES

TABLE 1: RESEARCH QUESTIONS .................................................................................................................. 2
TABLE 2: DEFINITION OF TERMS USED IN THIS THESIS REPORT ..................................................................... 4
TABLE 3: PRESCRIPTIVE FRAMEWORKS WITH THEIR FORMAL ASSESSMENT GUIDELINES .......................... 14
TABLE 4: MAPPING OF RESEARCH QUESTIONS AND RESEARCH METHODOLOGIES .................................... 28
TABLE 5: SPI INITIATIVE FORM  ................................................................................................................... 31
TABLE 6: EVALUATION AREA TABLE  ........................................................................................................... 37
TABLE 7: DETERMINATION OF MEASURES USING IRON TRIANGLE METAPHOR .............................................. 39
TABLE 8: GQM CONSTRUCTION TABLE  ....................................................................................................... 40
TABLE 9: MEASUREMENT GOAL, QUESTION(s) AND METRICS  ..................................................................... 41
TABLE 10: METRIC FORM  ............................................................................................................................ 43
TABLE 11: EVALUATION SCHEDULING FORM  ............................................................................................... 51
TABLE 12: HOLISTIC VIEW CALCULATION TABLE  ......................................................................................... 57
TABLE 13: EXAMPLE OF HOLISTIC VIEW CALCULATION FOR PROCESS ....................................................... 59
TABLE 14: EVALUATION FEEDBACK CHECKLIST ............................................................................................ 63
TABLE 15: MAPPING BETWEEN SPI-MEF’s STAGES, SUPPORT ARTIFACTS, STEPS AND CONCEPTS ............... 65
TABLE 16: SPI INITIATIVE FORM FOR ONE SOLUTION .................................................................................. 69
TABLE 17: PHASE I: EVALUATION AREA TABLE ............................................................................................ 72
TABLE 18: PHASE I: GQM CONSTRUCTION TABLE FOR BG1 ........................................................................... 72
TABLE 19: PHASE I: MEASUREMENT GOAL, QUESTION(s) AND METRICS FOR BG1 ......................................... 73
TABLE 20: PHASE I: GQM CONSTRUCTION TABLE FOR BG2 ........................................................................... 76
TABLE 21: PHASE I: MEASUREMENT GOAL, QUESTION(s) AND METRICS FOR BG2 ......................................... 76
TABLE 22: PHASE I: METRIC FORM FOR M01_PRC ......................................................................................... 78
TABLE 23: PHASE I: METRIC FORM FOR M04_PRC ......................................................................................... 79
TABLE 24: PHASE I: METRIC FORM FOR M03_PRJ ......................................................................................... 80
TABLE 25: PHASE I: EVALUATION SCHEDULING FORM .................................................................................. 81
TABLE 26: PHASE I: NEW AND REUSABLE METRICS ..................................................................................... 82
TABLE 27: PHASE II: EVALUATION AREA TABLE ........................................................................................... 86
TABLE 28: PHASE II: GQM CONSTRUCTION TABLE FOR BG1 ........................................................................... 87
TABLE 29: PHASE II: MEASUREMENT GOAL, QUESTION(s) AND METRICS FOR BG1 ......................................... 87
TABLE 30: PHASE II: GQM CONSTRUCTION TABLE FOR BG2 ........................................................................... 90
TABLE 31: PHASE II: MEASUREMENT GOAL, QUESTION(s) AND METRICS FOR BG2 ......................................... 90
TABLE 32: PHASE II: METRIC FORM FOR M04_PRD ......................................................................................... 91
TABLE 33: PHASE II: METRIC FORM FOR M08_PRD ......................................................................................... 92
TABLE 34: PHASE II: EVALUATION SCHEDULING FORM FOR PRODUCT .......................................................... 93
TABLE 35: PHASE II: NEW AND REUSABLE METRICS ..................................................................................... 94
TABLE 36: IMPACT RATING GUIDELINES ........................................................................................................... 96
TABLE 37: HOLISTIC VIEW CALCULATION FOR PRODUCT .................................................................................. 97
TABLE 38: SUCCESS INDICATORS AND METRICS FOR ORGANIZATION AND EXTERNAL LEVEL ................... 99
TABLE 39: TOPICS/concepts FOR VALIDATION AND THEIR AIMS .................................................................. 103
TABLE 40: ACADEMIA / INDUSTRY INTERVIEW QUESTIONS ......................................................................... 107
TABLE 41: QUESTIONS FOR INDUSTRY QUESTIONNAIRE .............................................................................. 109
TABLE 42: ACADEMIA INTERVIEWEES’ BACKGROUND .................................................................................... 111
TABLE 43: INDUSTRY INTERVIEWEES’ AND RESPONDENTS’ BACKGROUND ...................................................... 118
TABLE 44: ISSUES/suggestions CONSIDERED FOR REFINEMENTS ............................................................... 130
TABLE 45: ISSUES/suggestions NOT CONSIDERED FOR REFINEMENTS AND THE RATIONALES ....................... 132
TABLE 46: ISSUES/suggestions NOT ADDRESSED IN THE SPI-MEF ................................................................. 133
TABLE 47: ISSUES AND IMPLEMENTED SOLUTIONS ........................................................................................ 138
TABLE 48: SUMMARY OF THE VALIDATION RESULTS ..................................................................................... 141

VII
1 INTRODUCTION

Software plays an important role in our everyday lives since more and more products in the market incorporate software that drives the products’ functionality into their operation [1]. With this importance, the software engineering discipline and the study of the processes involved in software development have started to gain more popularity among researchers and practitioners in industry [2] [3] [4] [5]. One of the common research areas in software engineering is in the field of Software Process Improvement (SPI) [6]. SPI involves the understanding of the software processes as they are used within an organization and suggests areas for improvements in achieving specific goals such as increasing product quality, operation efficiency and cost reduction [7].

Assessing the outcomes of SPI is as important as their actual implementation since without a clear picture of gains or losses, it is impossible to reason about the performance of an SPI initiative [8]. The SPI literature provides many case studies of successful companies and descriptions of their SPI programs [9], and recent examples are presented in [10] [11] [12] [13] [14] [15] [16] [17] [18] [19]. Measurement is a primary facility to enable the software process to perform with predictable performance and capability and to ensure that process artifacts meet their requirements [20] [21]. However, to have an effective and meaningful measurement for the evaluation of the improvement, the correct metrics need to be determined [22]. Ambiguities in evaluating the results of SPI arise when the metric used is developed without thorough understanding of the object of study [23]. The success of improvement initiatives also means different things to different people [24] and therefore various stakeholders’ point of view (e.g. software developer, change agent or manager) have to be taken into consideration to make the evaluation adequate to assess the outcome from the SPI program [25]. Aside from that, the causal relationship between the improvement initiative and its effect is complex, and it is hard to determine whether the effect being measured is coming solely from the improvement initiative [26].

The aforementioned complexities and the lack of guidelines for conducting SPI measurements have raised the challenge to develop and implement effective performance measurement programs for SPI [27]. Since the evaluation of the outcome of SPI initiative is complex but also crucial to the organization, there is a need for a measurement and evaluation framework which guides SPI practitioners in their work, preserving effort and cost. The framework shall promote an evaluation which considers the improvement from different views and, in sum, increases the visibility, and consequentially alleviates the assessment of the achieved benefits. This thesis presents the continuation work of “Towards an Evaluation Framework for Software Process Improvement” [28] that proposed a model for the evaluation of SPI outcome. Motivated by the findings in [28], a measurement and evaluation framework called Software Process Improvement Measurement and Evaluation Framework (SPI-MEF) is constructed based on the proposed model. The framework is then validated by expert opinion originating from both academia researchers and industry practitioners.

1.1 Aims and objectives

Two aims are followed in this thesis work. The first aim is to build a framework based on the proposed model in [28] for the systematic evaluation of SPI initiatives. The second
aim is to validate the framework within industry context and incorporate the gathered feedback to refine the framework. These aims will be achieved by addressing the following objectives:

- Study the proposed model in [28] in order to investigate the properties and concepts that shall be incorporated into an evaluation framework for SPI initiatives.
- Formulation of the proposed framework based on the findings from the previous investigation.
- Statically validate the proposed framework from both the academia researchers’ and industry practitioners’ point of view.
- Incorporate recommendations gathered from the feedback of the static validation to refine the proposed framework.

1.2 Research questions

The research questions that will be answered during the course of this thesis together with their aims are listed in Table 1. The details on how each of these research questions is addressed can be found in Chapter 4.

Table 1: Research questions

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Aim</th>
</tr>
</thead>
<tbody>
<tr>
<td>RQ1: How can the evaluation model in [28] be implemented in a measurement and evaluation framework of SPI initiatives?</td>
<td>To analyze the concepts proposed in the evaluation model and incorporate them into a measurement and evaluation framework for assessing the outcome of SPI initiatives.</td>
</tr>
<tr>
<td>RQ2: To what extent is the model applied in the framework able to assist SPI evaluation in an industrial context?</td>
<td>To find out and verify whether the framework is able to provide a systematic evaluation of the outcome of SPI initiatives in an actual industrial environment.</td>
</tr>
<tr>
<td>RQ3: To what extent is the proposed framework applicable in practice, in terms of its flexibility and reusability in different industrial contexts?</td>
<td>To find out whether the framework is usable by different organizations, regardless of size, domain, and SPI initiative. It is also of interest to find out if the framework is adaptable to different industry settings.</td>
</tr>
<tr>
<td>RQ4: What are the refinements to be incorporated into the proposed framework based on the findings from RQ2 and RQ3?</td>
<td>To find out what recommendations are gathered as a result from the static validation to improve the framework. The recommendations will be incorporated to refine the framework.</td>
</tr>
</tbody>
</table>

1.3 Expected outcomes

The expected outcome from this thesis is a measurement and evaluation framework that is statically validated in an industrial context. The framework will consist of evaluation steps that guide practitioners to perform the evaluation of SPI outcomes. A list of support artifacts accompanying the evaluation steps and an elaborated scenario are provided to further assist the application of the framework.
1.4 Structure of the thesis

The overall structure of this thesis is depicted in Figure 1. The content is logically divided into three main parts.

The Background Research gives the reader the required information and understanding on the topic before reading the following chapters. The Background (Chapter 2) gives a brief introduction to SPI. This is followed by an introduction to software measurement, background about measurement frameworks, different approaches in software process improvement evaluation, and a summary of the proposed evaluation model in [28]. A brief summary on the related work in this research area is highlighted in Related Work (Chapter 3). The chapters in Background Research can be skipped if the reader is already familiar with these topics.

The Research Design presents the details on how the research is conducted. The Research Methodology (Chapter 4) presents the methodologies used to answer the indicated research questions. In Validity Threats (Chapter 8), a discussion on the threats to this research work is given.

The Research Contribution illustrates the details of the research work that is carried out in this thesis. Proposed Framework – SPI-MEF (Chapter 5) presents the framework for the systematic evaluation of SPI initiatives. Then, an elaborated scenario in which the framework is instantiated is illustrated in Scenario Applying SPI-MEF (Chapter 6). The design and execution of the validation together with the major changes that were incorporated in the framework after the validation are given in Static Validation (Chapter 7). In the Conclusion (Chapter 9), the conducted work is summarized and potential future work is discussed.
### 1.5 Terminology

Table 2: Definition of terms used in this thesis report

<table>
<thead>
<tr>
<th>Terms</th>
<th>Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPI initiative</td>
<td>All software engineering methods or activities which are intended to improve the performance of the software process. SPI initiatives can be categorized into frameworks (e.g. CMM, CMMI, SPICE, QIP, Six Sigma, etc), software engineering practices (e.g. inspections, test-driven development, etc.) or tools that support software engineering practices. For a more in-depth discussion see Section 2.1.2.</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Analysis on any kind of data with the aim to increase the knowledge on the evaluated entity.</td>
</tr>
<tr>
<td>Metric</td>
<td>A real objective measurement describing the structure or content of software products or software processes that has a standard unit of measure.</td>
</tr>
<tr>
<td>Validation</td>
<td>The process which determines the degree to which the concepts/features introduced by a model/framework can actually fulfill the needs of practitioners in the real-world.</td>
</tr>
<tr>
<td>Static validation</td>
<td>In the context of this thesis, static validation is done through a presentation of the framework to experts whose feedback is gathered regarding its applicability in real projects, instead of applying the framework in live projects. For a detailed discussion see Chapter 7.</td>
</tr>
<tr>
<td>Support artifacts</td>
<td>Support artifacts are materials in the form of guidelines, template forms and reference tables to guide organizations with the application of the proposed framework. For a detailed discussion see Chapter 5.</td>
</tr>
<tr>
<td>Evaluation steps</td>
<td>Evaluation steps are logical steps found in the proposed framework for the evaluation of the outcome of SPI initiatives. For a detailed discussion see Chapter 5.</td>
</tr>
<tr>
<td>Confounding factors</td>
<td>Usually unobserved variables that can distort the evaluation result and that hide or amplify the effect of the observed variables. For a detailed discussion see Section 2.4.3.</td>
</tr>
</tbody>
</table>
2 BACKGROUND

This section is aimed to provide sufficient background information to get a good understanding on the remaining chapters in this thesis report. The first section (Section 2.1) gives the general overview of software process improvement concepts. This is followed by a brief introduction to software measurements in the context of SPI (Section 2.2) and some widely used measurement frameworks in the industry will be presented in the subsequent section (Section 2.3). The next section (Section 2.4) provides background on the evaluation of SPI using formal assessment methods, the evaluation of benefits from improvement initiatives and an overview on confounding factors and their relation to SPI evaluation. In the end of this chapter (Section 2.5), a summary of the process improvement evaluation model proposed in [28] is given.

2.1 Software process improvement

2.1.1 Software process concepts

Software development has become a challenge for software practitioners since the boom of computing power in the early 1970's, and Dijkstra, in his ACM Turing Award Lecture in 1972, termed this challenge as "software crisis" [29]. Later on, more effort has been put into the software engineering field to make the software development process more disciplined, driven by the fundamental belief that the quality of a software system is governed by the quality of the process used to develop it [30] [31] [32]. In order to achieve a disciplined process, the process must be defined, trained, enforced and followed [32]. A disciplined process has the properties of an ordered pattern of collective behavior and increased team capability [32]. In addition, continuous improvement is needed to maintain a disciplined and mature process [30].

The formal definition from IEEE defines process as "a sequence of steps performed for a given purpose" [33]. A more elaborated definition is given by ISO 9000-1 which specifies process as "a set of interrelated resources and activities which transform inputs into outputs, and resources in this sense include personnel, finance, facilities, equipment, techniques and methods" [34]. Similar to the definition given by 9000-1, Paulk et al. [35] define software process as "a set of activities, methods, practices, and transformations that people use to develop and maintain software and the associated products (e.g., project plans, design documents, code, test cases, and user manuals)".

The establishment of discipline in the software process happened gradually and it was driven by two major "waves" as referred by Robert Lai [36]. The first wave was the introduction of the waterfall life-cycle model [37] in 1970 that depicts the activities in software development [32]. The second wave is regarded as the conception of process improvement models like the Capability Maturity Model (CMM) in the late 1980's that drive the process maturity movement [32]. Since then, several software process improvement initiatives have been introduced with the aim to promote the engineering discipline in software industry.

2.1.2 Software process improvement initiatives

Software process improvement (SPI) is defined as the implemented changes within the software process that are aimed to gain an improvement [38]. In other words, SPI is aimed to
assess and improve the processes and practices involved in software development. In general, continuous software process improvement consists of four iterative steps [39] [40]:

- Assessment of the current process status
- Elaboration of an improvement plan
- Implementation of the improvement plan
- Evaluation of the improved process

An effective software process improvement environment should provide a process infrastructure supporting the improvement activities [32]. The infrastructure consists of organization and management infrastructures (roles and responsibilities) and a technical infrastructure (technical tools and facilities) [32].

Many activities can be considered as SPI initiatives. In this thesis, SPI initiatives are defined as all software engineering methods or activities that are aimed to improve the performance of the software process. The SPI initiatives are grouped into three categories and are described below:

i. **Frameworks**
   SPI frameworks are models that provide guidelines for organizations to improve their software process capability and maturity. The guidelines provided by SPI frameworks can help organizations to prioritize their improvement and to focus on the area where more attention is needed. Several SPI frameworks define their own formal assessment methods to evaluate the organization’s process adherence with the standard requirements. The assessments give solid evidence of their increasing capability and maturity in software development. Examples of well-known SPI frameworks are Capability Maturity Model Integrated (CMMI) [41], ISO/IEC 15504 (also known as SPICE – Software Process Improvement & Capability dEtermination) [42], Quality Improvement Paradigm (QIP) [43] and Six Sigma [12].

ii. **Practices**
   Practices are software engineering activities that are planned and performed on a certain process or process area within the software development process in order to achieve particular goals. A process area is a cluster of related practices that, when performed collectively satisfy a set of goals considered important for making significant improvement [41].

iii. **Tools**
   Tools are software applications that support software development activities or the implementation of certain practices for software process improvement. Using tools in the software development process can increase the performance of the process, thus tools are considered as SPI initiatives. Examples are requirements management, test automation and configuration management tools.

### 2.2 Introduction to software measurement

#### 2.2.1 Software measurement

The dependency on software products has risen over the last decades [44]. Many of today's devices, whether they are home appliances, industrial machineries or personal devices, embed some kind of software [44]. This dependency demands industry to produce
high quality software and this condition has drawn the attention of software researchers and practitioners to improve software quality [1]. According to Fenton there are two views towards software quality [45]:

i. **External viewpoint**
   This view characterizes the software quality from the user’s perception of the final product.

ii. **Internal viewpoint**
   This view characterizes software quality by looking at the criteria that can be used to control the software quality during product development.

The quality of the software can be determined by applying software measurements [1]. Software measurements provide a quantification of software quality, not just in terms of the software as a product but also by the performed process and spent resources for producing the software [46]. Software measurement provides better control and visibility in the software development process and in the resulting product, and thus can be helpful in decision making [47]. The implementation of process improvement requires software measurement since if the results (whether in the development process or in the final product) of the process change are not measured, it is difficult to conclude that the improvement initiatives address the right issue [32].

2.2.2 Categorization of software measurement

A categorization of software measurements can be done in several ways [46]. Software measurements can be grouped into three categories based on the measured entities [48]:

i. **Product measures** are measurements that are collected from the software product, e.g. number of defects after release.

ii. **Process measures** are measurements that are collected from the methods, activities and practices used in developing a software product, e.g. the number of defects found during testing.

iii. **Resources measures** are measurements that are collected from the time, cost, effort, personnel or other kinds of resources used in the activities for developing a software product, e.g. the effort in man-months expended in the coding phase.

Software measurements can also be categorized based on the dependency between them [49]. According to this categorization, there are two types of software measurements [49]:

i. **Direct or fundamental measurements** are measurements of attributes that do not depend on other measurements and can stand-alone. An example for this type of measurement is the size of the software expressed in lines of code.

ii. **Indirect or derived measurements** are measurements that depend on the measurement of one or more other attributes. An example is productivity, expressed as lines of code (LOC) per hour, which requires the measurement of attribute size and time.

The measured attributes can also be grouped into two categories [48]:

i. **Internal attributes** can be measured purely from the entities themselves without relating them to their behavior, e.g. size of the software in lines of code.
ii. *External attributes* can only be measured from the entities by relating them to their environment. An example is the number of failures experienced by users of a software product.

### 2.2.3 Challenges in software measurement

Software measurement is difficult to implement and is a relatively new discipline [50]. The immaturity and the lack of standards in software measurement negatively impacts software engineering in general [51] and problems in data collection arise because of the poor definition of software measures [52]. Software measurement can only be *effective* and *meaningful* in a disciplined software process environment although software measurement may also be required in a non-disciplined environment [32].

Two factors that can cause difficulties in the implementation of software measurement and in the interpretation of the results: a human and a technical factor [53]. The human factor is characterized by the issue that people are involved in the data collection and serve also as data sources [53]. The technical factor is related to the issues in the process of collecting data for software measurement [53]. These two factors can also be considered as confounding factors (see Section 2.4.3 for a more in-depth discussion of confounding factors).

A study by Pfleeger in [54] revealed as a cause of the problems the different and sometimes conflicting motivations among the interest groups in software measurement. She identified three main interest groups, namely researchers, practitioners, and customers [54]. Researchers, mostly from academia, are motivated by publications they can produce and which provide highly theoretical results and are never used in the real world or tested empirically [54]. Practitioners are typically eager to achieve results in a short-term and are not always willing to be a test bed for new studies or providing their data to researchers since they fear that corporate information will be revealed to competitors [54]. Customers are often not involved in the development process and moreover, they cannot influence the development process, thus the customers can only hope that all the requirements are fulfilled as they expected [54].

Regardless of these obstacles, software measurement is still required in software engineering [55]. Software measurement plays an important role in the software engineering field [56] since the achievement of maturity in software technology depends on it [57]. As long as software measurement is well-understood and utilized properly it is still considered as beneficial [58].

### 2.3 Measurement frameworks

Mendonça et al. define a measurement framework as a "set of related metrics, data collection mechanisms, and data uses inside a software organization" [59]. Measurement must be based on goals and models [60]. There are a variety of frameworks available for measurement and defining measurable goals. The Goal-Question-Metric (GQM) approach, Quality Function Deployment (QFD), Practical Software Measurement (PSM) and the Balanced Scorecard (BSC) approach are few of the major well-known frameworks. These measurement frameworks are discussed briefly in the following sections.
2.3.1 Goal-Question-Metric (GQM)

GQM proposes a goal oriented approach where all collected data is based on a rationale that is explicitly documented [61]. The definition of goals and the decision of what to measure is based on a top-down refinement of measurement goals into questions and then into metrics. The interpretation of data is based on a bottom-up analysis of the collected metrics according to the previously defined context by goals and questions [62]. The GQM measurement model has three levels as described below:

i. **Conceptual level (Goal):** In this level, a goal is defined for an object (e.g. products, process, resource) for any purpose (e.g. characterize, evaluate) with respect to particular characteristics (e.g. defect removal, reliability) of the object of study, from point of view of the person needing the information, relative to a particular environment [60].

ii. **Operational level (Question):** In this level, a set of questions is formulated to characterize the way the assessment/achievement of the defined goal in the conceptual level is going to be performed by the selected viewpoints [60].

iii. **Quantitative level (Metric):** Each question defined in the operational level is then refined into metrics in this level. The metrics helps to objectively or subjectively assess the specific characteristics of the object of study and provide information to answer the specified questions in the operation level [60].

Figure 2: The GQM paradigm (inspired by [60])

As mentioned in [61], the GQM method contains four phases:

i. **Planning phase:** In the planning phase a project is selected for measurement. Measurement plans for the project are defined and integrated in the project plan. This phase also includes all basic requirements like training and management involvement to make the GQM measurement program a success.
ii. **Definition phase:** The definition phase specifies and documents the measurement program by describing questions, related metrics and expectations (hypothesis) of the measurements.

iii. **Data collection phase:** In this phase, actual data collection takes place. Data collection forms are defined, filled-in and stored in a measurement database.

iv. **Interpretation phase:** In this phase, collected data is processed with respect to the defined metrics to answer the stated questions, and these answers are again used to evaluate whether the stated goals have been attained.

![Figure 3: The four phases of GQM method (inspired by [61])]()
The “House of Quality” in QFD consists of six main parts [66], as shown in Figure 4, and presents basically the intersection of two dimensions [67]. The horizontal dimension located in Part 1 refers to the WHATs (also known as “Voice of the customer”) that identify the characteristics of the product desired by the customer while the vertical dimension located in Part 2 refers to the HOWs, that is the way (e.g. technical requirements) identified by the QFD team to achieve the WHATs [67]. The QFD team consists of people from different parts of the organization (e.g. marketing, design, project management, QA, development, etc.) [64]. Part 3 specifies the trade-offs in the technical specification of the product identified in Part 2. The relationship matrix, Part 4, located in the central part of the “House of Quality”, represents the strength of the relationship between WHATs and HOWs, that is, it correlates what customers want from the product and how the company can meet those requirements. Part 5 describes how customers and users perceive the competitors’ systems’ abilities in meeting the requirements (the WHATs). In Part 6, the technical feasibility is assessed and a technical competitive analysis is conducted by designers and developers with respect to the perceived competitors’ systems’ abilities in meeting the HOWs. By integrating all those parts mentioned above in the “House of Quality”, customer's voice can be incorporated throughout all the development activities providing the traceability between development activities and customer requirements [66].

2.3.3 Practical Software Measurement (PSM)

PSM is a “process for designing and implementing a project-based software measurement program that provides essential information on scheduling, resource allocation, and technological performance” [68]. It assists the decision making process that will affect the projects outcome positively by software managers and developers [68]. PSM derives from a measurement framework that is built up from four elements [69]:

![Figure 4: A simplified “House of Quality” structure (inspired by [66] and [67])](attachment:house_of_quality.png)
i. Information (measurement) needs
ii. Information (measurement) user
iii. Measurement process model
iv. Measurement information model

The last two elements, the measurement process model and measurement information model, have been codified as the international standard ISO/IEC 15939: Software Measurement Process [69]. The PSM measurement process model (Figure 5) consists of four basic activities [69]:

i. Establish commitment
   This activity obtains the necessary sponsorship and support to sustain a measurement program.

ii. Plan measurement
    Planning the measurement involves the understanding of the information needs and defining appropriate measurement constructs to address them.

iii. Perform measurement
    Measurements are collected and data is analyzed as according to the defined measurement constructs.

iv. Evaluate measurement
    The measurement program is assessed and improved.

![Figure 5: The PSM measurement process model (inspired by [69])](image)

A consistent set of terms and concepts for describing measurement constructs and activities is defined by the PSM measurement information model [69]. There are three levels of measures defined by the information model [69]:

i. Base measures
   Base measures refer to the quantifications of a single attribute.

ii. Derived measures
   Derived measures are the combinations of values of base measures.
iii. **Indicators**

Indicators are the base and/or derived measures with associated analysis models and decision criteria.

2.3.4 **Balanced Scorecard (BSC)**

BSC is a strategic planning and management system that helps to guide organizations to transfer abilities and specific knowledge held by people throughout the organization in order to achieve long-term strategic goals [70]. It provides a comprehensive framework for executives in an organization to translate company visions and strategies into a set of coherent performance measures [70].

![Figure 6: The Balanced Scorecard (inspired by [70])](image)

The BSC proposes to view the organization from four perspectives (Figure 6) [70]:

i. **Financial perspective:**
   It is of crucial importance to measure economic consequences of actions taken in the organization. The core measures include operating income, return-on-investment, sales growth and generation of cash flow.

ii. **Customer perspective:**
   This perspective measures the customer and market segment performance. The typical measures are customer satisfaction, customer retention, new customer acquisition, customer profitability and market share.

iii. **Internal-business-process perspective:**
   The internal-business-process perspective measures focus on internal processes that have impact on customer satisfaction and achieving an organization's financial objectives. Generic measures include quality, time and cost.

iv. **Learning and growth perspective:**
   This perspective identifies the infrastructure that the organization must build to create a long-term growth and improvement. The measures principally come from three sources: people, systems and organizational procedures.
2.4 Evaluation of software process improvement

2.4.1 Formal assessment

In prescriptive frameworks, in order to achieve a certain maturity level, organizations need to adhere with a set of practices proposed by the framework [40]. Formal assessment can be used to assess whether an organization is at a certain maturity level [71]. Formal assessment is an appraisal aimed to determine the maturity level of an organization based on its current condition, to provide prioritization on the organization’s issues, and to gain commitment/support for software process improvement [35].

A study by Zahran discusses three typical phases in conducting a formal assessment [72]:

i. **Phase 1**
   This phase is called preparation phase. The main objective of this phase is to obtain commitment and support from top level management to start the process and to take action based on the results of the assessment process. One of the activities in this phase is training for the people involved in the assessment.

ii. **Phase 2**
   This phase is called on-site assessment. In this phase the assessment process is conducted to assess the current processes in the organization. The result of this phase is a preliminary report of the findings.

iii. **Phase 3**
   This phase is called recommendation phase. In this phase the results of the assessment is presented to the managers and a team to plan and implement the recommendations is formed.

Several prescriptive frameworks are accompanied with guidelines to conduct formal assessments [40]; some examples are listed in Table 3.

<table>
<thead>
<tr>
<th>Prescriptive Framework</th>
<th>Formal Assessment Guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capability Maturity Model (CMM)</td>
<td>CMM Appraisal Framework (CAF) [73]</td>
</tr>
<tr>
<td>CMMI</td>
<td>Appraisals Requirements for CMMI (ARC) [74]</td>
</tr>
<tr>
<td>ISO/IEC 15504</td>
<td>ISO/IEC 15504 [32]</td>
</tr>
</tbody>
</table>

2.4.2 Actual benefit evaluation

In actual benefit evaluation the assessment of improvement is conducted by utilizing the results of metrics derived from a certain measurement derivation method (e.g. GQM, see Section 2.3.1) and not by assessing the conformance to a set of rules or process areas [75]. Even though several prescriptive frameworks provide formal assessment guidelines, compliance to standards does not guarantee that actual benefits are achieved [1].

Organizations embarking SPI initiatives must be able to quantify the benefits obtained from the initiatives [75]. The quantification of benefits is required in order to know whether the organization’s goals (e.g. improved product quality, increased customer satisfaction, return on investment) have been achieved or not [75]. Actual benefit evaluation helps organization to determine their achievement level on the targeted goals [75]. The framework presented in this thesis is aimed to provide support for actual benefit evaluation.
2.4.3 Confounding factors

Confounding factors, in the context of experimental design, were defined as “variables that may affect the dependent variables without the knowledge of the researcher” [76]. Similar definitions in [77], [78], [79], and [80] mention that confounding factors need to be associated with the independent variable in order to distort the measured effect on the dependent variable (Figure 7). Confounding factors can influence the conclusions regarding causal relationships between independent and dependent variables (they represent an internal validity threat) since the effect on dependent variables cannot be attributed solely to the independent variables. Simpson’s paradox, as described in [80], can be seen as an extreme example which illustrates how confounding factors distort the analysis and lead to conclusions which are a reverse of the actual situation.

The control of confounding variables in comparative research is the major concern in “Statistical Methods for Comparative Studies: Techniques for Bias Reduction” [81]. The authors propose study designs and statistical techniques to adapt the analysis in order to compensate for known confounding variables. In principle, the effect of confounding factors on the dependent variable can be limited by designing the study appropriately, e.g. by a random allocation of the subjects in the treatment and control groups. Such a design implies that the confounding variables are equally distributed in each group, i.e. that there is a high probability that the groups have similar properties. Hence, if the distribution of the dependent variable is similar in both the control and treatment group, it can be concluded that the treatment has no effect regardless the existence of confounding factors.

If it is not possible to create randomized samples, the matching technique can be used to group the samples in a way to minimize the effect of confounding variables. The basic idea inherent in all the presented matching techniques in [81] is to form study groups which are homogeneous in regard to possible confounding variables (that is, analogous to randomization, increase the probability that the association between confounding variable and independent variable is similar in the treatment and control group).

The above described methods can be applied before the study is conducted. The authors in [81] describe also adjustments techniques which are applicable during the analysis, that is, post-factum and if the groups were indeed not comparable. In order to effectively apply any of these methods, it is necessary to specify a valid mathematical model which describes the relationships between dependent, independent and confounding variables.

A major drawback of the previously discussed techniques is the prerequisite that the confounding variable is known in advance in order to control its effect. Without the
knowledge about the existence of the confounding variables neither an adequate design of the study nor the selection of an appropriate adjustment technique can minimize their effect.

2.5 Proposed evaluation model

For the implementation of the framework presented in this thesis the process improvement measurement and evaluation model proposed by Cheng et al. [28] is adapted and extended. The model consists of seven core concepts as shown in Figure 8. The core concepts are summarized below.

![Evaluation model for the evaluation of SPI initiatives (inspired by [28])](image)

2.5.1 Measurement levels

Measurement levels represent the spectrum of entities which are affected by SPI initiatives and need to be measured in order to achieve a holistic evaluation of the SPI initiative’s outcome. Five levels are defined in this model: Process, Project, Product, Organization and External. Each of those is illustrated briefly below:

i. **Process level**

The process is the main target of SPI initiatives. Therefore, it is important to measure the performance of the process to find out the improvement in its efficiency and effectiveness. Improvement in the process alone is not sufficient to judge the SPI success since there could be side effects in other processes or in other entities at different measurement levels which are also required to be taken into consideration.
ii. **Project level**

The assessment at this level is mainly focused on project control by monitoring budget, schedule and resources. It is common to measure the effects of newly introduced processes by assessing the work products created during the project. Adherence to project estimates and a gain in project quality can indicate process improvement.

iii. **Product level**

Most improvement activities are driven by the need to increase the internal and external product quality (see Section 2.2.1). Although quality is one of the most measured entities [28], other entities such as cost and time-to-market should be also included when evaluating the outcome of SPI initiatives in this level.

iv. **Organization level**

The implementation of SPI initiatives must be coherent with the organization’s vision and business goals. Therefore, it is important to assess the alignment between the improvement gained by implementing SPI with the organization’s strategy and business goal.

v. **External level**

The improvements experienced within the organization can also affect other entities outside the organization such as suppliers and the users of the provided services or products. Measurements in this level assess the effect of implementing SPI to the entities outside the organization.

The measurement levels are ordered from inner level to outer level starting from Process, Project, Product, Organization and External according to temporal, aggregation and traceability arguments as described in [28].

2.5.2 **Evaluation viewpoints**

Evaluation viewpoints represent the stakeholders within the organization who are interested and need to see the evaluation results of an SPI implementation. The model defines three evaluation viewpoints: Implementer, Coordinator, and Sponsor [28].

i. **Implementer**

The Implementer viewpoint considers those roles that are responsible to implement the software development process and the SPI initiatives. This viewpoint needs to know the evaluation results in order to be motivated and support the SPI program.

ii. **Coordinator**

The Coordinator viewpoint includes people that are in charge in coordinating and managing the implementers of the software development process and the SPI initiative. This viewpoint requires knowing the evaluation results such that it can be gauged whether the improvement goals are achieved or not, to guide the next improvement activities, and to provide feedback to superiors.

iii. **Sponsor**

The Sponsor viewpoint considers roles that provide funding and motivation to the SPI implementation. These people are interested to see the evaluation results in terms of costs and benefits in a long-term period.
2.5.3 Evaluation area

The evaluation area is the cornerstone of the entire evaluation model. The evaluation area guides the evaluator to determine where to measure and for whom the measurement results are important to see [28]. The area is defined by two dimensions: measurement levels and evaluation viewpoints [28]. These two dimensions are mapped orthogonally to help the elicitation of the roles in the organization that are interested to see the evaluation results on each measurement level.

2.5.4 The measurements

The measurement concept in the model proposes a method to obtain the needed metrics. The concept addresses three problems in the measurement and evaluation of process improvements [28].

- Benefits of SPI are visible only partially if one measurement level alone is assessed. This problem is addressed by applying the evaluation area concept. This concept provides support in measurements elicitation by defining where to measure (measurement levels) and who is interested to see the measurement results (evaluation viewpoints).

- Data validity and the confidence level of the evaluation results based on the collected data. Using a single metric to measure a certain attribute can affect the validity of the forthcoming evaluation. In order to countermeasure this issue, the measurement concept incorporates cross-examination which can be used to increase data validity in two ways. First, by using more than one data collection mechanism, and second, by using more than one data source or collecting the same data at different occasions.

- Side-effects of the implemented SPI initiative are not assessed and evaluated. The elicitation of measurements should consider two kinds of measurements: primary and complementary measurements. The former are measurements that gauge the accomplishment level of an improvement goal. The latter, on the other hand, are measurements that are needed to capture potential side-effects of the process change on attributes and entities that are not directly related to the improvement goal. The elicitation of complementary measurements can be conducted by using auxiliary improvement goals.

2.5.5 Evaluation methods

The evaluation methods give a guideline on how to analyze and evaluate the measurement results in order to determine the improvement. The model defines four evaluation methods: basic comparison, statistical based analysis, survey, and cost-benefit analysis [28]. Each of them is described briefly below:

i. **Basic comparison**

Basic comparison is the most commonly used evaluation method [28]. The main idea is to compare between measurement results of attributes. This evaluation method requires baseline values such that the improvement can be evaluated. In order to achieve valid evaluation results, confounding factors need to be considered [28].
ii. **Statistical based analysis**
Statistical analysis can be done using either descriptive statistics or inferential statistics. The data in statistical based analysis can be expressed in the form of numerical values (mean, median, estimation, etc) or graphically (charts and graphs). An example of an evaluation and monitoring method that uses statistical methods is Statistical Process Control (SPC). Confounding factors also need to be considered when using statistical based analysis [28].

iii. **Survey**
Survey is an evaluation method that collects and evaluates input from human subjects either qualitatively or quantitatively. Human subjects can be the internal people of the organization or external, e.g. customers. A survey can be an option to construct baselines if no historical data is available for comparison [28].

iv. **Cost-benefit analysis**
Cost-benefit analysis is an evaluation method that assesses the business impact of implementing SPI by considering the cost of implementing SPI initiatives and the benefit gained. One difficulty when using this method is to identify and quantify all relevant costs and all gained benefits. One popular metric that can be used in this evaluation method is Return-on-Investment (ROI) [28].

2.5.6 **Time to evaluate**
The model does neither provide a concrete schedule nor concrete guidelines to determine the appropriate time. Instead, it proposes two factors that need to be considered when determining the time to evaluate, namely latency and decay. Latency is the minimum time an improvement initiative has to be applied in order to show a visible and measurable outcome [28]. Decay is defined as the maximum time for which an evaluation result can be regarded as valid and used to base decisions on it [28].

2.5.7 **Holistic evaluation**
The model also proposes a simple representation of the evaluation results that can address the visibility issue of the improvement results as a whole. This representation is mostly suitable for the Sponsor viewpoint. The representation consists of representative values from every measurement level and can be broken down per measurement levels to see the improvement for each viewpoint respectively.

Three aspects are mentioned that need to be considered when constructing the representation model [28]:

i. Normalization of different measurements/metrics such that they can be aggregated.

ii. Compensation of different orders of magnitude of the metrics values.

iii. Consideration of the viewpoints opinion in judging the importance of a specific metric.
3 RELATED WORK

Based on the results of a preliminary study on the literature, there are no frameworks for outcome measurement and evaluation of software process improvement which incorporate the model proposed in [28]. However, several studies related to frameworks for measurement and evaluation of SPI were identified. The studies are briefly discussed in the paragraphs below.

A study by Saeidian [83] proposes a framework for evaluating SPI models (e.g. CMM, CMMI, SPICE, ISO9000, AMI). The framework is based on questions that are grouped into four groups, namely WHAT, WHICH, HOW and WHERE. WHAT questions are intended to characterize the SPI models based on their goals, base structure, the role of management, metrics used, and benefits of the model. WHICH questions are intended to gather information about principles, methods and standards from which the SPI model is developed. HOW questions are used to investigate how the SPI model rates the capability of an organization and how it can be used for an effective SPI program. WHERE questions address the information about organizations that have used the SPI model and in which countries those organizations are situated. Despite of those characteristics, the framework is rather meant for evaluating the SPI models in order to determine which one is suitable to be implemented by an organization instead of evaluating the actual benefit of implementing an SPI model. Furthermore, the framework does not address software engineering practices and tools which can also be considered as SPI.

Wilson et al. conducted a study in order to build a framework for evaluation and prediction of SPI success [84]. The framework is based on questions that are grouped according to four perspectives which were proposed by Jeffery [85]. The perspectives are Context, Inputs, Process, and Products. The Context perspective is aimed to gather information about the environment where SPI is implemented. The Inputs perspective is intended to collect information on the resources that are needed to conduct an SPI program. The Process perspective is targeted to obtain information regarding the method used to develop, maintain, and implement the SPI program. The Products perspective is intended to obtain data on the improvement and output of the SPI program. Although the framework has been validated, the purpose of the framework is for assessing the readiness of an organization to embark upon an SPI program and not to evaluate the outcome after implementing SPI.

Abrahamsson conducted a study to propose a method for measuring the SPI’s level of success [25]. The method comprises of two main components. The Dimensions which are proposed by Shenhar [86] and the stakeholders involved in SPI program by Zahran [32]. The Dimensions represent different interests and views of the parties involved in the SPI project. The dimensions are: “Project efficiency”, “Impact on the process user”, “Business success”, “Direct operational success”, and “Process improvement fit and preparing for the future”. The stakeholders are grouped into five different roles namely “Sponsorship”, “Management”, “Coordination”, “Operational”, and “Object of change”. The method proposed by this study is to determine the importance level of each dimension based on the stakeholders opinions. Practitioners may find the method helpful in determining the measurements for the SPI’s success level. This study shows the results of applying the approach with the stakeholder “Object of Change” (e.g. programmers, testers, etc). Although
the proposed idea incorporates SPI stakeholders and the dimensions can represent the measurable entities in a software organization, the study does neither propose concrete metrics for the dimensions nor provide a method to obtain the appropriate metrics.

A technical report by the Practical Software and Systems Measurement (PSM) is intended to propose measurements that are needed for measuring process improvement [87]. The measurements are required for three purposes: to justify the action of implementing SPI, to assess the readiness of an organization for SPI and to monitor the progress of the SPI. Measurements that can be used for the first purpose, justification of implementing process improvement, are grouped according to the four measurable concepts of the Balanced Scorecard (the financial, customer, internal process, and innovation and learning perspectives of an organization; see Section 2.3.4). The measurements for the second purpose, assessment of the organization’s readiness for conducting SPI, are grouped into the alignment and commitment, and the process improvement capability measurable concept. The classification of measurements for the third purpose, monitoring the progress of SPI, is based on the measurement classification of software and system projects, such as size, resources and cost, product quality, and customer satisfaction. The study provides measurable entities and concrete metrics to measure them, but no discussion is provided about a procedure to select suitable metrics since a company may not afford to use all proposed metrics.

A technical report from Software Engineering Institute (SEI) proposes concepts for measuring the benefits of software process improvement [75]. The report emphasizes the importance of benefit measurement. Several measurements that can be used to measure the benefit of implementing SPI are explained in this report along with guidelines to collect and to interpret them. An evaluation method by comparing the measurement results before and after an SPI implementation is also introduced in this report. However, the report does not discuss how to select the appropriate measurements.

A study by Bhatti et al. [82] presented a model which can be used by organizations to select optimum measurements according to their constraints and limitations. The model is called Optimum Measures Set Decision (OMSD). OMSD is constructed by adding a heuristic approach to the GQM paradigm. The heuristic is implemented by incorporating factors that the organizations need to consider when they determine the most appropriate measurements. Bhatti et al. also mentioned that OMSD can be used to select the measurements for SPI.

This thesis is aimed to build a framework for evaluating the outcome of SPI initiatives based on the model proposed in [28] and validate it. This implies also that it is aimed to address the issues regarding SPI evaluation (see Section 9.1, Table 47). The following paragraph shows how the above discussed research relates to the concepts proposed to solve these issues and motivates the implementation of SPI-MEF.

The outcome of implementing SPI is also discussed in [83] and [84]. The difference is that the framework presented in this thesis is not intended for predicting the outcome before SPI implementation but rather for evaluating the outcome after the SPI implementation. The model on which the framework is based on proposed a concept that shows the involvement of stakeholders in the evaluation process to identify entities which are important to be measured [28]. This concept is coherent with the argument from Abrahamsson in [25], which however misses to explain how to elicit the appropriate metrics. Another concept discussed in the model is concerned with the measurements that are required in the evaluation process.
Measurements for SPI evaluation are also discussed in [75], [82], [83] and [87]. However, this thesis is intended to provide a more concrete way to incorporate the measurements in the evaluation process. Although the OMSD model [82] claims to be applicable in SPI context, it is limited to the selection of measures and does not address the evaluation of SPI. The concept regarding the evaluation methods as proposed in the model by Cheng et al. [28] is also incorporated in this thesis. From the related studies only [75] discusses evaluation methods. The appropriate times to conduct the evaluation and factors that can influence the evaluation results, which are not investigated in any related study, are discussed in more detail in this thesis. The proposed framework is also intended to provide a way to enhance the communication of evaluation results and to enhance the visibility of the improvement impact, which is to the best knowledge of the authors, not yet discussed in the context of SPI evaluation.
4 RESEARCH METHODOLOGY

To answer the questions mentioned in Section 1.2, two research methods were used: conceptual analysis and validation through expert opinion with the means of interview and questionnaire.

The definition and general idea of each of the research methods is presented in Section 4.1. Then, how the research methods were applied in this study to answer the research questions is illustrated by a flow chart in Section 4.2. In Section 4.3, a mapping of research questions and research methods is shown in a table along with the steps taken to answer the research questions.

4.1 Brief description of research methodologies

4.1.1 Conceptual analysis

According to Beaney, “conceptual analysis consists primarily in breaking down or analyzing concepts into their constituent parts in order to gain knowledge or a better understanding of a particular philosophical issue in which the concept is involved” [88]. According to the “Writing guides of content analysis” published by the Colorado State University, a concept is chosen for examination and in the analysis its presence is quantified and tallied [89]. Conceptual analysis was selected as one of the research methods in this thesis since it defines clear steps to conduct the analysis, which provides traceability.

Conceptual analysis can be conducted in the following steps as mentioned in [89]:

i. **Deciding the level of analysis:** Determining which word, set of words or phrases constitute the concept.

ii. **Deciding how many concepts to code for:** This step involves developing a pre-defined set of concepts and categories and determining the total number of individual concepts to code for.

iii. **Deciding whether to code for existence or frequency of a concept:** This step is to determine whether to simply look for the existence of a concept or to count it each time it occurs.

iv. **Deciding on how to distinguish among concepts:** This step determines the level of generalization that can be done for each concept. For example, to decide if two terms express the same idea, whether they can be grouped together as one item or if they should be interpreted differently.

v. **Developing rules for coding the texts:** To ensure that the coding of the concepts is done consistently and coherently, a set of rules are determined in this step. This step defines how to make an implicitly mentioned concept explicit so that coding is done exactly for what it was intended to code for.

vi. **Deciding what to do with "irrelevant" information:** This step determines whether to discard unwanted material or considering all information as relevant and important and using it to reexamine or reassess later.

vii. **Coding the texts:** This step is to code the text based on the decision for coding made in the previous steps.

viii. **Analyzing the results:** Based on the data gathered from the previous steps, analysis is done in this step to draw conclusions and generalizations where possible.
4.1.2 Expert opinion

Expert opinion can be gathered in several ways. The two methods that are used for gathering expert opinion in this thesis are interview and self-administered questionnaire. These two methods are briefly described in below:

4.1.2.1 Interview

The Interview is one of the commonly used techniques for collecting qualitative data in empirical software engineering research [90] [91]. Carolyn B. Seaman described in [90] that interviews are conducted with varying objectives like collecting historical data from the memories of interviewees, collecting opinions or impressions about something, or to help identifying the terminology used in a particular setting. The types of interview that are commonly found in the literature are:

- Structured interview
- Semi-structured interview and
- Unstructured interview

In structured interviews, the objective of the interview is very clear and specific. The interviewer knows well the type of information he/she is looking for, and therefore all the questions for the interview are very specific and prepared beforehand [90] [91]. On the contrary, in unstructured interviews, the objective of the interview is not to elicit specific information but to elicit as much information as possible on a broadly defined topic and the interviewer does not know the form of the information he/she is looking for. Therefore the interviewer might have only a few specific open-ended questions in mind to drive the interview or, in extreme cases, just mention the topic/theme to be discussed in the interview [90]. The semi-structured interview is a combination of these two types. Semi-structured interviews can have specific questions to elicit the information foreseen and, on the other hand, also allow bringing up new questions to elicit unexpected types of information during the interview as a result of what the interviewee says [90].

Interviews can also be divided into other two types depending on the communication between the partners [92]. These are:

- Personal interview: the interviewer conducts the interview face-to-face with the interviewee.
- Telephone interview: the interview is done by communicating with the interviewee through telephone lines.

Kvale presented seven stages of an interview investigation in [93]. These seven stages are briefly described below:

i. *Thematizing*: This step is to clarify the why (purpose) and what (the concept to be investigated) of the interview before the interviews start.

ii. *Designing*: Designing of the study includes planning the study with regard to the knowledge intended to obtain, taking into account the moral implications of the study (e.g. confidentiality of the interviewee) in all seven stages of the investigation before the interview starts.

iii. *Interviewing*: In this step, the actual interview is conducted based on an interview guide.

iv. *Transcribing*: The conversation in the interview is transcribed to written text for the purpose of analysis.
v. Analyzing: In this step, an appropriate method for analysis is chosen and interview results are analyzed based on the purpose of interview.

vi. Verifying: This step ascertains the generalizability, reliability (how consistent the results are), and validity (whether the interview investigates what is intended to be investigated) of the interview findings.

vii. Reporting: This step is for communicating the findings of the interview study through a scientific report or publication.

4.1.2.2 Self-administered questionnaire

A self-administered questionnaire is usually used as an instrument for data collection in survey research [94]. In this thesis, this method is used to gather expert opinions from those industry practitioners with whom face-to-face interviews are not possible.

Kasunic described in [94] four main steps in developing the questionnaire. The steps are described below:

i. Determining the questions to be asked: In this step, the research team identifies the questions that they need to answer for themselves (also known as internal questions). The selected identified internal questions will be used later to prepare the actual questionnaire items that the respondent will answer.

ii. Selecting the question type/format for each question and specify the wording: Each identified internal question of the previous step is transformed to one or more questions in the actual questionnaire. Questions are formulated in a way so that responses can be analyzed and the question wording is understandable to the targeted respondent. Questions can be of two forms: open-ended questions and closed-ended questions. Open-ended questions allow the respondent to answer questions freely in their own words without limiting their answer to fixed choices. On the contrary, closed-ended questions provide fixed answer choices and the respondent selects the one that best reflects his/her opinion.

iii. Designing the question sequence and overall questionnaire layout: This step involves determining the length of the questionnaire, organizing and sequencing questions, writing instructions and page layout considerations (e.g. font size, page size, graphics, text in header or footer etc.)

iv. Developing ancillary documents: This step involves the preparation of supportive documents (e.g. pre-notification letter/email, a questionnaire cover letter, a reminder letter/email (to remind respondents to complete the questionnaire before the deadline), a thank-you letter/email etc.)
4.2 Flow chart for the research methodology

Figure 9: Flow chart for the research methodology
The objective of research question one (RQ1) is to investigate how the model proposed for SPI evaluation in [28] can be incorporated in a measurement and evaluation framework. The result of the systematic review and conceptual analysis done in [28] is the basis for this investigation. The steps (step \(i \rightarrow vii\)) of the conceptual analysis mentioned in Section 4.1.1 were already applied in the systematic review in [28]. Step \(viii\), “analyze the result”, of conceptual analysis was also applied to come up with the SPI evaluation model. However, this step needs to be extended to explore the possibilities to incorporate the evaluation model in a concrete framework. Therefore, all the results from [28] are accumulated and raw data are collected from the authors for further analysis. Since, these results and the collected raw data is the outcome of the application of conceptual analysis and systematic review already, therefore, conceptual analysis was deemed to be appropriate method for further analyzing the previous work. Hence, step \(viii\), “analyze the result” of conceptual analysis is further extended in this thesis to come up with a concrete framework for SPI evaluation.

A framework for measurement and evaluation of SPI is proposed after the end of the conceptual analysis. Thereafter, to answer research question two and three (RQ2 and RQ3) the framework is validated in terms of its ability to assist in SPI evaluation and applicability in practice. The validation is done by presenting the framework to the experts and gathering their feedback regarding the applicability of the framework (static validation [95]). A validation by applying the framework in a live project in industry (dynamic validation [95]) cannot be done since the implementation of the framework in real company settings would exceed the time schedule of the thesis. For gathering expert opinion, two groups of people were identified: academic researchers who work closely with industry and industry practitioners who have experience related to SPI.

The method that is used for gathering expert opinion from academic people is “structured interview” (see Section 4.1.2.1). Two types of structured interviews are conducted: personal interview and telephone interview. The personal interview is conducted if it is possible to meet the researcher in person and a face-to-face meeting can be arranged. If a face-to-face meeting cannot be arranged with a researcher due to geographical distance or other reasons, a telephone interview is conducted.

For industry practitioners, two types of methods are used for gathering expert opinion: personal interview (structured type) and self-administered questionnaire. For all interviews, the steps for interview investigation mentioned in Section 4.1.2.1 are followed, while for the self-administered questionnaire, the steps mentioned in Section 4.1.2.2 are followed. The detailed design, results and analysis of the personal interview, telephone interview and self-administered questionnaire are given in Chapter 7. The actual material used for the questionnaire is shown in Appendix A.

Finally, conceptual analysis is conducted on the feedback received from experts in the validation step to answer research question four (RQ4). This assists in elaborating the refinement possibilities of the proposed framework. Findings from the validation step act as an input for the conceptual analysis. Hence, the conducted conceptual analysis is highly dependent on the information gathered from the validation. Therefore, the steps (step \(i \rightarrow \nu\)) of the conceptual analysis mentioned in Section 4.1.1 are applied and incorporated in the validation steps (validation scoping in Section 7.1 and validation design in Section 7.2) of the static validation so that the data can be collected in a systematic way and is usable for later use in conceptual analysis. Table 39 in Section 7.1 shows the categories of concepts that were selected for validation. Step \(vii\), “coding the text”, is done by transcribing the interview.
and self-administered questionnaire results into documents according to the pre-defined categorization of the concepts to code for. The results of this step are shown in Section 7.3 and 7.4. The results from the validation step drive step vii, “analyze the result”, of conceptual analysis. The analysis involves developing the meanings of the results by bringing the subjects’ opinion and understanding on the concepts into the light as well as providing new perspectives from the authors’ on the results. The outcome of the analysis step is presented in Section 7.5. Step vi of conceptual analysis, “Deciding what to do with ‘irrelevant’ information”, is done together with the analysis step. Based on the findings from the conceptual analysis, necessary refinements are incorporated in the proposed framework. The summary of the validation and conceptual analysis is given in Section 7.6.

4.3 Mapping of research questions and research methods

The research questions, corresponding research methodologies and steps followed to answer the research questions are described in Table 4.

<table>
<thead>
<tr>
<th>Research Questions</th>
<th>Steps</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RQ1:</strong> How can the evaluation model in [28] be implemented in a measurement and evaluation framework of SPI initiatives?</td>
<td>The concepts of the evaluation model in [28] is further investigated and elaborated to incorporate those in different steps of a measurement and evaluation framework.</td>
<td>Conceptual analysis</td>
</tr>
</tbody>
</table>
| **RQ2:** To what extent is the model applied in the framework able to assist SPI evaluation in industrial context? | The core concepts of the model applied in the framework are identified and presented to the expert for their judgment and opinions to see to what extent the model can assist in SPI evaluation. | Expert opinion:  
i. *Researcher:*  a. Personal interview  
b. Telephone interview  
ii. *Practitioner:*  a. Personal interview  
b. Questionnaire |
| **RQ3:** To what extent is the proposed framework applicable in practice, in terms of its flexibility and reusability in different industrial contexts? | The proposed framework is presented to the experts who provide their opinions regarding the applicability of the framework in practice. | Expert opinion:  
i. *Researcher:*  a. Personal interview  
b. Telephone interview  
ii. *Practitioner:*  a. Personal interview  
b. Questionnaire |
| **RQ4:** What are the refinements to be incorporated into the proposed framework based on the findings from RQ2 and RQ3? | Feedback from the experts regarding RQ2 and RQ3 are analyzed to make the necessary refinements in the proposed framework. | Conceptual analysis          |
5 PROPOSED FRAMEWORK – SPI-MEF

SPI-MEF is motivated by some key concepts that are considered as important in SPI evaluations. The key concepts exhibit their importance at different places in the evaluation process. Therefore, SPI-MEF is structured according to generic evaluation stages that start with the initialization stage, followed by the design and implementation stage and close with the evolution stage (Figure 10). Each of them has its specific steps that incorporate the key concepts and they are supported by some artifacts to assist the evaluation process. Since the nature of improvement is generally continuous, the evaluation reflects this by being iterative, that is, the stages represent a cycle in which also the evaluation is adapted, matured and improved.

![Diagram of SPI-MEF](image)

Figure 10: SPI-MEF overview

SPI-MEF provides a systematic way of evaluating the outcome of SPI initiatives. The evaluation of SPI initiatives is usually conducted after the initiative has been implemented and is done ad-hoc, without designing in advance any proper documented plans about what,
how and when to evaluate. The evaluation itself is as important as the implementation of the SPI initiative and organizations should draft out the evaluation plan before the start of the SPI initiative and not after the initiative has been implemented.

This approach enables organizations to think ahead and to be able to project their aims and expected goals as the result from the process improvement. Designing the plan before the embarkment of any SPI initiative also helps in pointing out metrics that should be collected during the implementation so that the intended evaluations can be performed effectively later on. Therefore, the first two stages in SPI-MEF are designed to address the need of evaluation planning.

The implementation stage is where the evaluation plan derived from the design stage is put into execution. Before evaluations can be conducted, organizations have to incorporate evaluation metrics into their current practice. The establishment of a measurement program, which addresses data validity, the human nature in the data collection process, and the incorporation of evaluation metrics, is dependent on the organization, i.e. different organizations have or need different approaches in their implementation. It is beyond the scope of SPI-MEF to address the establishment of a measurement program.

The evaluation of SPI initiatives is not complete without process improvement of the evaluation itself. After the implementation of the evaluation, the evolution stage is aimed to assess the evaluations that have taken place and elicit feedback for future evaluations. With this strategy, the evaluation process can be seen as an evolutionary process.

The steps that incorporate the key concepts in each stage of SPI-MEF are further assisted with support artifacts. Those are given in form of guidelines, template forms and reference tables to guide organizations with the proposed evaluation steps. In the following sections, the details of each evaluation stage are described.

---

Running scenario

In order to enhance the readers’ understanding, the stages and steps are accompanied by a running scenario as they are described. The background of the running scenario is described here so that the reader has contextual information about the organization mentioned in the scenario.

Company background

APOLLO International Ltd is a young medium-sized software development organization. APOLLO has 150 full-time software professionals developing off-the-shelf software applications for various industries ranging from manufacturing, food and beverages, car sales, etc. Software development in the company follows an iterative spiral life-cycle model and a typical software product is released after every 6 months of development.

Note: The running scenario will appear at the end of each section enclosed within a textbox similar to this one.
5.1 Initialization stage

The initialization is the first stage in SPI-MEF and takes place before organizations start to implement SPI initiatives in their current process (Figure 11). This is actually the setup stage that gathers the pre-requisites in information that are needed for evaluations. Basically, this stage helps to characterize two things: first, data about the SPI initiative itself is collected. Second, information on the organizational context is gathered in order to determine the targeted evaluation quality.

5.1.1 Gather SPI initiative information

In this step, important information about the SPI initiative that is relevant for the purpose of evaluation is gathered and documented. SPI-MEF proposes a template form called the SPI Initiative Form (Table 5) that contains some basic fields that are needed to characterize the SPI initiative that is going to be implemented. The current template form captures the most basic information that is usually present in most SPI initiatives and thus can be adapted by organizations to fit their specific needs.

<table>
<thead>
<tr>
<th>SPI Initiative Form</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of SPI Initiative</td>
<td>&lt; SPI Framework / Practices / Tools &gt;</td>
</tr>
<tr>
<td>Name</td>
<td>&lt; Name of the SPI initiative &gt;</td>
</tr>
<tr>
<td>Description</td>
<td>&lt; Description of the SPI initiative that states the purpose, what the initiative does and how it is implemented. In case of an SPI framework that consists of more than one practice, all the single practices have to be listed out. &gt;</td>
</tr>
<tr>
<td>Improvement Goal(s)</td>
<td>&lt; The goal(s) or aim(s) of the SPI initiative. &gt;</td>
</tr>
<tr>
<td>Process Areas / Phases</td>
<td>&lt; The process areas or developmental phases targeted by the SPI initiative. E.g. requirements engineering, verification and validation, design, coding, etc. &gt;</td>
</tr>
<tr>
<td>Target Entities for Implementation</td>
<td>&lt; The target entities for the SPI initiative implementation. E.g. specific projects, products, departments, divisions or even organization-wide. &gt;</td>
</tr>
<tr>
<td>Implementation Schedule</td>
<td>&lt; The tentative schedule of the implementation. The schedule can be broken down into phases depending on context of the organization. &gt;</td>
</tr>
</tbody>
</table>
The second aspect that is important to characterize is the organizational context in terms of the quality of the current evaluation process. This step is needed in order to determine how well the organization can conduct the measurement and the following evaluation, i.e., to assess the organizations’ measurement and evaluation capability. Currently, SPI-MEF does not prescribe a specific method, but a gap analysis technique seems to be the most appropriate approach. In general, gap analysis uses two sets of information: the current status and the aspired status [96]. The current status of measurement capability can be determined by following one of the approaches presented in [97] [98] [99] [100]. Then it has to be determined which is the aspired measurement and evaluation quality. The gap then shows which refinements in the measurement program need to be implemented. Since measurement capability assessment and implementation of measurement programs are currently not the core focus of SPI-MEF, these topics are not discussed in further detail.

However, SPI-MEF proposes to use a “2 x 2” matrix to support the decision process which long-term evaluation strategy should be followed. As shown in Figure 12, the evaluation quality in SPI-MEF is defined by two dimensions: accuracy and coverage. Accuracy is improved by considering confounding factors (Section 5.2.3.3), primary and complementary measures (Section 5.2.2.1) and by applying cross-examination measures (Section 5.2.2.3). Coverage is determined by what extent measurement levels and viewpoints

---

**Gather SPI initiative information**

The current software processes in APOLLO is still immature and it offers several opportunities for improvement. Lately, the President of APOLLO has voiced out his interest in alleviating the product quality of the company. Therefore, a small process team has been set up to investigate the pressing process area that needs immediate attention to bring up the first SPI initiative to the company. After one month of investigation, the process team has proposed to introduce code inspection in the software development. The *SPI Initiative Form* for code inspection is documented as below:

<table>
<thead>
<tr>
<th>SPI Initiative Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of SPI Initiative</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>Description</td>
</tr>
<tr>
<td>Improvement Goal(s)</td>
</tr>
<tr>
<td>Process Areas / Phases</td>
</tr>
<tr>
<td>Target Entities for Implementation</td>
</tr>
<tr>
<td>Implementation Schedule</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

5.1.2 Gap analysis of evaluation quality

The second aspect that is important to characterize is the organizational context in terms of the quality of the current evaluation process. This step is needed in order to determine how well the organization can conduct the measurement and the following evaluation, i.e., to assess the organizations’ measurement and evaluation capability. Currently, SPI-MEF does not prescribe a specific method, but a gap analysis technique seems to be the most appropriate approach. In general, gap analysis uses two sets of information: the current status and the aspired status [96]. The current status of measurement capability can be determined by following one of the approaches presented in [97] [98] [99] [100]. Then it has to be determined which is the aspired measurement and evaluation quality. The gap then shows which refinements in the measurement program need to be implemented. Since measurement capability assessment and implementation of measurement programs are currently not the core focus of SPI-MEF, these topics are not discussed in further detail.

However, SPI-MEF proposes to use a “2 x 2” matrix to support the decision process which long-term evaluation strategy should be followed. As shown in Figure 12, the evaluation quality in SPI-MEF is defined by two dimensions: accuracy and coverage. Accuracy is improved by considering confounding factors (Section 5.2.3.3), primary and complementary measures (Section 5.2.2.1) and by applying cross-examination measures (Section 5.2.2.3). Coverage is determined by what extent measurement levels and viewpoints

---

32
are included in the evaluation (Section 5.2.1). These two dimensions can be seen as a typical dilemma: accuracy promotes depth in the evaluation while coverage stands for breadth, and it has to be decided which of those are more important to the organization. Clearly, it is possible to address both dimensions simultaneously, but cost and effort constraints may prohibit such a strategy. The “2 x 2” matrix has been proven to be an excellent tool to address such decision dilemmas [101]. The roles involved in the discussion about the long-term strategy should include management which provides the funding for the improvement initiative and measurement program experts. The Evaluation Opportunity Matrix Template (Figure 12) serves as a starting point for the discussion.

![Evaluation Opportunity Matrix](image)

The evaluation opportunity matrix (“2 x 2”) is divided into four regions:

i. **Usual practice (Low accuracy and narrow coverage)**
   This region depicts the usual practice that many organizations follow [28], in which the evaluation is only conducted at certain measurement levels, e.g. process and project, and the accuracy aspect is rarely considered.

ii. **Accurate evaluation (High accuracy and narrow coverage)**
   The accuracy aspect is prioritized over the coverage in this region, where evaluating the right measures is perceived as more important than evaluating a wider scope of measures.

iii. **Wide evaluation (Low accuracy and wide coverage)**
   The characteristic of this region is the opposite of the previous region (ii), where having a wider scope of evaluation is considered as more valuable than the accuracy of evaluation.

iv. **Holistic evaluation (High accuracy and wide coverage)**
   The holistic evaluation is the ultimate region that is best at both aspects; accuracy and coverage. However, the cost involved might be higher; more discussion about cost is given in the next paragraph.
Aside from accuracy and coverage another important aspect to consider is the cost of the evaluation. Cost is denoted as a function $f$ into the quality and scope of the evaluation. The relationship between cost and quality is assumed to be proportional, i.e. achieving higher quality of evaluation will incur higher costs. Therefore, the resources an organization is willing to invest have to be taken into consideration when choosing the desired path to improve the quality of evaluation.

The cost of evaluation can arise from the amount of metrics which, are defined and need to be collected, the resources needed to manage the metrics, the number of people involved in the metric collection and evaluation, etc. Although achieving high accuracy and coverage in the context of SPI-MEF seems to inherently require more metrics, the potential reuse of metrics should be considered when evaluating the cost for evaluation implementation. During the decision process which strategy to follow, it is advisable to involve personnel who has expertise in implementing a measurement program and can estimate the cost of metric collection, and management, which sponsors the improvement program.

**Gap analysis of evaluation quality**

The process team had a meeting with the upper level management and some key people who are in-charge of the existing measurement program in APOLLO to discuss the gap analysis. After considering the budget for the process improvement, the company’s short-term goal is to focus on coverage first by addressing the Process, Project and Product measurement levels. To increase accuracy, it is targeted to consider primary and complementary measures and to control the major confounding factors by establishing a baseline from the appropriate historical data.
5.2 Design stage

The design stage (Figure 13) is regarded as the core part of SPI-MEF because the key concepts that constitute the model are mostly applied in this stage. Taking inputs from the initialization stage, the main goal of this stage is to create an evaluation blueprint that is tailored to the organization’s context and needs. In order to achieve this goal the design stage defines four steps: evaluation scoping, determination of measures, selection of evaluation methods and evaluation scheduling.

5.2.1 Evaluation scoping

The purpose of this step is to define the scope of evaluation by defining the measurement levels where data can be collected and by identifying the stakeholders who are interested to see the evaluation of the improvement. The strategy chosen in the previous step, gap analysis of evaluation quality (Section 5.1.2), defines to which extent measurement levels and viewpoints are covered.

The measurement levels (Process, Project, Product, Organization, and External) represent the measurable entities that can be assessed to evaluate the software process

---

**Tips for gap analysis of evaluation quality**

- It is better to start small and grow gradually in any measurement program. The path to be taken in the evaluation opportunity matrix should not be a drastic, simultaneous change in both aspects.

- The path in the evaluation opportunity matrix depends upon the scope of the implementation. If the SPI initiative is only implemented in a small number of projects, the impact on the Organization and External level might not be visible.

- An SPI initiative can be implemented in phases, starting from pilot projects, which can expand to a larger scope in the organization. The path in the evaluation opportunity matrix should consider the implementation details and roll-out phases of specific SPI initiatives.

*Note:* Tips for some evaluation steps will be provided in text boxes similar to this, and they represent the authors’ personal opinion in assisting the evaluation process.
improvement (see [28] for a detailed discussion). The following list briefly characterizes those levels:

i. **Process**: Measurement of process efficiency (e.g. in terms of time and resources) and effectiveness (e.g. by assessing certain properties of the implemented process).

ii. **Project**: Measurement of project characteristics to assess the effect of the improvement (typical characteristics are cost, schedule, productivity and estimation adherence).

iii. **Product**: Measurement of product attributes (e.g. quality, cost and time-to-market)

iv. **Organization**: Measurement of the organizational effects of process improvement (e.g. economical benefits and effects on employees)

v. **External**: Measurement of externalities of the improvement (e.g. effects on the customers’ business and the society)

SPI-MEF defines three generic viewpoints that represent the stakeholders who are interested to see the results of the improvement:

i. The **Implementer** viewpoint reflects all stakeholders and roles that are involved in the software development and in the process improvement by implementing the changes [28].

ii. The **Coordinator** viewpoint reflects all stakeholders and roles that coordinate and control the software development and specifically the process improvement initiative [28].

iii. The **Sponsor** viewpoint reflects all stakeholders and roles that motivate and fund the improvement initiative; they are particularly interested in evaluating the improvement regarding its costs and benefits [28].

The scope of the evaluation, which is known as the *Evaluation Area* (Figure 14), is defined by considering both the measurement levels and viewpoints; both serve as the drivers for determining the appropriate measures (see Section 5.2.2).

![Figure 14: Evaluation area (inspired by [28])](image)
SPI-MEF provides a template table called the *Evaluation Area Table* (Table 6) that can be used to populate the viewpoints in their respective measurement levels. A typical evaluation area with exemplary stakeholders is shown in Table 6.

Table 6: *Evaluation Area Table* template with example of stakeholders

<table>
<thead>
<tr>
<th>Measurement Levels</th>
<th>Viewpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Implementer</strong></td>
</tr>
<tr>
<td>Process</td>
<td>Development teams (Programmer, Tester, etc)</td>
</tr>
<tr>
<td></td>
<td>Development teams (Programmer, Tester, etc)</td>
</tr>
<tr>
<td></td>
<td>Project teams (Project managers, development teams, etc)</td>
</tr>
<tr>
<td></td>
<td>Project teams (Project managers, development teams, etc)</td>
</tr>
<tr>
<td></td>
<td>Product departments (Product development teams, etc)</td>
</tr>
</tbody>
</table>

**Evaluation scoping**

By considering the key people that are involved and interested in the implementation of the SPI initiative, the process team in APOLLO has come out with the *Evaluation Area Table* shown below. The measurement levels that are involved are Process, Project and Product.

<table>
<thead>
<tr>
<th>Measurement Levels</th>
<th>Viewpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Implementer</strong></td>
</tr>
<tr>
<td>Process</td>
<td>Development team (Analyst, programmer and tester)</td>
</tr>
<tr>
<td>Project</td>
<td>Development team (Analyst, programmer and tester)</td>
</tr>
<tr>
<td>Product</td>
<td>Development team (Analyst, programmer and tester)</td>
</tr>
</tbody>
</table>
5.2.2 Determination of measures

This is the major step in the design stage where the required metrics for the evaluation are derived. This crucial step is built up by considering a few key concepts that specifically tackle the aspects in determining the appropriate measures for improvement evaluation.

5.2.2.1 Primary and complementary measures

It is common that SPI initiatives are evaluated by looking only at the gained benefits, considering the targeted improvement goal. However, the impact of SPI initiative may go beyond the expected outcome, i.e. certain side-effects may change the picture of the improvement and it is necessary to assess them to evaluate the true effects of the implemented change. In order to highlight this idea, SPI-MEF categorizes the measures used in the evaluation into two types, namely the primary and complementary measures (Figure 15):

i. **Primary measurements** are the set of measurements that are instrumented to assess the achievement of the improvement goal [28].

ii. **Complementary measurements** are the set of measurements that are instrumented to assess the effects of the process change that are not directly related to the improvement goal [28]. They assess the potential side-effects caused by the improvement initiative. They are not to be confused with optional or “good-to-have” measurements and they are as important as the primary measurements.

![Figure 15: Primary and complementary measures (inspired by [28])](image)

The aim of categorizing the measures into these two categories is to create a mindset that supports the elicitation of appropriate measures to evaluate the improvement. The “Iron Triangle” metaphor (Figure 16) is used to illustrate the relationship between primary and complementary measurements and helps to derive the needed measures [28]. In project
management, the iron triangle is used to visualize the challenge to optimize the three major attributes cost, quality and schedule contemporaneously.

![Iron Triangle Diagram](image)

Figure 16: The famous iron triangle metaphor (inspired by [28] and [102])

For example, one can try to optimize schedule and cost, but most probably quality will then be negatively affected. The same idea is applied to the example shown in Table 7. Defect density is the primary measure since it is directly related to the improvement goal (increase product quality). Effort in man-hours and project cycle time are complementary measures which need to be monitored in order to make the improvement an overall success. The example is based on the project measurement level, but the same principle can be applied on every measurement level.

<table>
<thead>
<tr>
<th>SPI initiative:</th>
<th>Peer review</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvement goal:</td>
<td>Improve product quality</td>
</tr>
<tr>
<td>Measures:</td>
<td></td>
</tr>
<tr>
<td>Success indicator:</td>
<td>Quality (Defects)</td>
</tr>
<tr>
<td>Metric:</td>
<td>Defect density</td>
</tr>
</tbody>
</table>

Table 7: Determination of measures using iron triangle metaphor

In order to ease the application of the iron triangle metaphor, a list of success indicators for each measurement level would be helpful. A pool of metrics that is grouped according to different success indicators can be found in [28]. The list is called the List of Success Indicators and Metrics and it serves as one of the support artifacts provided by SPI-MEF in the design stage. Organizations have the liberty to expand and adapt the list suiting their specific needs. The list can be found in Appendix B.

5.2.2.2 Derivation of measures using GQM

The previously introduced concepts provide additional information that facilitates the usage of the GQM paradigm. The GQM approach is a systematic way to tailor and integrate an organization’s objectives into measurement goals and refine them into measurable values (for a more information about GQM, refer to Section 2.3.1). It provides a template for defining measurement goals and guidelines for top-down refinement of measurement goals into questions and then into metrics, and a bottom-up analysis and interpretation of the collected data [60]. SPI-MEF provides an interface with the conceptual level of GQM to define the appropriate measurement goals for improvement evaluation (Figure 17). Note that in Figure 17, “Evaluate” is a constant keyword to be used in the GQM that maps to the “Purpose” of the measurement goal.
By using the input from SPI-MEF, it is possible to define very specific measurement goals which facilitate the elicitation of the appropriate questions and right metrics to answer those questions. SPI-MEF introduces the *GQM Construction Table* (Table 8) template that can be used to populate the elements of SPI-MEF onto the different facets of GQM.

**Table 8: GQM Construction Table template**

<table>
<thead>
<tr>
<th>Purpose:</th>
<th>Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Goal (Improvement Goal):</td>
<td>&lt; State the improvement goal. If there is more than one improvement goal, separate the improvement goals into different table &gt;</td>
</tr>
<tr>
<td>Object of Study:</td>
<td>&lt; SPI initiative &gt;</td>
</tr>
<tr>
<td>Context:</td>
<td>&lt; State the scope or target entities for the SPI initiative implementation &gt;</td>
</tr>
<tr>
<td>Measurement Level</td>
<td>Point of View (Viewpoint)</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Process</td>
<td>&lt; Stakeholders per viewpoint &gt;</td>
</tr>
<tr>
<td></td>
<td>.....</td>
</tr>
<tr>
<td>External</td>
<td>.....</td>
</tr>
</tbody>
</table>

The *GQM Construction Table* only acts as a bridge that pulls together the elements from SPI-MEF to produce an instance of GQM for the purpose of evaluating an SPI initiative. The final output of the table will lead to the derivation of specific measurement goals. With the input from the *GQM Construction Table*, specific measurement goals along with its questions and metrics will be derived using the *Measurement Goal, Question(s) and Metrics*
*Template* (Table 9). Practitioners can refer to the Appendix B to get to specific metrics for the evaluation as it provides a list of commonly used metrics reported in the industry.

<table>
<thead>
<tr>
<th>Measurement goal ID</th>
<th>Success indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Measurement Goal will be constructed using the input from the GQM Construction Table</td>
<td></td>
</tr>
</tbody>
</table>

**Question 1:** < The question that helps to derive the necessary metrics to answer the measurement goal >

**Metrics:**
- < Metric ID 1 >:
- < Metric Name >
  - < The description of the metric that includes the purpose, formula/algorithm, data collection and etc >

- < Metric ID n >:
  - ............................................................................................................................

**Question <n>:**
- ........................................................................................................................................
**Determination of measures (Part 1)**

In each measurement level, the stakeholders (viewpoints) who are involved in the derivation of measures discussed and agreed on the primary and complementary (if any) foci (success indicators) for the evaluation of code inspection. The complete *GQM Construction Table* that is populated for all the measurement levels is shown below:

<table>
<thead>
<tr>
<th>Purpose:</th>
<th>Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Business Goal (Improvement Goal):</strong></td>
<td>BG: Improve product quality</td>
</tr>
<tr>
<td><strong>Object of Study:</strong></td>
<td>Code inspection</td>
</tr>
<tr>
<td><strong>Context:</strong></td>
<td>All projects in the organization</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Level</th>
<th>Point of View (Viewpoint)</th>
<th>Category</th>
<th>Focus (Success Indicator)</th>
<th>Measurement Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Process</strong></td>
<td>Process team</td>
<td>Primary</td>
<td>Effectiveness</td>
<td>MG01_PRC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complementary</td>
<td>Efficiency</td>
<td>MG02_PRC</td>
</tr>
<tr>
<td></td>
<td>Development teams</td>
<td>Primary</td>
<td>Effectiveness</td>
<td>MG03_PRC</td>
</tr>
<tr>
<td></td>
<td>Head of development</td>
<td>Primary</td>
<td>Effectiveness</td>
<td>MG04_PRC</td>
</tr>
<tr>
<td><strong>Project</strong></td>
<td>Project manager</td>
<td>Primary</td>
<td>Defects</td>
<td>MG01_PRJ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complementary</td>
<td>Cost</td>
<td>MG02_PRJ</td>
</tr>
<tr>
<td></td>
<td>Development teams</td>
<td>Primary</td>
<td>Defects</td>
<td>MG03_PRJ</td>
</tr>
<tr>
<td></td>
<td>Head of development</td>
<td>Primary</td>
<td>Defects</td>
<td>MG04_PRJ</td>
</tr>
<tr>
<td><strong>Product</strong></td>
<td>Product manager</td>
<td>Primary</td>
<td>Quality</td>
<td>MG01_PRD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complementary</td>
<td>Cost</td>
<td>MG02_PRD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Time-to-market</td>
<td>MG03_PRD</td>
</tr>
<tr>
<td></td>
<td>Development teams</td>
<td>Primary</td>
<td>Quality</td>
<td>MG04_PRD</td>
</tr>
<tr>
<td></td>
<td>Head of development</td>
<td>Primary</td>
<td>Quality</td>
<td>MG05_PRD</td>
</tr>
</tbody>
</table>
Determination of measures (Part 2)

Using the Measurement Goal, Question(s) and Metrics Template, the measurement goals, questions and metrics for code inspection in APOLLO were derived; an extract of the produced artifact for the Process level is shown below:

MG01_PRC: Effectiveness

Evaluate the process of code inspection with respect to its effectiveness on pre-test defect detection in project from the viewpoint of the process team.

**Question 1:** What is the number of defects found by the used code inspection method?

**Metrics:**

**M01_PRC:**

Percentage of defects found by inspection

Percentage of defects found by inspection = I / T  
I = Number of defects found during inspection  
T = Total number of defects found in code inspection, unit test, integration test, system test and acceptance test.

This metric shows the effectiveness of code inspection by evaluating the percentage of defects found during code inspection with respect to the total defects found in all verification and validation activities.

**M02_PRC:**

Phase Containment Effectiveness (PCE) for coding phase

PCE for coding phase = F / (F + E)  
F = Number of defects found in coding phase  
E = Number of defects escaped from coding phase

This metric shows the phase containment effectiveness for coding phase. Conducting code inspection is expected to contain more defects from escaping to the next phase during development.

The metrics derived from the GQM need to be documented in order to be easily usable for the subsequent evaluations. SPI-MEF proposes a template form called the Metric Form (Table 10) for this purpose. This form will also be populated in the next evaluation step (Section 5.2.3) to describe the evaluation method, metrics baselining and for the consideration of confounding factors.

**Table 10: Metric Form template**

<table>
<thead>
<tr>
<th>Metric Form</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric ID</td>
<td>&lt; The metric ID from the GQM &gt;</td>
</tr>
<tr>
<td>Metric Name</td>
<td>&lt; The metric name &gt;</td>
</tr>
<tr>
<td>Success Indicator</td>
<td>&lt; The success indicator &gt;</td>
</tr>
<tr>
<td>Measurement Goal (Viewpoint)</td>
<td>&lt; List out all the measurement goals together with its viewpoint that this metric is trying to address &gt;</td>
</tr>
<tr>
<td>Measurement Level</td>
<td>&lt; Process / Project / Product / Organization / External &gt;</td>
</tr>
</tbody>
</table>
| Algorithm / Description | < The formula or equation of the metric >  
< The description of the metric > |
5.2.2.3 Cross-examination

Cross-examination is one of the elements for accuracy in the evaluation opportunity matrix (see Figure 12). It is a strategy in software measurement that assesses a particular success indicator by using more than one measure in order to increase the confidence level in the validity of the evaluation [28].

In deriving the metrics using GQM, more than one metric that address the same success indicator can be determined. As pointed out in [28], cross-examination can be achieved using methodological triangulation, in which different data collection mechanisms are involved. Cross-examination can be exemplified in the evaluation of the product level. In evaluating the product quality, two types of measures can be used; the first is calculating the number of post-release defects found in operation (quantitative), while the second one is evaluating the product quality using survey (qualitative) to get the customer satisfaction. These two different ways in evaluating the product quality is considered as cross-examination.
5.2.3 Selection of evaluation methods

After the specific metrics for the evaluation of the SPI initiative are derived, the next task is to define the methods that can be used to carry out the evaluation. The evaluation methods, metrics baselining and confounding factors will be discussed in this section.
5.2.3.1 Metrics baselining

Baselining the value for the previously elicited metrics is essential because the baselines will serve as the initial point for evaluating the improvement. An accurate baseline is very important because the evaluation on the degree or magnitude of improvement depends upon the baseline value. There are various ways how organizations can set the baselines for their metrics.

The most typical way would be creating the baseline from historical data collected from previously conducted processes or projects and already finished products. Some derived metrics that consist of two or more elementary measurements can sometimes be easily acquired from historical data.

If there is no historical data available, the baseline can be obtained by start collecting data from active projects that are currently running in the organization. The data collected from the active projects would then serve as the baseline to evaluate the projects that are going to incorporate the SPI initiative.

There is also another option that depends upon the scope of SPI initiative implementation. If the SPI initiative is implemented partially in the company, i.e. for specific products, departments, etc., the baseline can be established at the same time with the SPI initiative implementation. The baseline can be obtained from projects that are running in parallel and be used for comparison with the target projects that implement the SPI initiative.

For a more in-depth discussion on baseline establishment, refer to Section 8.4.2.1 in [28] which also provides several references to publications which address baselines in general and also presents advanced baseline construction techniques.

5.2.3.2 Evaluation method

The improvement evaluation model proposed in [28] defines four general evaluation methods:

i. Basic comparison

This is the most commonly used evaluation method because it is the simplest and most straightforward way for evaluating an improvement. The evaluation is performed by comparing the change in a particular metric’s value attributed to the improvement initiative. The metric’s value before the process change (baseline) is compared against a new value measured after the SPI initiative has been implemented for a period of time.

ii. Statistics-based analysis

The evaluation of SPI initiatives can also be conducted by using descriptive and inferential statistics. Analyzing the metrics using statistics-based methods such as graphical charts, numerical sums and control charts enable a more elaborative evaluation outcome.

iii. Survey

A survey is defined as any method to collect, compare and evaluate quantitative or qualitative data from human subjects, and it is usually conducted using interviews or questionnaires [28]. Survey is a common evaluation method for assessing qualitative improvements such as customer and employee satisfaction.
iv. **Cost-benefit analysis**

Cost-benefit analysis is an evaluation method that targets the assessment of the economic value of an SPI initiative. This method is widely used in calculating the return on investment of SPI initiatives to justify to sponsors the investment put forth by assessing the benefits gained with respect to the cost of an SPI initiative.

To deepen the understanding of the above described evaluation methods refer to Section 8.4.2 in [28]. There, the evaluation methods are discussed in detail and references to their practical implementation in industrial cases are given.

5.2.3.3 **Confounding factors**

Confounding factors represent a fundamental threat for the evaluation of a process improvement initiative if any kind of comparison is used to assess its effects. A comparison is said to be confounded if the observed difference between two values (the effect of a treatment) is not solely caused by the treatment alone but can be partially attributed to an external factor [28]. Figure 18 shows how this affects the evaluation of software process improvement initiatives. The target entity of the initiative is usually one or a set of projects or products. Confounding factors hide, distort or amplify the measured effects in the target entity and the improvement cannot be accurately assessed. The challenge in evaluating SPI initiatives is therefore to identify and possibly control confounding factors in order to achieve valid results.

![Confounding factors in the evaluation of SPI initiatives](adapted from [28])

SPI-MEF aims to create awareness of the potential threats caused by confounding factors and, since they are dependent on the context in which the improvement initiative is carried out, exemplify those which are most commonly encountered.

Typical confounding factors that should be taken into consideration for evaluation planning or during the evaluation:

i. **Project type**

New development and enhancement/maintenance projects have different properties and they should not be treated as the same during evaluation, i.e. comparison of success indicators from different project types should be avoided.

ii. **Development model**

Different project development life-cycles such as waterfall model and spiral model have different project characteristics and potentially confound the evaluation.

iii. **Product size and complexity**

Size (lines of code, function point, etc.) and complexity (number of features, number of control flow, etc.) of the product have to be taken into consideration during the
evaluation. Some metrics can be normalized to the product size to control this confounding factor.

iv. **Product domain**
The product domain difference also affects the evaluation. Front-end applications, server-side systems and embedded software are different types of product domains that should not be put on par during the evaluation.

v. **Employee factors**
The staffs working in the project might have a different experience level and some metrics such as productivity metrics and efficiency metrics should take the staff experience level into consideration. In addition to that, the employee turnover in the organization also affects the evaluation.

vi. **Technology factors**
Technological factors such as the programming language and tool support can influence indicators like productivity and effort.

vii. **Process compliance**
The degree to which the standard process is followed in the actual implementation should be considered in the evaluation as this can give indications to what extent the improvement can be attributed to the SPI initiative.

viii. **Evaluation sampling**
The sample of target entities for evaluation can affect the validity of evaluation results. The sampling size and technique shall be taken into consideration during the evaluation.

ix. **Time factors**
Time can be seen as a factor that can affect the evaluation result. When conducting a customer survey on product quality, the total usage time of the product has to be considered in the evaluation. Further details about the time to evaluate will be discussed in the next section (Section 5.2.4).

x. **Subjective evaluation**
Surveys are prone to subjective evaluations because different subjects/people have different level of expectation depending on the role and their personal preference.

xi. **Instrumentation methods**
The data collection methods such as automated vs. human intervention methods would also affect the validity of evaluation results.

xii. **Multiple improvement initiatives**
It is difficult to ensure that a particular improvement is attributed to a specific SPI initiative. Several improvement initiatives that run in parallel would create traceability issues in the evaluation. For example, when calculating the cost saving from a specific improvement initiative, care should be taken not to count the saving twice as the saving might also be attributed to another improvement initiative.

There exist several techniques to control confounding factors. The most basic one (and also used in industrial practice, see Section 7.2.6 in [28]) is to group the to-be-compared entities (processes, projects or products) into categories with similar properties.
Tips to address confounding factors

- When selecting target entities (processes, projects or products) for basic comparison, try to pay attention for selecting target entities that have similar properties in project type, development model, product domain, staffs’ experience level and development environment.

- In order to address confounding factors related to size or complexity, the metrics used for evaluation can be normalized with the size or complexity of the target entities.

- To increase validity and confidence in traceability of the cause-and-effect from the SPI initiative, the degree of process compliance needs to be measured in the actual implementation of the SPI initiative.

- When selecting target entities for evaluation (in case of multiple target entities), the best representative of the target entities should be given higher priority. Criteria such as investment unit, size or complexity, number of employees and project type can be considered in the selection.

- The “burn-in” phase for the improvement should be considered when determining the time to evaluate (Section 5.2.4).

- To address the subjectivity of surveys, a larger sample (target entities) should be taken from the population. The sampling process should also be carefully conducted so that a diversified sample is obtained, i.e. the sample should be randomized.

- In order to provide traceability of improvement from multiple SPI initiatives, the timeline of each SPI initiative should be documented so that if a significant change in improvement is detected, the timelines can be analyzed to attribute the observed improvement to the specific SPI initiative.
5.2.4 Evaluation scheduling

The determination of the appropriate time to evaluate the improvement is important [28]; however, it is difficult to suggest a general rule which is universally applicable to determine the suitable time. Therefore, SPI-MEF introduces the concepts of Lag Factor (LF) and Degradation Factor (DF) which support the practitioner to determine the time bounds for an
improvement evaluation. The LF is the minimum time required for the improvement initiative to propagate a visible impact on the measurable entities (measurement levels). The DF, on the other hand, is the maximum time for which an evaluation result can be considered as valid, or, more specifically, can be considered as representing the actual status of the improvement [28].

The value of LF and DF can be defined by an organization based on the type of the SPI initiative taken, the organization’s experience and historical data. Driven by these two factors, SPI-MEF proposes a periodic improvement evaluation, that is, an evaluation indexed by time to see the trend of the improvement.

SPI-MEF proposes for practitioners a template called the Evaluation Scheduling Form (Table 11) to document the scheduling information of SPI initiative evaluations.

Figure 19: Illustration of lag factor and degradation factor

Figure 19 illustrates the concepts of Lag Factor and Degradation Factor. LF and DF are defined as four months and six months respectively for a certain SPI initiative at the Project level. E1 is the evaluation at the first point in time (four months after the introduction of the SPI initiative) when the impact of the SPI initiative is assumed to be visible at the Project level. As DF is estimated as six months, E1 is assumed not to be a valid representation of the improvement anymore (after the 10th month). That is the reason why the organization needs to conduct another evaluation (E2) to get the current representation of improvement. After the 8th month, another evaluation (E3) is conducted within the DF time frame.

SPI-MEF proposes for practitioners a template called the Evaluation Scheduling Form (Table 11) to document the scheduling information of SPI initiative evaluations.

Table 11: Evaluation Scheduling Form template

| Duration: | < The duration of the implementation of SPI initiative > |
| Context: | < The scope of the target entity that the SPI initiative is to be implemented > |
| Time to evaluate / Periodic evaluation: | |
| Measurement Levels | Lag Factor (LF) | Degradation Factor (DF) | Time to Evaluate |
| Process | < State the value of LF and briefly describe how the value is derived > | < State the value of DF and briefly describe how the value is derived > | < Considering the LF and DF, state the time bound for evaluations > |
| .......... | .......... | .......... | .......... |
| External | .......... | .......... | .......... |

51
**Evaluation scheduling**

APOLLO has decided to roll out the code inspection practice in two phases, where the first phase is trying out the practice in pilot projects while the second phase expands the implementation throughout the development projects within the organization. The evaluation scheduling details for the pilot implementation are shown below:

<table>
<thead>
<tr>
<th>Duration:</th>
<th>Starting from Jan 2008 until Jun 2008 (6 months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context:</td>
<td>There are two pilot projects that have been selected for the evaluation. The two projects have the following schedule:</td>
</tr>
<tr>
<td></td>
<td>ii) Project II: 14th February 2008 – 30th May 2008</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time to evaluate / Periodic evaluation:</th>
<th>Measurement Levels</th>
<th>Lag Factor (LF)</th>
<th>Degradation Factor (DF)</th>
<th>Time to Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>4 months</td>
<td>Since this is a pilot implementation, DF does not apply.</td>
<td>After the first four months from the SPI initiatives introduction.</td>
<td></td>
</tr>
<tr>
<td>Project</td>
<td>4 months</td>
<td>Since this is a pilot implementation, DF does not apply.</td>
<td>After the first four months from the SPI initiatives introduction.</td>
<td></td>
</tr>
</tbody>
</table>

Pilot projects' evaluation schedule:
5.3 Implementation stage

The implementation stage (Figure 20) is where the evaluation blueprint created during the design stage is put into realization. By referring to the evaluation plan, the derived metrics will be incorporated into the existing measurement program (if any) in the organization and the evaluations are conducted as scheduled.

![Figure 20: SPI-MEF implementation stage](image)

5.3.1 Establishment of a measurement program

Organizations incorporate the metrics derived in the design stage into their existing measurement program. In certain company settings, some metrics are already used and collected throughout their developments. Therefore, it is practical to assess the current measurement program in the organization in order to identify metrics that can be reused for the improvement evaluation purpose.

During the implementation of SPI initiative, data for the defined metrics should be collected carefully so that the validity of the measurements is not compromised. A measurement workshop can be conducted to transfer the basic knowledge on measurement and train the users on the use of organizational policies, standards, and procedures for measurement [103]. In order to have a smooth evaluation process later on, support systems such as database applications and statistical tools are recommended so that data collection and data manipulation can be done with ease [61]. It would also be useful to adopt some common lessons learnt and best practices from industries to support the implementation of the measurement program [104]. Further suggestions and experiences for implementing effective measurement programs are given in [105].

5.3.2 Conduct evaluation

According to the planned schedule, the evaluations are conducted at the specified date. Different measurement levels will have their evaluation conducted separately depending on the planned schedule. Before the evaluation, the to-be-evaluated target entities (target projects, specific products, etc.) have to be selected and the evaluator has to be assigned. The proposed evaluator specified in the Metric Form (Table 10) can be referred when assigning the evaluator to conduct the evaluation.

During the evaluation, the actual values of metrics are collected and evaluated using the evaluation method specified in the Metric Form. The metric is evaluated with the respect to the baseline and as a result of the evaluation, the measurements are classified as either “Improvement gained”, “No significant impact” or “Loss recorded” (see Figure 21). The deviation range of metric values for the improvement classification has to be defined by organizations to ensure a consistent evaluation. The Metric Form contains a field called
“Deviation Range” for this purpose and it documents the variance of a metric value around its baseline which represents neither significant improvement nor decline. This field can be defined by a percentage and the viewpoints have to agree on a unique value for each metric. A stakeholder with expertise on the particular metric should be consulted to define the value for this field.

SPI-MEF provides an Evaluation Result Template (Figure 21) that can be used to present the evaluation results visually. The template can be viewed as a “dashboard” with three regions colored differently to show the status of the evaluated metric:

i. **Green (“Improvement gained”)**
   If the value of the evaluated metric shows an improvement with respect to the baseline value, then the “dashboard’s needle” will point to this region.

ii. **Yellow (“No significant impact”)**
   If the value of the evaluated metric lies within the deviation range, then the “dashboard’s needle” will point to this region.

iii. **Red (“Loss recorded”)**
    If the value of the evaluated metric records a loss with respect to the baseline value, then the “dashboard’s needle” will point to this region.

![](image)

**Figure 21: Evaluation Result Template**

The position of the “dashboard’s needle” in the respective regions have different meanings. Figure 22 illustrates these differences:

1. If the evaluated metric has the same value as the baseline value, the “dashboard’s needle” will point exactly to the middle of the “Yellow” region.
2. Within the “Yellow” region, if the “dashboard’s needle” points to the left side of the region, this indicates that the evaluated metric value is near to recording a loss.
3. Within the “Yellow” region, if the “dashboard’s needle” points to the right side of the region, this indicates that the evaluated metric value is near to showing an improvement.
4. The “dashboard’s needle” falls in the “Red” region and this indicates that a loss is recorded.
5. The “dashboard’s needle” falls in the “Green” region and this indicates that an improvement is observed.

6. Within the “Green” region, if the “dashboard’s needle” points to the far right side of the region, this indicates that a very significant impact is observed. The same way of reading applies to the “Red” region as well. The determination of where the “dashboard’s needle” fall within each region is judged by the proposed evaluator who has the knowledge on the evaluated metric.

Figure 22: Interpretation of different evaluation results

Conduct evaluation

APOLLO has conducted the evaluations for the pilot implementation of code inspection. The visual presentation below shows the evaluation results of three metrics that have been evaluated at the Process and Project level in the second evaluation conducted on 1st June:

From the figure above, the effectiveness of code inspection shows a positive improvement while the efficiency and cost both recorded no significant impact.
5.3.3 Holistic view creation

The final outputs from the previous step are the individual evaluation results for each metric. If the number of metrics grows, evaluating the SPI initiative by judging the evaluation results of each single metric might be difficult. Therefore, SPI-MEF tries to consolidate the evaluation results by proposing a “one-view” representation of the overall improvement evaluation results. In this way, the impact of the improvement initiative can be presented in a compact and easy to communicate form.

5.3.3.1 Holistic view components

The following components are used to construct the holistic view:

i. Subjective weight (SW)

Each viewpoint (that is, each stakeholder of the improvement initiative grouped by coordinator, sponsor and implementer) gives a weight of subjective importance to every measurement. The weights sum up to 1.

\[
\sum_{i}^{n} SW_i = 1
\]

ii. Impact rating (IR)

Each viewpoint rates the impact of the improvement initiative on the respective measurement according to an 11 point Likert Scale shown as in Figure 23. The reason for choosing a relative high number of response categories is that low numbered scales have generally less valid scores and have less discriminating power than scales with six or more response categories [106]. Since the impact rating of the improvement is subjective in nature, organization shall discuss and agree upon guidelines on how to perform the rating with the aim to improve the consistency in the rating between different viewpoints.

Figure 23: Impact rating scale

iii. Subjective value of improvement (SVI)

The subjective value of improvement is calculated per viewpoint taking the SW and IR into consideration using the following formula:

\[
SVI_{vw} = \sum (SW_i * IR_i)
\]

\(vw\): Sponsor / Coordinator / Implementer

\(i\): Metric 1, Metric 2, ……, Metric n
If there is more than one stakeholder present in each viewpoint, the stakeholders shall agree on the same values given for SW and IR. SPI-MEF proposes a Holistic View Calculation Table template (Table 12) for gathering the required values for the components of the holistic view and support the systematic calculation of SVI.

Table 12: Holistic View Calculation Table template

<table>
<thead>
<tr>
<th>Viewpoint</th>
<th>Metrics</th>
<th>Subjective Weight (SW)</th>
<th>Impact Rating (IR)</th>
<th>Subjective Value of Improvement (SVI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; List the stakeholders &gt;</td>
<td>&lt; Metric ID i &gt;</td>
<td>SW_i</td>
<td>IR_i</td>
<td>Σ (SW_i * IR_i )</td>
</tr>
<tr>
<td>(Coordinator)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; Metric ID n &gt;</td>
<td></td>
<td>SW_n</td>
<td>IR_n</td>
<td></td>
</tr>
<tr>
<td>(Implementer)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Sponsor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The calculation of SVI is questionable in software measurement theory because it involves mathematical operations which, in a strict sense, are not applicable on ordinal scales. On the other hand, it is pointed out by Stevens [107] that it can be practicable to treat an ordinal scale as interval scale. Therefore, the authors deem the application of statistical methods on the SVI, as the calculation of the mean, as permissible. A similar problem regarding the measurement of Function Points is discussed in [108]. Since the aim of the holistic view is to provide an overview of the improvement (the individual evaluations are a more appropriate data source for decision making), SVI has to be seen as an index or score. It gives an indication of the improvement, rather than a metric, which in its formal definition has to fulfill the representation condition of measurement [48].

5.3.3.2 Aggregation of multiple evaluation results

Driven by the Lag Factor (LF) and Degradation Factor (DF), SPI-MEF proposes a periodic improvement evaluation, that is, an evaluation indexed by time to see the trend of the improvement. However, periodic evaluation produces multiple evaluation results through time and an issue might arise when choosing which evaluation results to represent the actual improvement during the holistic view creation.

If multiple evaluations are conducted on the same target entities (e.g. a certain process or the organization as a whole) the latest evaluation results shall be taken as the representation of the actual improvement for the holistic view creation. However, if the target entities are different (e.g. different projects or products), an aggregation of the evaluation results needs to be performed.

By recapitulating from evaluation scheduling (Section 5.2.4), DF is the maximum time for which an evaluation result can be considered as valid. Based on this definition, DF can be used to solve the issue of choosing which evaluation results should be included in the aggregation.

The illustration in Figure 24 shows how this issue is resolved. If the holistic view is going to be created in October and based on a DF of four months, three evaluation results (E2, E3, and E4) should be considered and included in the aggregation.
After knowing which evaluation results to select, the next task is to aggregate the multiple evaluation results into one representative SVI. The simplest way to achieve this task is to take the arithmetic mean of the individual SVI’s. However, by following such an approach the properties from the entities (e.g. effort put into the process or project, cost and size of a product) from which the SVI’s are calculated are not taken into consideration. For example, the SVI from a small three month project would have the same weight as the one from a 24 month project. Therefore, the aggregated SVI should be weighted by an “investment unit” (IU). IU can be regarded as the resources which were spent in the implementation of processes, projects or products on which the individual evaluations are based [28].

Aggregated SVI for holistic view = \( \sum (SVI_i \times IU_i / \sum IU) \)

\( i: \) Evaluation result 1, Evaluation result 2, …., Evaluation result n

By referring to the illustration in Figure 24, the aggregated SVI for holistic view is computed as follow:

\[
\sum IU = IU_1 + IU_2 + IU_3 \\
= 5 + 3 + 2 \\
= 10
\]

Aggregated SVI for holistic view = \( \sum (SVI_i \times IU_i / \sum IU) \)

\[= [(1.4)(5/10) + (1.4)(3/10) + (1.8)(2/10)]\]

\[= 1.48\]

5.3.3.3 Presentation of the holistic view

In order to better understand the components of the holistic view and the calculation of SVI, an example for the Process level is shown in Table 13. The SW and IR values are rated by the corresponding stakeholders from their point of view for each metric.
Table 13: Example of holistic view calculation for Process

<table>
<thead>
<tr>
<th>Viewpoint</th>
<th>Metrics</th>
<th>Subjective Weight (SW)</th>
<th>Impact Rating (IR)</th>
<th>Subjective Value of Improvement (SVI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process manager</td>
<td>M01_PRC</td>
<td>0.40</td>
<td>+3.0</td>
<td>[(0.40)(+3.0) + (0.20)(+2.0) + (0.40)(-2.0)] = 0.00</td>
</tr>
<tr>
<td>(Coordinator)</td>
<td>M02_PRC</td>
<td>0.20</td>
<td>+2.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M03_PRC</td>
<td>0.40</td>
<td>-4.0</td>
<td></td>
</tr>
<tr>
<td>Development team</td>
<td>M01_PRC</td>
<td>0.60</td>
<td>+4.0</td>
<td>[(0.60)(+4.0) + (0.40)(+2.0)] = +3.20</td>
</tr>
<tr>
<td>(Implementer)</td>
<td>M02_PRC</td>
<td>0.40</td>
<td>+2.0</td>
<td></td>
</tr>
<tr>
<td>Head of R&amp;D</td>
<td>M01_PRC</td>
<td>1.00</td>
<td>-1.0</td>
<td>[(1.00)(-1.0)] = -1.00</td>
</tr>
<tr>
<td>(Sponsor)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

After all the SVI’s are computed for each viewpoint, the overall view of the evaluation results can be depicted. By consolidating the values of SVI from all viewpoints, SPI-MEF recommends the visualization of the holistic view in a radar (Kiviat) chart called the Holistic View Chart as shown in Figure 25.

![Holistic View Chart](image)

Figure 25: Example of the holistic view presentation for the Process level

The radar chart is constructed by considering the values of SVI from three different viewpoints. Figure 26 illustrates the interpretation of different holistic views presented in radar charts:

1. If the SVI values are computed as 0.00 in all three viewpoints, then the triangle in the radar chart is drawn exactly in the middle of the radar chart’s full range (-5.00 to +5.00).
2. The different SVI values denote the views of the improvement as perceived by different viewpoints. In this radar chart, all the three viewpoints have positive SVI values, but the Coordinator perceives the gain higher than the Sponsor than the Implementer.
3. If all the SVI values are negative, then the triangle in the radar chart is drawn in the inner range (-5.00 to 0.00). As observed in this radar chart, the area of the triangle drawn is significantly smaller than the previous ones (as shown in 1. and 2.).

4. In some situations, there can be mix of positive (gain) and negative (loss) values of SVI from different viewpoints. In this radar chart, the triangle is skewed towards the Implementer (high gain perceived by Implementer than Sponsor, while loss perceived by Coordinator).

![Radar charts](image)

Figure 26: Interpretation of different holistic views presented in radar charts

### 5.3.3.4 Combined holistic view

The combined holistic view represents the improvement at the various measurement levels. In this way, the impact of the improvement initiative can be made visible and put into relation at the levels where the change ought to have an effect. The combined holistic view can be constructed after at least two measurement levels have been assessed and can be augmented iteratively as soon as new evaluation results are available. A new component called the average value of improvement (AVI) is required for the combined holistic view creation. Assuming that each viewpoint has an equal contribution (or weight) to SVI, the value of AVI is calculated by taking the arithmetic mean of all the SVI's.
AVI_{ml} = \sum SVI_{vw} / n

\textit{ml}: \text{Process} / \text{Project} / \text{Product} / \text{Organization} / \text{External}

\textit{vw}: \text{Coordinator} / \text{Implementer} / \text{Sponsor}

\textit{n}: \text{Number of viewpoint involved (1 / 2 / 3)}

After all the AVI are computed for measurement levels that are involved in the evaluation, the overall view of the improvement evaluation can now be illustrated. Similarly to the individual holistic view, a radar (Kiviat) chart can be used to visualize the combined holistic view by consolidating the values of AVI from all measurement levels. An example is shown in Figure 27.

![Figure 27: Example of the combined holistic view](image-url)
Holistic view creation

Based on the pilot evaluation results obtained, a meeting with the key people representing each viewpoint were held in APOLLO. During the meeting, the SW and IR values were rated based on the evaluation results. The table below shows the computation of SVI for the Process level:

<table>
<thead>
<tr>
<th>Viewpoint</th>
<th>Metrics</th>
<th>Subjective Weight (SW)</th>
<th>Impact Rating (IR)</th>
<th>Subjective Value of Improvement (SVI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process team (Coordinator)</td>
<td>M01_PRC</td>
<td>0.30</td>
<td>+3.0</td>
<td>[(0.30)(+3.0) + (0.20)(+3.0) + (0.20)(+1.0) + (0.30)(0.0)] = +1.70</td>
</tr>
<tr>
<td></td>
<td>M02_PRC</td>
<td>0.20</td>
<td>+3.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M03_PRC</td>
<td>0.20</td>
<td>+1.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M04_PRC</td>
<td>0.30</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Development team (Implementer)</td>
<td>M01_PRC</td>
<td>0.50</td>
<td>+3.0</td>
<td>[(0.50)(+3.0) + (0.50)(+2.0)] = +2.50</td>
</tr>
<tr>
<td></td>
<td>M02_PRC</td>
<td>0.50</td>
<td>+2.0</td>
<td></td>
</tr>
<tr>
<td>Head of development (Sponsor)</td>
<td>M01_PRC</td>
<td>0.40</td>
<td>+2.0</td>
<td>[(0.40)(+2.0) + (0.60)(+3.0)] = +2.60</td>
</tr>
<tr>
<td></td>
<td>M02_PRC</td>
<td>0.60</td>
<td>+3.0</td>
<td></td>
</tr>
</tbody>
</table>

In order to have a “one-view” presentation of the overall improvement, the holistic view is created by consolidating all the SVI from different viewpoints into a radar chart:

Based on the created holistic view, the development team (Implementer) and the head of development (Sponsor) rated the improvement gained from code inspection relatively higher than the process team (Coordinator). This is largely due to that the development team and the head of development are interested to see the impact of code inspection only on the process effectiveness while the process team has to look into both the effectiveness and efficiency of the process.
5.4 Evolution stage

The evolution stage (Figure 28) is the last stage in SPI-MEF, and it is aimed to improve the evaluation process by analyzing feedback and lessons learnt gathered throughout the evaluation process. Since the evaluation process is iterative, the gathered feedback should be used to refine and improve and mature the evaluation process itself, and support the long-term measurement and evaluation strategy defined in the evaluation opportunity matrix (Section 5.1.2).

![Figure 28: SPI-MEF evolution stage](image)

5.4.1 Evaluation feedback analysis

After every complete cycle of the evaluation process, a post-mortem meeting should be held among the involved parties to gather feedback on every evaluation step. The agenda of the meeting shall cover at least the aspects found in the Evaluation Feedback Checklist (Table 14) provided by SPI-MEF.

The questions in the checklist are aimed to bring out potential areas where feedback can be gathered and discussed. The checklist is a growing list and it should be populated with more questions through several evaluation cycles and adapted by organizations according to their needs.

<table>
<thead>
<tr>
<th>Evaluation Feedback Checklist:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initialization stage</strong></td>
</tr>
<tr>
<td>i. Were there any changes to the SPI initiative implementation?</td>
</tr>
<tr>
<td>ii. Was the implementation of the SPI initiative aligned with the improvement goal?</td>
</tr>
<tr>
<td>iii. Was the planned evaluation opportunity path being followed in the actual evaluation?</td>
</tr>
<tr>
<td>iv. Has the actual cost for the evaluation remained within the range of the planned cost?</td>
</tr>
<tr>
<td>v. Does the evaluation opportunity matrix require revision? If yes, what are the change(s)?</td>
</tr>
<tr>
<td><strong>Design stage</strong></td>
</tr>
<tr>
<td>i. Were the identified stakeholders able to give input for the evaluation?</td>
</tr>
<tr>
<td>ii. Is there any stakeholder who should be included / excluded in the viewpoints?</td>
</tr>
<tr>
<td>iii. Were the goals and guidelines made clear during the derivation of measures?</td>
</tr>
<tr>
<td>iv. Were there any metrics found not appropriate for evaluating the SPI initiative?</td>
</tr>
<tr>
<td>v. Is there any additional metric that needs to be considered for evaluating the SPI initiative?</td>
</tr>
<tr>
<td>vi. Were the baselines accurately established?</td>
</tr>
<tr>
<td>vii. Were confounding factors being considered during the evaluation?</td>
</tr>
<tr>
<td>viii. Are there any revisions needed to the values of LF and DF?</td>
</tr>
<tr>
<td>ix. Was the periodic evaluation conducted according to the schedule?</td>
</tr>
</tbody>
</table>
### Implementation stage

i. Was there any problem in collecting the data for the elicited metrics?

ii. What were the biggest obstacles in meeting the scheduled periodic evaluations (if any)?

iii. Was there any resource problem throughout the evaluation process?

iv. Was there any communication problem between the parties involved in the evaluation process?

v. Was there any problem in conducting the evaluation based on the collected data?

vi. Were the impact ratings done consistently by the viewpoints across all measurement levels?

vii. Was there any difficulty in creating the holistic view?

### Refinements

i. Are there any refinements (addition, correction, deletion) needed to the support artifacts throughout the initialization, design and implementation stages?

ii. Are there any changes to the evaluation steps proposed by SPI-MEF that require further adaptation?

### Evaluation feedback analysis

A post-mortem meeting was held between the parties who involved in the pilot implementation of code inspection in APOLLO. Among the major issues and refinements that were discussed are:

i. The head of department felt that a new metric shall be added to evaluate the effectiveness of code inspection as to increase the confidence in the accuracy of the evaluation.

ii. The LF of four months was found too short to evaluate the impact of code inspection in the project and process level. A suggestion was to increase the LF to five months in the next evaluation cycle.

iii. The development team felt that there was a mix of opinions in the impact rating and subjective weight of the metrics. A suggestion was to limit the representatives from the development team to only senior and experienced staffs in order to reduce the variations in the ratings.
5.5 Summary

SPI-MEF is aimed to provide a systematic way for evaluating the outcome of SPI initiatives. The framework is implemented based on several key concepts for evaluating the outcome of SPI initiatives. The key concepts are incorporated within the steps of SPI-MEF. All of these steps belong to the stages that represent the logical flow of the evaluation process. In order to assist the evaluation process, SPI-MEF provides support artifacts. Figure 29 shows an overview of SPI-MEF. The mapping among the incorporated concepts, the stages, steps, and support artifacts are presented in the Table 15.

Table 15: Mapping between SPI-MEF’s stages, support artifacts, steps and concepts

<table>
<thead>
<tr>
<th>Stage</th>
<th>Support artifacts</th>
<th>Evaluation step</th>
<th>SPI-MEF concept</th>
</tr>
</thead>
</table>
| Initialization | • SPI Initiative Form  
• Evaluation Opportunity Matrix Template | Gather SPI initiative information              | --                                       |
|             |                                          | Gap analysis of evaluation quality             | Evaluation opportunity matrix            |
| Design      | • Evaluation Area Table  
• GQM Construction Table  
• Measurement Goal, Question(s), Metrics Template  
• Metric Form  
• List of Success Indicators and Metrics  
• Evaluation Scheduling Form | Evaluation scoping  
Determination of measures  
Selection of evaluation methods  
Evaluation scheduling | Evaluation area  
(Measurement levels and evaluation viewpoints)  
Primary and complementary measures  
Derivation of measures using GQM  
Cross-examination  
Evaluation method  
Confounding factors  
Time to evaluate |
| Implementation | • Evaluation Result Template  
• Holistic View Calculation Table  
• Holistic View Chart | Establishment of measurement program  
Conduct evaluation  
Holistic view creation | --  
--  
Holistic evaluation |
| Evolution   | • Evaluation Feedback Checklist          | Evaluation feedback analysis                  | --                                       |
Figure 29: Software Process Improvement Measurement and Evaluation Framework (SPI-MEF)
6 SCENARIO APPLYING SPI-MEF

6.1 Background

One Solution is a software company specialized in banking and financial solution software. The company has expertise in producing software for banks, insurances, mortgage and financial applications with clients dispersed all across Scandinavian countries. Software produced by One Solution includes Internet Banking System, Loan System, Deposit System, Mobile Banking System and front-end office for banking and also systems for a central mainframe platform.

One Solution has approximately 900 employees dispersed geographically in Sweden, Denmark, Norway and Finland. Product development at One Solution is divided into varied sizes of projects. A typical product consists of 1-3 projects with approximately 5-10 full-time staff working in one project a time. All projects follow the Waterfall model as software development life-cycle.

One Solution’s organizational structure consists of three layers. The top layer consists of corporate shareholders and top level managers (President, Vice President, etc.). Top level managers coordinate the middle organization layer which consists of head of departments. Departments within One Solution are divided into two functional units, the Research & Development (R&D) and the Support.

R&D is responsible for the development of One Solution products, whereas Support is not directly related with One Solution’s core business but supports and implements operational activities (e.g. Marketing Department, Human Resource Department, Finance Department, etc). R&D is led by one head of department which coordinates departments that are responsible for One Solution products (e.g. Department of Internet Banking, Department of Loan System, Department of Deposit System, etc.). R&D includes also a process group that is responsible for the implementation of software process improvements in product development.

Each of the development departments under R&D and the process group has low-level managers, which occupy the roles of project managers and product managers. These managers are responsible for the operation of project and product development. Within the process group are SPI coordinators who are responsible for the assimilation and adoption of new processes or technologies in all developments within the R&D departments.

One Solution’s highest priority is customer satisfaction. Their customers demand high quality software and fast delivery time to achieve their goals as the domain is quite competitive. In order to fulfill their customers’ needs, One Solution aims to increase the maturity of their processes by following the CMMI standard. Recently, One Solution has been assessed at CMMI Maturity Level 2 according to CMMI scale, and their goal is to increase their CMMI maturity level to Level 3 with the objective to deliver better products to their customers and increase their satisfaction.
A preliminary assessment for CMMI Level 3 showed that One Solution was still lacking in three key process areas (KPA). Those KPAs are listed below:

- **Product Integration (PI)**
  The purpose of Product Integration (PI) is to assemble the product from the product components, ensure that the product, as integrated, functions properly, and deliver the product.

- **Verification (VER)**
  The purpose of Verification (VER) is to ensure that selected work products meet their specified requirements.

- **Validation (VAL)**
  The purpose of Validation (VAL) is to demonstrate that a product or product component fulfills its intended use when placed in its intended environment.

To address these KPAs, One Solution has decided to adopt several improvement actions. Those actions are listed below:

i. **Peer Review Process**
   Peer review is an important activity in verification for effective defect detection. The type of peer review taken in this improvement is inspection. Inspection is a review of work products by trained individuals who want to find defects using a well-defined process. The inspection process requires training on the process change, methodologies and tools.

ii. **Unit Test Planning**
   Unit testing is found not to be properly conducted and thus many critical defects that should have been detected and resolved in the implementation phase escaped to later phases or were even shipped to the customer. Therefore, the process is expanded to include unit test plans so that the task can be performed more effectively. The unit test plan will contain a systematic approach on the creation of unit test cases during the design and coding phases and a specified test coverage goal must be achieved before going to the next development phase. Unit test planning also requires training on the process change, methodologies and tools.

iii. **Developmental Integration Testing (DIT)**
   The components of the various software products are developed by different development teams in different projects. An integration of these components requires developmental integration testing (DIT). DIT is intended to expose defects in the interfaces and the interaction among integrated components. DIT includes the planning and execution of tests in an environment when the various components are integrated. The developmental integration testing also requires training on the process change, methodologies and tools.

iv. **Support Tool for Software Development**
   The current bug tracking was done manually, which led to chaotic bug documentation. In order to solve this problem, a support tool for bug tracking needs to be implemented. The support tool will provide a systematic way of bug tracking and thus improve the software development process.

The initiatives will be rolled out in three phases according to the target entities where the actions are implemented (Figure 30). Later phases have broader scope, i.e. they address a
higher number of target entities. The reason for choosing this strategy is to gain better control on the improvement process. Starting with a smaller scope and then expand the initiatives to more target entities will help to manage the risk of failing initiatives or changes that do not perform as well as expected. Implementing the initiatives in three phases also allows applying refinements from one phase to another such that successive implementations succeed with a higher probability.

Figure 30: SPI-MEF stages applied in scenario phases

6.1.1 Initialization – Gather SPI initiative information

In the initialization stage, the information about the SPI initiative in One Solution is gathered and the SPI Initiative Form is filled in:

Table 16: SPI Initiative Form for One Solution

<table>
<thead>
<tr>
<th>SPI Initiative Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of SPI Initiative</td>
</tr>
<tr>
<td>SPI Frameworks</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>Capability Maturity Model Integrated (CMMI)</td>
</tr>
<tr>
<td>Current Maturity Level</td>
</tr>
<tr>
<td>Level 2</td>
</tr>
<tr>
<td>Description</td>
</tr>
<tr>
<td>Improvement Practices (see Section 6.1):</td>
</tr>
<tr>
<td>1. Peer Review Process</td>
</tr>
<tr>
<td>2. Unit Test Planning</td>
</tr>
<tr>
<td>3. Development Integration Testing</td>
</tr>
<tr>
<td>4. Support Tool for Software Development</td>
</tr>
<tr>
<td>Improvement Goals</td>
</tr>
<tr>
<td>The improvement goal is aimed to meet customer satisfaction in two aspects:</td>
</tr>
<tr>
<td>1. Improve product quality</td>
</tr>
<tr>
<td>2. Decrease time-to-delivery</td>
</tr>
<tr>
<td>Key Process Area (KPA)</td>
</tr>
<tr>
<td>The following KPAs are being addressed with the improvement practices (see Section 6.1):</td>
</tr>
<tr>
<td>1. Product Integration</td>
</tr>
<tr>
<td>2. Verification</td>
</tr>
<tr>
<td>3. Validation</td>
</tr>
<tr>
<td>Target Entities for</td>
</tr>
<tr>
<td>Implementation</td>
</tr>
<tr>
<td>The implementation of the process change will be carried out in three phases. Each phases are distinguished as follows:</td>
</tr>
<tr>
<td>Phase 1: Pilot projects in Internet Banking System development</td>
</tr>
<tr>
<td>Implementation Schedule</td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td>Phase 2: Expand to all projects in Internet Banking System development</td>
</tr>
<tr>
<td>Phase 3: Further expand to all projects in the following development:</td>
</tr>
<tr>
<td>✓ Loan System</td>
</tr>
<tr>
<td>✓ Deposit System</td>
</tr>
</tbody>
</table>

### 6.1.2 Initialization – Gap analysis of evaluation quality

In this step, the historical data gained in previous evaluations conducted by One Solution is analyzed. The assessment was conducted when the company decided to acquire the CMMI Level 2 certification. From the available historical data it is evident that One Solution only evaluated the performance of their processes at the Process and Project levels. Meetings with the employees in charge of the current measurement program within the SPI process group revealed that the collected measurements were defined ad-hoc, following best practices established in industry. In particular, following the definition of SPI-MEF, they only considered primary measurements, without considering complementary measurements and confounding factors. Knowing this fact, the current position of One Solution in the evaluation opportunity matrix is located in the area where accuracy is low and coverage is at an average level.

One Solution decides to achieve a holistic evaluation with high accuracy and wide coverage in order to benefit in the long-term from the improved evaluation process, which is, achieving confidence that the improvement initiatives provide the expected outcome and provide indeed value to their customers. A meeting with representatives of the stakeholders of the SPI initiatives, the SPI process group and the measurement team resulted in the definition of the path that One Solution is going to follow to evaluate the forthcoming SPI initiatives. The decision is based on the company's setting (primarily cost, time and available resources) and is also constrained by the phases that One Solution has already defined to roll-out the initiatives. Based on this information, the path that One Solution takes for evaluating the improvement is starting from achieving accuracy first and then addressing coverage. The path that One Solution is going to take towards the holistic view is shown in the Figure 31.

The evaluation of the first and second phases in the initiatives implementation is aimed to achieve high accuracy. Coverage is not aimed at this point since in the scope of those phases in terms of target entities is limited and it is assumed that the effect of the improvement will not yet be visible at the Product or Organizational Level. Besides, cost is also considered for this decision since the top-level management of One Solution is still reluctant to approve the resources for a broader evaluation before the accuracy of evaluations is improved. Therefore, starting with the evaluation in two measurement levels, Process and Project, in an accurate way is deemed as more reasonable than choosing to evaluate all measurement levels.
The evaluation of the third phase is anticipated to have a wider coverage as compared to the previous ones. The considered measurement levels include all measurement levels defined by SPI-MEF. The reason for having a wider coverage is that the effects of the improvement have propagated to the Product and Organizational level. Furthermore, the evaluation team assumes at that moment, that the top-level management is willing to provide more resources for the evaluation after seeing the results of the previous two evaluations.

6.2 Phase I

In this phase the initiatives are implemented in pilot projects. The objective of having one phase for piloting the initiatives in projects is to see whether the changed processes are applicable to the typical projects in One Solution and to discover any adaptability issues which can then be addressed in the following phases. Four pilot projects within Internet Banking Department are selected for this phase’s implementation.

6.2.1 Design – Evaluation scoping

In this step, the scoping is done by considering two measurement levels, Project and Process, and all stakeholders that are interested to see the evaluation result of those two levels. There are four stakeholders interested in this evaluation: SPI Coordinator, Head of Internet Banking Department, Development Teams and the Project Managers of the pilot projects.

In the Process level, the SPI Coordinator needs to know the compliance and adaptability of the process to define what refinements need to be implemented in order to allow a smooth introduction of the changes in the following phases. The SPI Coordinator needs to report to the Head of Internet Banking Department. The Head of Internet Banking Department needs to see the result of evaluation as he is the one who takes decisions for the next implementation phase in his department. The Development Teams who are implementing the

![Evaluation Opportunity Matrix for One Solution](image)

Figure 31: Evaluation Opportunity Matrix for One Solution
initiatives see the evaluation results as it motivates to adapt the changes and increases their commitment.

In the Project level, the Project Managers, as the ones who coordinate the entire processes within the pilot projects, need to see the evaluation results because they need to judge and gauge the change from the introduced processes as how they impact the projects in general. The Head of Department, as the superior of the Project Managers, is interested to see the result in this level because he decides which of the initiatives will be implemented in the following phase. The Development Teams of the pilot projects who are coordinated by the Project Managers need to know evaluation results to know the effect of the initiatives at the project scope so that they can be motivated for these improvement initiatives.

Table 17: Phase I: Evaluation Area Table

<table>
<thead>
<tr>
<th>Measurement Levels</th>
<th>Viewpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Implementer</td>
</tr>
<tr>
<td>Process</td>
<td>Development Team (Architect, Analyst, Programmer, Tester)</td>
</tr>
<tr>
<td>Project</td>
<td>Development Team (Architect, Analyst, Programmer, Tester)</td>
</tr>
</tbody>
</table>

6.2.2 Design – Determination of measures

From the SPI Initiative Form, it is known that there are two improvement goals for the introduced SPI initiatives. Since there are two separate improvement goals, the determination of measures will be performed separately for each improvement goal. For BG1: Improve product quality the GQM Construction Table (Table 18) is populated first and then the GQM approach is applied as shown in Table 19. The same approach is followed for BG2: Decrease time-to-deliver as shown in Table 20 and Table 21 respectively.

Table 18: Phase I: GQM Construction Table for BG1

<table>
<thead>
<tr>
<th>Purpose:</th>
<th>Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Goal (Improvement Goal):</td>
<td>BG1: Improve product quality</td>
</tr>
<tr>
<td>Context:</td>
<td>Pilot projects in Internet Banking System</td>
</tr>
<tr>
<td>Measurement Level</td>
<td>Point of View (Viewpoint)</td>
</tr>
<tr>
<td>Process</td>
<td>SPI Coordinator</td>
</tr>
</tbody>
</table>
(*) Process conformance is not a success indicator for the evaluation but in this phase where the SPI initiatives are piloted, it is an important Focus to assist the evaluation. This is done by looking at the degree of the process conformance and it gives the confidence on how much the impact can be attributed to the SPI initiatives.

Table 19: Phase I: Measurement Goal, Question(s) and Metrics for BG1

<table>
<thead>
<tr>
<th>Project</th>
<th>Complementary</th>
<th>Efficiency</th>
<th>MG02_PRC</th>
<th>Process Conformance*</th>
<th>MG03_PRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development Team</td>
<td>Primary</td>
<td>Effectiveness</td>
<td>MG04_PRC</td>
<td></td>
<td>MG05_PRC</td>
</tr>
<tr>
<td>Head of Department</td>
<td>Primary</td>
<td>Effectiveness</td>
<td>MG01_PRC</td>
<td></td>
<td>MG02_PRJ</td>
</tr>
<tr>
<td>Project Manager</td>
<td>Primary</td>
<td>Defects</td>
<td>MG03_PRJ</td>
<td></td>
<td>MG04_PRJ</td>
</tr>
<tr>
<td>Development Team</td>
<td>Primary</td>
<td>Defects</td>
<td>MG03_PRJ</td>
<td></td>
<td>MG04_PRJ</td>
</tr>
<tr>
<td>Head of Department</td>
<td>Primary</td>
<td>Defects</td>
<td>MG03_PRJ</td>
<td></td>
<td>MG04_PRJ</td>
</tr>
</tbody>
</table>

MG01_PRC: Effectiveness
Evaluate the SPI initiatives in terms of their effectiveness in defect detection on pilot projects from the SPI Coordinator point of view.

**Question 1:** What is the number of defects found by the peer review process in terms of its effectiveness?

**Metrics:**

- **M01_PRC:**
  Phase Containment Effectiveness for Phase \(i\) (PCE \(i\))
  \[
  \text{PCE}_i = \frac{F}{F + E}
  \]
  \(F\) = Number of defects found in phase \(i\)
  \(E\) = Number of defects escaped from phase \(i\)
  \(i\) = Requirement, Design, Coding, Unit Test

  This metric shows the phase containment effectiveness for a certain phase during the development. Conducting peer reviews is expected to retain more defects from escaping to the next development phases.

**Question 2:** What is, in overall, the number of defects found by the introduced SPI initiatives (peer review process, unit test planning and development integration testing) in terms of their effectiveness?

**Metrics:**

- **M02_PRC:**
  Defect detection effectiveness before system test
  \[
  \text{Defect detection effectiveness before system test} = \frac{A}{B}
  \]
  \(A\) = Total number of actual high severity defects found before system test
  \(B\) = Total number of actual high severity defects found before delivery to customer
  High severity defect = Defect that causes crash or misbehaviour of the intended function.

  This metric shows the effectiveness of the defect detection processes (peer review process, unit test planning and development integration testing) during the development before
entering the system test phase.

Notes: Since the initiatives are introduced simultaneously it is impossible to reliably assess the effect of a single initiative. Therefore, in Question 2, the initiatives are addressed as a whole. The development life-cycle of One Solution allows an exception: since the Waterfall model is applied, the development phases are conducted in sequence; therefore it is assumed to be possible to assess the effectiveness of the peer reviews (Question 1) since no other initiative will interfere with it and influence the outcome. The same principle applies to the next measurement goal, process efficiency.

### MG02_PRC: Efficiency
Evaluate the SPI initiatives in terms of its efficiency in defect detection on pilot projects from the SPI Coordinator point of view.

**Question 1:** What is the number of defects found by the peer review process in terms of its efficiency?

**Metrics:**

**M03_PRC:**
Defect removal efficiency, DRE before Unit Test  
DRE before Unit Test = \( \frac{E}{E+D} \)  
\( E \) = Total number of defects found before Unit Test  
\( D \) = Total number of defects found after Unit Test

This metric shows the defect removal efficiency with the help of the peer review process before entering the unit test phase.

**Question 2:** What is the number of defects found by the introduced SPI initiatives (peer review process, unit test planning and development integration testing) in terms of their efficiency?

**Metrics:**

**M04_PRC:**
Defect removal efficiency, DRE  
DRE = \( \frac{E}{E+D} \)  
\( E \) = Total number of defects found before delivery to customer  
\( D \) = Total number of defects found after delivery to customer

This metric shows the defect removal efficiency of the SPI initiatives (peer review process, unit test planning and development integration testing) cumulatively. The initiatives are expected to increase the number of defects found before delivering the product to customers.

### MG03_PRC: Process Conformance
Evaluate the SPI initiatives with respect to its conformance to the specified standard process in the pilot projects from the SPI Coordinator point of view.

**Question 1:** What is the degree of process compliance of the SPI initiatives when they are implemented?

**Metrics:**

**M05_PRC:**
Checklist-based rating for each process  
Each SPI initiative will have a pre-defined checklist containing a list of standards or requirements that need to be followed. The rating will not be used for evaluating the success of the SPI implementation but it is intended to be used along-side with the evaluation results to judge how much the improvement can be attributed to the implemented process change.
### MG01_PRJ: Defects
Evaluate the SPI initiatives in terms of their impact on the number of encountered defects in the pilot projects from the Project Manager point of view.

**Question 1:** What is the number of defects in the work products in the pilot projects?

**Metrics:**

<table>
<thead>
<tr>
<th>M01_PRJ:</th>
<th>Defect density, DD for Phase $i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD for Phase $i = N / S$</td>
<td></td>
</tr>
<tr>
<td>$N =$ Number of defects in the work products of the respective development phases</td>
<td></td>
</tr>
<tr>
<td>$S =$ Project size calculated in lines-of-codes (LOC)</td>
<td></td>
</tr>
<tr>
<td>$i =$ Requirement, Design, Coding, Unit Test, Integration</td>
<td></td>
</tr>
</tbody>
</table>

This metric shows the number of defects in each phase normalized by the project size. With the introduction of the new SPI initiatives, the defect density of later development phases (e.g. Coding and Unit Test) is expected to be lowered as compared to previous projects because the peer review activities are expected to prevent defects from escaping to the next phases.

### MG02_PRJ: Cost/Effort
Evaluate the SPI initiatives in terms of their impact on the resources needed in the pilot projects from the Project Manager point of view.

**Question 1:** What is the actual effort that is required to develop the pilot projects?

**Metrics:**

<table>
<thead>
<tr>
<th>M02_PRJ:</th>
<th>Total project effort normalized by project size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total project effort normalized by project size $= \frac{\text{Actual project effort}}{\text{Actual project size}}$</td>
<td></td>
</tr>
</tbody>
</table>

This metric shows the total effort in the development of the pilot projects. This metric will indicate if there is any significant effort that is required to incorporate the introduction of the new SPI initiatives.

### MG03_PRJ: Defects
Reuse of the same Questions and Metrics from MG01_PRJ, point of view is the Development Team.

### MG04_PRJ: Defects
Reuse of the same Questions and Metrics from MG01_PRJ, point of view is the Head of Department.
Table 20: Phase I: *GQM Construction Table* for BG2

| Purpose: | Evaluate |
| Business Goal (Improvement Goal): | BG2: Decrease time-to-deliver |
| Context: | Pilot projects in Internet Banking System |

<table>
<thead>
<tr>
<th>Measurement Level</th>
<th>Point of View (Viewpoint)</th>
<th>Category</th>
<th>Focus (Success Indicator)</th>
<th>Measurement Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>SPI Coordinator</td>
<td>Primary</td>
<td>Efficiency</td>
<td>MG06_PRC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complementary</td>
<td>Effectiveness</td>
<td>MG07_PRC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Process Conformance*</td>
<td>MG08_PRC</td>
</tr>
<tr>
<td></td>
<td>Development Team</td>
<td>Primary</td>
<td>Efficiency</td>
<td>MG09_PRC</td>
</tr>
<tr>
<td></td>
<td>Head of Department</td>
<td>Primary</td>
<td>Efficiency</td>
<td>MG10_PRC</td>
</tr>
<tr>
<td>Project</td>
<td>Project Manager</td>
<td>Primary</td>
<td>Schedule</td>
<td>MG05_PRJ</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complementary</td>
<td>Defects</td>
<td>MG06_PRJ</td>
</tr>
<tr>
<td></td>
<td>Development Team</td>
<td>Primary</td>
<td>Schedule</td>
<td>MG07_PRJ</td>
</tr>
<tr>
<td></td>
<td>Head of Department</td>
<td>Primary</td>
<td>Schedule</td>
<td>MG08_PRJ</td>
</tr>
</tbody>
</table>

(*) Process conformance is not a success indicator for the evaluation but in this phase where the SPI initiatives are piloted, it is an important Focus to assist the evaluation. This is done by looking at the degree of the process conformance and it gives the confidence on how much the impact can be attributed to the SPI initiatives.

Table 21: Phase I: *Measurement Goal, Question(s) and Metrics* for BG2

**MG06_PRC: Efficiency**
Reuse of the same Questions and Metrics from MG02_PRC, point of view is the SPI Coordinator.

**MG07_PRC: Effectiveness**
Reuse of the same Questions and Metrics from MG01_PRC, point of view is the SPI Coordinator.

**MG08_PRC: Process Conformance**
Reuse of the same Questions and Metrics from MG03_PRC, point of view is the SPI Coordinator.

**MG09_PRC: Efficiency**
Reuse of the same Questions and Metrics from MG02_PRC, point of view is the Development Team.
MG10_PRC: Efficiency
Reuse of the same Questions and Metrics from MG02_PRC, point of view is the Head of Department.

MG05_PRJ: Schedule
Evaluate the SPI initiatives in terms of their impact on project schedule of the pilot projects from the Project Manager point of view.

**Question 1:** What is the actual project schedule for the pilot projects?

**Metrics:**

**M03_PRJ:**
Project cycle-time normalized by project size

\[
\text{Project cycle-time normalized by project size} = \frac{\text{Actual project cycle-time}}{\text{Actual project size}}
\]

This metric gives the total project cycle-time in the development of the pilot projects, normalized by the project size. This metric will indicate if there is any significant delay in the usual project cycle-time after incorporating the new SPI initiatives.

MG06_PRJ: Defects
Reuse of the same Questions and Metrics from MG01_PRJ, point of view is the Project Manager.

MG07_PRJ: Schedule
Reuse of the same Questions and Metrics from MG05_PRJ, point of view is the Development Team.

MG08_PRJ: Schedule
Reuse of the same Questions and Metrics from MG05_PRJ, point of view is the Head of Department.

Figure 32 illustrates the GQM Tree for Phase I which enhances the traceability of the elicited metrics to the respective measurement and business goal.

**GQM Tree for Phase I**

![GQM Tree for Phase I](image)

Figure 32: GQM tree for Phase I
6.2.3 Design – Selection of evaluation methods

In this step, the evaluation methods for each metric are defined. The metrics alongside with their properties and applied evaluation methods is documented in the Metric Form. Table 22, Table 23 and Table 24 below show examples of the Metric Form for metrics M01_PRC, M04_PRC and M03_PRJ defined in the previous step.

Table 22: Phase I: Metric Form for M01_PRC

<table>
<thead>
<tr>
<th>Metric ID</th>
<th>M01_PRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric Name</td>
<td>Phase Containment Effectiveness for Phase i (PCE i)</td>
</tr>
<tr>
<td>Success Indicator</td>
<td>Effectiveness</td>
</tr>
</tbody>
</table>
| Measurement Goal (Viewpoint) | i. MG01_PRC (SPI Coordinator)  
ii. MG04_PRC (Development Team)  
iii. MG05_PRC (Head of Department)  
v. MG07_PRC (SPI Coordinator) |
| Measurement Level | Process |
| Algorithm / Description | PCE i = F / (F + E)  
F = Number of defects found in phase i  
E = Number of defects escaped from phase i  
i = Requirement, Design, Coding, Unit Test |

This metric shows the phase containment effectiveness for certain phase during the development. Conducting peer reviews is expected to retain more defects from escaping to the next development phases.

Example: The number of defects found in Coding phase is 30 defects and the number of defects that managed to escape from Coding phase is 15. Therefore, the PCE Coding = 30 / (30 + 15) = 66%

| Baseline Value | The baselines for this metric can be computed based on the data collected from previous projects or calculated from current projects (without the SPI initiative implemented) that are running. Based on the historical data in the organization, the baseline values for PCE are:  
PCE Requirement = 60%  
PCE Design = 50%  
PCE Coding = 55%  
PCE Unit Test = 40% |
| Deviation Range | The deviation range for the respective phases is ±10%. |
| Evaluation Method | Pre-post comparison  
Objective: The evaluation is aimed to find out if the SPI initiatives are able to increase the value of PCE in the development phases of the pilot projects. |
| Confounding Factor(s) | During pre-post comparison, the following factors should be taken into consideration in the evaluation:  
1. Project size and complexity (LOC, No. of features) |
2. Technology factors (Language, tool, environment, platform)
3. Employee factors (Level of experience, turnover)
4. Process Compliance (Degree of process conformance implemented and followed)

Table 23: Phase I: Metric Form for M04_PRC

<table>
<thead>
<tr>
<th>Metric Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric ID</td>
</tr>
<tr>
<td>Metric Name</td>
</tr>
<tr>
<td>Success Indicator</td>
</tr>
</tbody>
</table>
| Measurement Goal (Viewpoint) | i. MG02_PRC (SPI Coordinator)
   ii. MG06_PRC (SPI Coordinator)
   iii. MG09_PRC (Development Team)
   iv. MG10_PRC (Head of Department) |
| Measurement Level | Process |
| Algorithm / Description | DRE = E / (E+D)
E = Total number of defects found before delivery to customer
D = Total number of defects found after delivery to customer
This metric shows the defect removal efficiency of the SPI initiatives (peer review process, unit test planning and development integration testing) cumulatively. The initiatives are expected to increase the number of defects found before delivering the product to customers.
Example: The number of defects found before delivery is 120 defects and the number of defects that are reported by the customer is 45. Therefore, the DRE = 120 / (120 + 45) = 73% |
| Baseline Value | The baselines for this metric can be computed based on the data collected from previous projects or calculated from current projects (without the SPI initiative implemented) that are running. Based on the historical data in the organization, the baseline value for DRE is 75%. |
| Deviation Range | The deviation range is set to ±10%. |
| Evaluation Method | Pre-post comparison
Objective: The evaluation is aimed to find out if the SPI initiatives are able to increase the value of DRE in the pilot projects. |
| Confounding Factor(s) | During pre-post comparison, the following factors should be taken into consideration in the evaluation:
1. Project size and complexity (LOC, No. of features)
2. Technology factors (Language, tool, environment, platform)
3. Employee factors (Level of experience, turnover)
4. Process Compliance (Degree of process conformance implemented and followed) |
Table 24: Phase I: *Metric Form* for M03_PRJ

<table>
<thead>
<tr>
<th>Metric Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric ID   : M03_PRJ</td>
</tr>
<tr>
<td>Metric Name : Project cycle-time normalized by project size</td>
</tr>
<tr>
<td>Success Indicator : Schedule</td>
</tr>
<tr>
<td>Measurement Goal (Viewpoint) :</td>
</tr>
<tr>
<td>i. MG05_PRC (Project Manager)</td>
</tr>
<tr>
<td>ii. MG07_PRC (Development Team)</td>
</tr>
<tr>
<td>iii. MG08_PRC (Head of Department)</td>
</tr>
<tr>
<td>Measurement Level : Project</td>
</tr>
<tr>
<td>Algorithm / Description : Project cycle-time normalized by project size = Actual project cycle-time / Actual project size</td>
</tr>
<tr>
<td>This metric gives the total project cycle-time in the development of the pilot projects, normalized by the project size. This metric will indicate if there is any significant delay in the usual project cycle-time after incorporating the new SPI initiatives.</td>
</tr>
<tr>
<td>Example: The project size is 10 KLOC and the actual project cycle-time is 3.5 months. Therefore, the project cycle-time normalized by project size = 3.5 / 10 = 0.35 months per KLOC.</td>
</tr>
<tr>
<td>Baseline Value : The baselines for this metric can be computed based on the data collected from previous projects or calculated from current projects (without the SPI initiative implemented) that are running. Based on the historical data in the organization, the baseline value for project cycle-time normalized by project size is 0.20 months per KLOC.</td>
</tr>
<tr>
<td>Deviation Range : The deviation range is set to ±20%.</td>
</tr>
<tr>
<td>Evaluation Method : Pre-post comparison</td>
</tr>
<tr>
<td>Objective: The evaluation is aimed to find out if the SPI initiatives affects the project cycle-time in the pilot projects.</td>
</tr>
<tr>
<td>Confounding Factor(s) : During pre-post comparison, the following factors should be taken into consideration in the evaluation:</td>
</tr>
<tr>
<td>1. Project size and complexity (LOC, No. of features)</td>
</tr>
<tr>
<td>2. Technology factors (Language, tool, environment, platform)</td>
</tr>
<tr>
<td>3. Employee factors (Level of experience, turnover)</td>
</tr>
<tr>
<td>4. Process Compliance (Degree of process conformance implemented and followed)</td>
</tr>
</tbody>
</table>
6.2.4 Design – Evaluation scheduling

After the required metrics have been defined, the evaluation on the pilot projects can now be scheduled as shown in Table 25.

Table 25: Phase I: Evaluation Scheduling Form

<table>
<thead>
<tr>
<th>Duration:</th>
<th>Starting from June 2006 until December 2006 (6 months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context:</td>
<td>There are 4 pilot projects in the Internet Banking System Department that have been selected for the evaluation. The 4 projects have the following schedule:</td>
</tr>
<tr>
<td></td>
<td>i) Project I: 30th June 2006 – 25th August 2006</td>
</tr>
<tr>
<td></td>
<td>iii) Project III: 5th September 2006 – 5th November 2006</td>
</tr>
<tr>
<td></td>
<td>iv) Project IV: 9th November 2006 – 28th December 2006</td>
</tr>
</tbody>
</table>

Note: In order to effectively pilot out the SPI initiatives in the projects within the Internet Banking System Department, the above projects are selected based on the same staffs who are working on these projects, i.e. staffs working in Project I and III, and II and IV are the same. Proper rolling out of the SPI initiatives is performed that includes training sessions and written guidelines for staffs who are working on these projects.

<table>
<thead>
<tr>
<th>Time to evaluate / Periodic evaluation</th>
<th>Measurement Levels</th>
<th>Lag Factor (LF)</th>
<th>Degradation Factor (DF)</th>
<th>Time to Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project</td>
<td>4 months</td>
<td>The usual project cycle-time for Internet Banking development is around three months and one month in the beginning is allocated for training and learning. Thus, LF is set to four months.</td>
<td>The purpose of the evaluation at this pilot phase is to assess if there is any significant impact of the SPI initiatives to the Process and Project level. Therefore, DF which is needed to assist the determination of periodic evaluation is not required at this phase.</td>
<td>After the first four months from the SPI initiatives implementation.</td>
</tr>
<tr>
<td>Process</td>
<td>4 months</td>
<td>Due to the fact that the software process is tightly coupled with the projects in the development, the same LF is taken from Project and being applied here.</td>
<td></td>
<td>After the first four months from the SPI initiatives implementation.</td>
</tr>
</tbody>
</table>

Note: Considering the LF, only Project III and Project IV will be evaluated.

Figure 33: Evaluation scheduling for Phase I
As illustrated in Figure 33, Project I and Project II are not considered for evaluation because they are within the time bound of the LF. Thus, only Project III and Project IV will be evaluated at the end of December 2006.

6.2.5 Implementation – Establishment of measurement program

The metrics from the derivation of metrics step are investigated to find out which ones are reusable and which ones are new and have not been implemented or used within One Solution (Table 26). The metrics are then established using previous knowledge and guidelines for measurement program establishment that One Solution already has.

Table 26: Phase I: New and reusable metrics

<table>
<thead>
<tr>
<th>Measurement levels</th>
<th>Reusable/existing metrics</th>
<th>New metrics to be implemented</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Process</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. M01_PRC: PCE for Phase i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. M03_PRC: Defect removal efficiency before Unit Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. M04_PRC: Defect removal efficiency</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Project</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. M01_PRJ: Defect density for Phase i</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. M02_PRJ: Total project effort normalized by project size</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. M02_PRC: Defect detection effectiveness before system test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. M05_PRC: Checklist-based rating for each process</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6.2.6 Implementation – Conduct evaluation

After the metrics that are required for the purpose of evaluations have been incorporated in the existing measurement program of the Internet Banking Department, the evaluations are conducted according to the plan and schedule. One evaluation is conducted both at the Process and Project level at the end of December 2006. The results of the evaluation are grouped according to the measurement level and they are depicted in the illustrations below:
Figure 34: Process level evaluation results for Phase I
Figure 34 shows the summary of the evaluation results in the Process level. Each of the used metrics in the evaluation is shown with its properties, success indicator, actual evaluation result, baseline value (if any) and the stakeholders (viewpoints) who are interested to see the result. Each metric is displayed in a dashboard diagram with three segments. The midpoint of the yellow segment is the baseline value against which the actual value is compared. The red and green segments represent decline and improvement respectively and are defined by the Deviation Range defined in the Metric Form.

A total of 10 metrics are evaluated in the Process level; three of them (M05_PRC) are not meant for success indication but are used to assess the process conformance. From the obtained evaluation results, the effectiveness and efficiency show an indication that the SPI initiatives are giving a positive impact to the process. The stakeholders (SPI Coordinator, Development Team and Head of Department) are all pleased with the results. They feel confident that the SPI initiatives will be able to improve their current processes when they are going to be implemented in Phase II to all projects in the Internet Banking System development.

Figure 35: Project level evaluation results for Phase I
Figure 35 above shows the summary of the evaluation results in the Project level. A total of seven metrics are evaluated in the Project level and five of them (M01_PRJ) belong to the same category of metrics (one for each development phase). The defect density metrics show a positive indication because the number of defects that are found during the each phase decreased. On the other hand, the effort and schedule metrics show a slight increase and this is probably due to the overhead of the introduction of the SPI initiatives. However, the extra overhead was found to be not significant with respect to the magnitude of the process change that was implemented. In general, the evaluation results show a positive impact as a result of the SPI initiatives in the Project level.

6.2.7 Implementation – Holistic view creation

In this phase’s evaluation, no holistic view is created because the purpose of this phase is to pilot the SPI initiatives in the Internet Banking Department. The piloting is intended to assess the adaptability of the SPI initiatives in the projects of One Solution. Furthermore, the improvement from the SPI initiatives seen in this phase’s evaluation is not significant to be presented to upper management for cost-benefit analysis or any kind of decision making.

6.2.8 Evolution – Evaluation feedback analysis

In this stage, the entire evaluation process for the first phase is evaluated. The post-mortem analysis is held in a meeting with the evaluation team who conducted the evaluation and with the representatives of the stakeholders (viewpoints). The meeting is aimed to detect problems in the evaluation process and to elaborate solutions for those issues in order to refine the evaluation process for the next phase.

During the meeting, issues about the sufficiency of the metrics used are discussed. The Project Managers feel that there is a need to identify more metrics for Phase II implementation in order to increase the confidence level on the gained improvements. The Head of Department also agrees with the Project Managers and suggests including productivity metrics to assess the impact of the SPI initiatives on the staff’s performance. The SPI Coordinator, on the other hand, suggests to include statistical analysis such as control charts to evaluate the trend of the defect metrics over the time.

Aside from the above discussions, the stakeholders also discuss problems in the data collection mechanisms on the target entities. The evaluation team who was supposed to gather the data relies too much on the people from the projects which are supposed to report the data. This has led to a delay in the evaluation process because people from the projects tend to forget and delaying the task. In the next phase, the evaluation team is advised to act more actively by collecting data in the target entity, paying attention not to interrupt the ongoing development process while doing it.

6.3 Phase II

In this phase, the SPI initiatives are expanded to be implemented in all projects within Internet Banking System development. The aim of the implementation is aligned with the improvement goals of the SPI initiatives, that is, to improve the product quality and achieve a shorter time-to-deliver. In order to shorten the scenario and limit redundancy, only the Product level is discussed in detail for each step since Process and Project level were already shown in Phase I.
6.3.1 Design – Evaluation scoping

In this step, scoping for Process, Project and Product levels is conducted. Stakeholders that are interested to see the results in these levels are the Development Team of Internet Banking projects, Project Managers of Internet Banking projects, the SPI Coordinator, Internet Banking Product Managers, the Head of the Process Group, the Head of the Internet Banking Department, and the Head of R&D. Table 27 shows how these viewpoints are mapped to the measurement levels and the following paragraphs describe the rationale for this mapping.

In Process level, the SPI Coordinator coordinates the implementation of initiatives for all projects within the Internet Banking Department. Therefore, he needs to see the evaluation results in order to judge the impact in this larger scale. In this phase, the Head of Internet Banking Department is convinced about the improvement and he passes the sponsorship to the Head of the Process Group. He, as the superior of the SPI Coordinator and the person who decided which initiatives to implement, needs to know how they perform in this larger scope, that is, in the whole Internet Baking Department. The Development Teams of Internet Banking projects are also interested to see the evaluation results at this level so that they can see the effect of their actions and the results can motivate them to work along with the improvement initiatives.

In the Project level, Project Managers of Internet Banking projects, as the ones who coordinate the entire processes in the Internet Banking projects, need to see the evaluation results such that they can judge the improvement and impact on the projects in general. The Head of Internet Banking Department as the superior of Project Managers needs to see the results of the evaluation since he wants to know whether the initiatives bring benefits in his department or not. The Development Teams of Internet Banking need to see the results in this level to get motivated for and stay committed with the improvement initiatives.

In the Product level, Product Managers in Internet Banking Department need to see the evaluation results so they can gauge the impact of the initiatives in their products. The Head of R&D who supervises all products of One Solution also needs to know the improvement in Internet Banking products because Internet Banking is one of the companies’ major products and has to be monitored closely. Another reason is that the Head of R&D has to take the decision whether the improvement initiatives are worth to be implemented in Phase III, in which the initiatives will be implemented in the remaining departments. The Project Team of Internet Banking needs to see the evaluation results because they participate in the projects that build up the products. Therefore, they must be aware of the effects of the initiatives in the products such that they stay committed to the initiatives in their daily work.

<table>
<thead>
<tr>
<th>Measurement Levels</th>
<th>Viewpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project</td>
<td>Implementer: Development Team (Architect, Analyst, Programmer, Tester)</td>
</tr>
</tbody>
</table>

Table 27: Phase II: Evaluation Area Table
6.3.2 Design – Determination of measures

Similar to Phase I, the determination of measures is separated according to the two improvement goals. Since the determinations of metrics for the Process and Project measurement levels has been shown in Phase I, only the necessary steps for the Product level are shown. For BG1: Improve product quality the GQM Construction Table (Table 28) is created and the GQM approach applied as shown in Table 29. The same approach is followed for BG2: Decrease time-to-deliver as shown in Table 30 and Table 31 respectively.

Table 28: Phase II: GQM Construction Table for BG1

<table>
<thead>
<tr>
<th>Purpose:</th>
<th>Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Goal (Improvement Goal):</td>
<td>BG1: Improve product quality</td>
</tr>
<tr>
<td>Context:</td>
<td>Projects in Internet Banking System development</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Level</th>
<th>Point of View (Viewpoint)</th>
<th>Category</th>
<th>Focus (Success Indicator)</th>
<th>Measurement Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product</td>
<td>Product Manager</td>
<td>Primary</td>
<td>Quality</td>
<td>MG01_PRD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complementary</td>
<td>Time-to-market</td>
<td>MG02_PRD</td>
</tr>
<tr>
<td></td>
<td>Project Team</td>
<td>Primary</td>
<td>Quality</td>
<td>MG04_PRD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complementary</td>
<td>Time-to-market</td>
<td>MG06_PRD</td>
</tr>
<tr>
<td></td>
<td>Head of R&amp;D</td>
<td>Primary</td>
<td>Quality</td>
<td>MG05_PRD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complementary</td>
<td>Time-to-market</td>
<td>MG07_PRD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cost</td>
<td>MG03_PRD</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Cost</td>
<td>MG03_PRD</td>
</tr>
</tbody>
</table>

Table 29: Phase II: Measurement Goal, Question(s) and Metrics for BG1

**MG01_PRD: Quality**
Evaluate the SPI initiatives in terms of their impact on the quality on the products developed within the Internet Banking from the Product Manager point of view.

**Question 1:** What is the product quality of the final software delivered to customers?

**Metrics:**
<table>
<thead>
<tr>
<th>M01_PRD:</th>
<th>Post-release fault density</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post-release fault density = $F / S$</td>
<td></td>
</tr>
<tr>
<td>$F$ = Number of failures encountered by customer during operation</td>
<td></td>
</tr>
<tr>
<td>$S$ = Size of the delivered product in KLOC</td>
<td></td>
</tr>
</tbody>
</table>

This metric shows the product quality in terms of the failures during operation. The introduced SPI initiatives are expected to help in capturing more defects in the products before releasing them to customers and therefore reduced the customer reported failures.

**Question 2:** How reliable is the final delivered product during operation?

**Metrics:**

<table>
<thead>
<tr>
<th>M02_PRD:</th>
<th>Mean Time Between Failures (MTBF)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTBF = $\frac{\text{Sum of all differences between uptime and downtime}}{\text{Number of failures}}$</td>
<td></td>
</tr>
</tbody>
</table>

This metric shows the reliability of the product during operation. The longer the time period between failures, the higher is the reliability of the system. The metric is using the total uptime of the system as the numerator and divides it by the number of failures to get an average value.

**Question 3:** What is the customer satisfaction on the features delivered in the final product?

**Metrics:**

<table>
<thead>
<tr>
<th>M03_PRD:</th>
<th>Percentage of delivered features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage of delivered features = $\frac{I}{R}$</td>
<td></td>
</tr>
<tr>
<td>$I$ = Number of features implemented</td>
<td></td>
</tr>
<tr>
<td>$R$ = Total number of features in the requirements</td>
<td></td>
</tr>
</tbody>
</table>

This metric can be used to indicate the customer satisfaction in terms of the fulfilment of the requirements that are agreed in the contract.

<table>
<thead>
<tr>
<th>M04_PRD:</th>
<th>Customer satisfaction survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>In order to effectively evaluate the customer satisfaction, one of the best ways is to hear it from the voice of the customers themselves. This can be done by preparing a customer satisfaction survey addressing the achieved product quality.</td>
<td></td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>MG02_PRD:</th>
<th>Time-to-market</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluate the SPI initiatives in terms of their impact on the time-to-delivery of the products within the Internet Banking development from the Product Manager point of view.</td>
<td></td>
</tr>
</tbody>
</table>

**Question 1:** What is the time that is needed to deliver a product to customer?

**Metrics:**

<table>
<thead>
<tr>
<th>M05_PRD:</th>
<th>Time-to-deliver</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-to-deliver = Time elapsed between the system requirements conception and deployment.</td>
<td></td>
</tr>
</tbody>
</table>

The typical product development in One Solution consists of 1-3 projects. Therefore, the product development time is highly dependent on the schedule of the projects involved in the product development.
Question 1: What is the cost involved of developing the product?

Metrics:

**M06_PRD:**
Relative cost of System Test
Relative cost of System Test = \( \frac{T}{P} \times 100\% \)
- \( T \) = Product testing effort involved, that refers specifically to System Testing after the features from all the projects are integrated.
- \( P \) = Overall product development effort
Effort is calculated in man-hours

The SPI initiatives are aimed to contain more defects in earlier phases of the development and therefore it is expected that System Testing requires less effort and resources.

**M07_PRD:**
Cost of rework
Cost of rework normalized by size = \( \frac{E}{S} \)
- \( E \) = Total effort of fixing the defects; in man-hours
- \( S \) = Product size in KLOC

The SPI initiatives are expected to reduce the numbers of defects and thus the cost of rework, i.e. the cost of fixing defects, is expected to be reduced too.

**MG04_PRD: Quality**
Reuse of Question 1 and its Metric from MG01_PRD, point of view is the Project Team.

**MG05_PRD: Quality**
Reuse of Question 3 and its Metric from MG01_PRD, point of view is the Head of R&D.

**MG06_PRD: Time-to-market**
Reuse of the same Question and Metric from MG02_PRD, point of view is the Head of R&D.

**MG07_PRD: Cost**
Evaluate the SPI initiatives in terms of their impact on the cost of the products within the Internet Banking development from the Head of R&D point of view.

Question 1: What is the total cost of the product?

Metrics:

**M08_PRD:**
Product cost normalized by size
Product cost normalized by size = \( \frac{(E + M)}{S} \)
- \( E \) = Total effort for the whole product life cycle in man-month
- \( M \) = Maintenance cost that involves the cost of fixing post-release defects and all the necessary support cost calculated in man-month
- \( S \) = Product size in KLOC

Although cost is not the improvement goal of the SPI initiatives, it is still a major concern to the Head of R&D. He would be interested to see if the SPI initiatives would give any positive impact to the cost of the product in the long-run.
Table 30: Phase II: *GQM Construction Table* for BG2

<table>
<thead>
<tr>
<th>Purpose:</th>
<th>Evaluate</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Business Goal (Improvement Goal):</strong></td>
<td>BG2: Decrease time-to-deliver</td>
</tr>
</tbody>
</table>
| **Object of Study:** | CMMI:  
1. Peer Review Process  
2. Unit Test Planning  
3. Development Integration Testing  
4. Support Tool for Software Development |
| **Context:**       | Projects in Internet Banking System development |
| **Measurement Level** | **Product** |
| **Point of View (Viewpoint):** | **Category** | **Focus (Success Indicator):** | **Measurement Goal** |
| Product Manager    | Primary | Time-to-market | MG08_PRD |
|                    | Complementary | Quality | MG09_PRD |
|                    |            | Cost | MG10_PRD |
| Project Team       | Primary | Time-to-market | MG11_PRD |
| Head of R&D        | Primary | Time-to-market | MG12_PRD |
|                    | Complementary | Quality | MG13_PRD |
|                    |            | Cost | MG14_PRD |

Table 31: Phase II: *Measurement Goal, Question(s) and Metrics* for BG2

| **MG08_PRD: Time-to-market** | Reusing the same Question and Metric from MG02_PRD, point of view is the Product Manager. |
| **MG09_PRD: Quality** | Reusing the same Questions and Metrics from MG01_PRD, point of view is the Product Manager. |
| **MG10_PRD: Cost** | Reusing the same Questions and Metrics from MG03_PRD, point of view is the Product Manager. |
| **MG11_PRD: Time-to-market** | Reusing the same Question and Metric from MG02_PRD, point of view is the Project Team. |
| **MG12_PRD: Time-to-market** | Reusing the same Question and Metric from MG02_PRD, point of view is the Head of R&D. |
| **MG13_PRD: Quality** | Reusing Question 3 and its metrics from MG01_PRD, point of view is the Head of R&D. |
| **MG14_PRD: Cost** | Reusing the same Question and Metric from MG07_PRD, point of view is the Head of R&D. |
In Figure 36 the GQM Tree for the Product measurement level is given. It illustrates how the elicited metrics can be traced back to the respective measurement and business goals.

**GQM Tree for Phase II (Product only)**

**BG1: Improve product quality**

- **Quality**
  - MG01_PRD
  - MG09_PRD

- **Time-to-market**
  - MG02_PRD
  - MG06_PRD
  - MG08_PRD
  - MG11_PRD
  - MG12_PRD

- **Cost**
  - MG04_PRD
  - MG15_PRD

- **Quality**
  - MG03_PRD
  - MG07_PRD
  - MG10_PRD
  - MG14_PRD

- **Cost**
  - MG05_PRD
  - MG09_PRD

**BG2: Decrease time-to-deliver**

- **Quality**
  - MG03_PRD
  - MG07_PRD
  - MG10_PRD
  - MG14_PRD

- **Cost**
  - MG05_PRD
  - MG09_PRD

Figure 36: GQM tree for Phase II (Product only)

### 6.3.3 Design – Selection of evaluation methods

In Phase I, some examples of metrics in *Metric Form* for Process and Project level have been shown. Here, only two metrics for the Product level are shown (Table 32 and Table 33) for the sake of brevity.

**Table 32: Phase II: Metric Form for M04_PRD**

<table>
<thead>
<tr>
<th>Metric ID</th>
<th>M04_PRD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric Name</td>
<td>Customer satisfaction survey</td>
</tr>
<tr>
<td>Success Indicator</td>
<td>Quality</td>
</tr>
</tbody>
</table>
| Measurement Goal (Viewpoint) | i. MG01_PRD (Product Manager)  
ii. MG05_PRD (Head of R&D)  
iii. MG09_PRD (Product Manager)  
iv. MG09_PRD (Head of R&D) |
| Measurement Level | Product |
| Algorithm / Description | In order to effectively evaluate the customer satisfaction, one of the best ways is to hear it from the voice of the customers themselves. This can be done by preparing a customer satisfaction survey addressing the achieved product quality.  
Example: The customer satisfaction survey can be designed using Likert |
Scales consisting of questions about the functionality, usability, and other quality attributes while using the product.

Baseline Value: The baseline for the customer satisfaction survey can be obtained from similar surveys that were previously conducted from past product releases. Based on the data from past surveys, the customer satisfaction was rated as Average.

Deviation Range: N/A

Evaluation Method: Survey (Questionnaire)

Objective: The objective of the evaluation is aimed to find out if the quality of the product got affected in a positive way as a result of the implementation of the SPI initiatives.

Confounding Factor(s): While conducting survey, the following factors should be taken into consideration in the evaluation:
1. Evaluation sampling (New or existing target customers for survey)
2. Subjects’ domain (Different industry or domains have different levels of expectation towards the product)
3. Subjective evaluation (Different customer have different levels of expectations)
4. Degree of usage (Number of features actually used, total usage time, maturity of usage)

Table 33: Phase II: Metric Form for M08_PRD

<table>
<thead>
<tr>
<th>Metric Form</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric ID</td>
<td>M08_PRD</td>
</tr>
<tr>
<td>Metric Name</td>
<td>Product cost normalized by size</td>
</tr>
<tr>
<td>Success Indicator</td>
<td>Cost</td>
</tr>
<tr>
<td>Measurement Goal (Viewpoint)</td>
<td>i. MG07_PRD (Head of R&amp;D) ii. MG14_PRD (Head of R&amp;D)</td>
</tr>
<tr>
<td>Measurement Level</td>
<td>Product</td>
</tr>
</tbody>
</table>
| Algorithm / Description | Product cost = (E + M) / S  
E = Total effort for the whole product life cycle in man-month  
M = Maintenance cost that involves the cost of fixing post-release defects and all the necessary support cost calculated in man-month  
S = Product size in KLOC  
Although cost is not the improvement goal of the SPI initiatives, it is still a major concern to the Head of R&D. He would be interested to see if the SPI initiatives would give any positive impact to the cost of the product in the long-run.  
Example: The total effort needed for the development of the product is 10 man-months and the maintenance/support cost after the delivery of the product to customer is 4 man-months. If the product size is 20 KLOC, then the product cost normalized by size is calculated as (10 + 4) / 20 = 0.7 man-months per KLOC. |
Baseline Value: The baselines for this metric can be computed based on the data collected from previous projects or calculated from current projects (without the SPI initiative implemented) that are running. Based on the historical data in the organization, the baseline value for product cost is 1.2 man-months per KLOC.

Deviation Range: The deviation range is set to ±10%

Evaluation Method: Pre-post comparison

Objective: The objective of the evaluation is aimed to find out if the SPI initiatives are able to reduce the product cost in general as a result from the savings of defects containment effectiveness during development.

Confounding Factor(s): During pre-post comparison, the following factors should be taken into consideration in the evaluation:
1. Project size and complexity (No. of features)
2. Technology factors (Language, tool, environment, platform)
3. Employee factors (Level of experience, turnover)
4. Process Compliance (Degree of process conformance implemented and followed)

6.3.4 Design – Evaluation scheduling

After the required metrics have been defined, the evaluation on the products within the Internet Banking System development can now be scheduled as shown in Table 34.

Table 34: Phase II: Evaluation Scheduling Form for Product

| Duration: | Starting from January 2007 onwards |
| Context: | The SPI initiatives are implemented in the whole Internet Banking System Department and are applied across all the products that are developed in that department. At the time to evaluate, the candidate products to be evaluated will be chosen based on the products schedule that are active at that point in time. |
| Time to evaluate / Periodic evaluation | Measurement Levels | Lag Factor (LF) | Degradation Factor (DF) | Time to Evaluate |
| Product | 5 months | 4 months | 4 months |
| The usual product cycle time is four months. However, it needs to be considered when the customer feedback might be available to integrate it in the evaluation. Therefore, the LF is set to be one month later after the product is released to customer. | Since the usual product release cycle is four months, the product release cycle time is taken as the DF to enforce a periodic evaluation after every product release. | The first evaluation shall be done after the first five months and periodic evaluations shall be done no later than four months after the first evaluation. In this case, the evaluation will be conducted every four months. |
Figure 37 illustrates the product’s evaluation scheduling for Phase II. The candidate target entity (product) to be evaluated at the scheduled date will be chosen based on the product’s schedule that are active at that point in time.

### 6.3.5 Implementation – Establishment of measurement program

As in Phase I, the elicited metrics are investigated to find out which ones are reusable and which ones are new and have not been implemented or yet used within One Solution (Table 35). The metrics are then established using previous knowledge and guidelines for measurement program establishment that One Solution already has.

Table 35: Phase II: New and reusable metrics

<table>
<thead>
<tr>
<th>Measurement levels</th>
<th>Reusable/existing metrics</th>
<th>New metrics to be implemented</th>
</tr>
</thead>
</table>
| Product            | 1. M01_PRD: Post-release fault density  
3. M05_PRD: Time-to-Deliver  
4. M08_PRD: Product cost normalized by size | 1. M02_PRD: Mean Time Between Failures  
2. M03_PRD: Percentage of delivered features  
3. M06_PRD: Relative cost of System Test  
4. M07_PRD: Cost of rework normalized by size |

### 6.3.6 Implementation – Conduct evaluation

For the sake of brevity, only the Product evaluation results are shown. The evaluations are conducted as according to the schedule on the Products that were active at the time of evaluation.

Figure 38 shows the summary of the evaluation results in the Product level. From the evaluation results obtained on the four metrics, the overall product quality shows a positive indication of improvement. The customer satisfaction survey turned out to give a very positive indication whereas the percentage of delivered features slightly dropped. Regarding the other two success indicators (time-to-market and cost), both show a good sign of improvement after one year. All the stakeholders, especially the Head of R&D, are satisfied.
with the results and the organization is ready to further expand the implementation of SPI initiatives to the other two major products of One Solution.

In order to see a better representation of evaluation results, the next step in the scenario presents the holistic view creation that shows the subjective value of the improvement from different viewpoints.

Product Level Evaluation Results – Phase II

6.3.7 Implementation – Holistic view creation

At the end of year 2007, One Solution created a holistic view for each measurement level, showing the different subjective values of the improvement from different viewpoints. Every stakeholder that is interested to see and evaluate the improvement initiatives was involved in the activity of rating the metrics in order to come out with the final holistic views.

Since the impact rating of the improvement is subjective in nature, there was a meeting between the stakeholders to define guidelines on how to perform the rating with the aim to
improve the consistency in the rating outcomes. The impact rating is based on a Likert Scale from -5.0 up to +5.0 as shown in Table 36.

Table 36: Impact rating guidelines

<table>
<thead>
<tr>
<th>Impact Rating (IR)</th>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>High gain</td>
<td>+5.0</td>
<td>Significant improvements observed. Actual values of improvements usually above 40% with respect to baseline.</td>
</tr>
<tr>
<td>Considerably high gain</td>
<td>+4.0</td>
<td>Improvements recorded in between high gain and moderate gain range.</td>
</tr>
<tr>
<td>Moderate gain</td>
<td>+3.0</td>
<td>The initiatives show an average gain. Actual values of improvements usually around 20% above baseline.</td>
</tr>
<tr>
<td>Considerably low gain</td>
<td>+2.0</td>
<td>Improvements recorded in between moderate gain and low gain range.</td>
</tr>
<tr>
<td>Low gain</td>
<td>+1.0</td>
<td>There is slight improvement visible. Actual values of improvements usually around 10% above baseline.</td>
</tr>
<tr>
<td>No impact</td>
<td>0.0</td>
<td>There is no significant impact observed. Usually the actual values that varies less than 5% above or below the baseline fall in this category.</td>
</tr>
<tr>
<td>Low loss</td>
<td>-1.0</td>
<td>There is slight loss visible from the improvement initiatives. Actual values usually around 10% below baseline.</td>
</tr>
<tr>
<td>Considerably low loss</td>
<td>-2.0</td>
<td>Decline in actual values recorded in between moderate loss and low loss range.</td>
</tr>
<tr>
<td>Moderate loss</td>
<td>-3.0</td>
<td>The initiatives show an average negative impact. Actual values of improvements usually worsen around 20% below baseline.</td>
</tr>
<tr>
<td>Considerably high loss</td>
<td>-4.0</td>
<td>Decline in actual values recorded in between high loss and moderate loss range.</td>
</tr>
<tr>
<td>High loss</td>
<td>-5.0</td>
<td>Significant decline observed. Actual values usually worsen by around 40% with respect to baseline.</td>
</tr>
</tbody>
</table>

Note: Table 36 only applies to One Solution and it cannot be used in other companies. The descriptions in the impact rating are mostly discussing on the percentage of changes. However, there are some metrics that are not objective in nature in which they cannot really be judged in an objective manner. The table only serves as a guideline and the impact rating is still subject to the interpretation of the individual viewpoints.

Table 37 shows the details of the subjective value of improvement calculation that takes into account the subjective weight of the metrics together with its impact rating. Similar calculations were performed for the Process and Project level and based on the obtained values; the holistic views for Process, Project and Product level are illustrated in the Figure 39.
### Table 37: Holistic view calculation for Product

<table>
<thead>
<tr>
<th>Viewpoint</th>
<th>Metrics</th>
<th>Subjective Weight (SW)</th>
<th>Impact Rating (IR)</th>
<th>Subjective Value of Improvement (SVI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Head of R&amp;D (Sponsor)</strong></td>
<td>M03_PRD</td>
<td>0.10</td>
<td>0.0</td>
<td>(0.10)(0.0) + (0.40)(+4.0) + (0.10)(+1.0) = +2.50</td>
</tr>
<tr>
<td></td>
<td>M04_PRD</td>
<td>0.40</td>
<td>+4.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M05_PRD</td>
<td>0.40</td>
<td>+2.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M08_PRD</td>
<td>0.10</td>
<td>+1.0</td>
<td></td>
</tr>
<tr>
<td><strong>Product Managers (Coordinator)</strong></td>
<td>M01_PRD</td>
<td>0.10</td>
<td>+3.0</td>
<td>(0.10)(+3.0) + (0.10)(0.0) + (0.15)(-1.0) + (0.25)(+3.0) + (0.25)(+2.0) + (0.05)(+2.0) + (0.10)(+1.0) = +1.60</td>
</tr>
<tr>
<td></td>
<td>M02_PRD</td>
<td>0.10</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M03_PRD</td>
<td>0.15</td>
<td>-1.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M04_PRD</td>
<td>0.25</td>
<td>+3.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M05_PRD</td>
<td>0.25</td>
<td>+2.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M06_PRD</td>
<td>0.05</td>
<td>+2.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>M07_PRD</td>
<td>0.10</td>
<td>+1.0</td>
<td></td>
</tr>
<tr>
<td><strong>Project Team (Implementer)</strong></td>
<td>M01_PRD</td>
<td>0.60</td>
<td>+3.0</td>
<td>(0.60)(+3.0) + (0.40)(+1.0) = +2.20</td>
</tr>
<tr>
<td></td>
<td>M05_PRD</td>
<td>0.40</td>
<td>+1.0</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 39: Holistic view presentation for the Process, Project and Product level**
From the holistic views in Figure 39, the SPI initiatives implementation can be rated as successful because all three Process, Project and Product level show a positive impact from the point of view of all the stakeholders. It is interesting to see that in different measurement levels, the stakeholders who are most satisfied in the respective levels differs from one another; Implementer is most satisfied in the Process, Coordinator is most satisfied in the Project while the Sponsor is most satisfied in the Project and Product level. The company is now in a good situation to roll out the initiatives to a wider scope since all stakeholders perceive benefits from the change.

6.3.8 Evolution – Evaluation feedback analysis

In this stage, the post-mortem analysis for the second phase evaluation is conducted. The stakeholders and the evaluation team had a meeting to discuss the existing issues along with the solutions.

The first discussed issue is concerning the Customer Satisfaction Survey. The evaluation team lacks of members who possess the required experience to conduct the survey efficiently. The proposed solution for this problem is to recruit personnel who are qualified in conducting surveys or to train the evaluation team for conducting future surveys.

The second discussed issue discussed in the meeting concerns the confidence on the evaluation results. The stakeholders feel that the confidence in the evaluation results should be improved. They suggest implementing cross-examination in the next phase for the three measurement levels.

6.4 Phase III

In this phase, the initiatives are expanded to two new products, Loan System and Deposit System products. The aim of this phase is to implement the initiatives organization-wide, as the requirement of CMMI is to implement standardized processes at the organizational level. The description about this phase’s evaluation will be presented by giving a summary of the stages and only the main outcomes of each stage are illustrated.

6.4.1 Design stage

This stage consists of four steps but only the outcome of two steps will be presented, evaluation scoping and determination of measures. The evaluation method and evaluation scheduling in this phase are not discussed here because the same activities were already explained in the previous two phases. For evaluation scoping and the determination of measures, only Organization and External levels are shown.

In evaluation scoping, three stakeholders are identified: the Board of Directors, Development Departments (Head of R&D, Head of Department, and Project Team) and the Company Shareholders. In determining the measures, these stakeholders come out with the success indicators that are considered as important from their point of view and the steps involved in deriving those measures are similar to the previous phases. Table 38 shows the combined outcome for the Organization and External level.
### Table 38: Success indicators and metrics for Organization and External level

| Improvement Goal:                      | 1. Improve product quality  
<table>
<thead>
<tr>
<th></th>
<th>2. Decrease time-to-deliver</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPI Initiatives:</td>
<td>CMMI:</td>
</tr>
<tr>
<td></td>
<td>1. Peer Review Process</td>
</tr>
<tr>
<td></td>
<td>2. Unit Test Planning</td>
</tr>
<tr>
<td></td>
<td>3. Development Integration Testing</td>
</tr>
<tr>
<td></td>
<td>4. Support Tool for Software Development</td>
</tr>
<tr>
<td>Context:</td>
<td>Development in the following products:</td>
</tr>
<tr>
<td></td>
<td>1. Internet Banking System</td>
</tr>
<tr>
<td></td>
<td>2. Loan System</td>
</tr>
<tr>
<td></td>
<td>3. Deposit System</td>
</tr>
<tr>
<td>Measurement Level</td>
<td>Point of View (Viewpoint)</td>
</tr>
<tr>
<td>---------------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Organization</td>
<td>Coordinator: Board of Directors</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Implementer: Development Departments (Head of R&amp;D, Head of Department, Project Team)</td>
</tr>
<tr>
<td></td>
<td>Sponsor: Company Shareholders</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>External</td>
<td>Coordinator: Board of Directors</td>
</tr>
<tr>
<td></td>
<td>Sponsor: Company Shareholders</td>
</tr>
</tbody>
</table>

### 6.4.2 Implementation stage

In this stage, only the holistic view creation step is presented as the other steps are similar with the ones from previous phases with the only difference on the used target entities and measures. At the end of year 2008, One Solution conducted another holistic view evaluation of the improvement in all measurement levels. In order to have a “one-view” representation of the improvement, the individual holistic views at each measurement level are combined into an average value of improvement, AVI, for each measurement level. Figure 40 shows the combined holistic view for all measurement levels. Overall, the SPI initiatives have shown a very positive impact to One Solution with all the measurement levels showing some degree of improvement after two years of implementation. The inner measurement levels, the Process and Project levels show the most visible impact from the improvement initiatives as compared to the other levels.
6.4.3 Evolution stage

In this stage, a post-mortem analysis is conducted to evaluate the evaluation process of Phase III.

The first issue discussed during the meeting is concerning the evaluation scheduling for the Organization and External level. The stakeholders feel that the evaluation for these two levels was conducted too early, and therefore, it failed to show significant improvements in the Organization level as they were expected. The evaluation team also realizes this problem and agrees to change the evaluation schedule for these two measurement levels.

The second discussed issue concerns the entities where the evaluation team collected the data for the External level. The evaluation team sampled two clients of One Solution to get the data. The stakeholders propose to get the data from all of their clients. However, this idea is quite difficult to be carried out as One Solution’s clients are more than 20 companies and dispersed in many countries. Therefore, it will cost too much to gather data from all clients for the evaluation process. Finally, it is agreed to sample four clients from different countries.
7 Static Validation

One of the early and most often cited definitions of validation (by unofficial count) is given by O'Keefe et al. [109]. According to O'Keefe et al. "validation refers to building the right system (that is, substantiating that the system performs with an acceptable level of accuracy)" [110]. However, this definition is rather general in a sense that it does not explain what accuracy is and what the benchmark should be for this [109]. Adrion et al. provided another early definition of validation, which is "determination of the correctness of the final program or software […] with respect to the user needs and requirements" [111]. This definition states what the baseline should be, although it is defined specifically for conventional computer software. The United States Department of Defense (DoD) defined validation as "the process of determining the degree to which a model is an accurate representation of the real-world from the perspective of the intended uses of the model" [112]. Several other definitions of validation exist in literature which are summarized and compared nicely in [109]. In the context of this thesis, motivated by the definition of DoD, validation is defined as the process which determines the degree to which the concepts/features introduced by a model/framework can actually fulfill the needs of practitioners in the real-world. Since the necessity of the framework introduced in this thesis was gathered through a systematic review of published peer reviewed articles of academic researchers and industry practitioners (industry reports) [28], the subjects who validate this framework are selected from the population of academic researchers and practitioners from industry.

![Diagram of Static Validation Steps](image)

**Figure 41: Static validation steps**

Validation can usually be realized in two ways: static validation and dynamic validation [95]. In static validation, instead of applying the framework in live projects in industry, a thorough presentation of the framework is given to the subjects who do the validation and then they give feedback regarding its applicability in real projects according to their own opinion [95]. On the other hand, in dynamic validation, a method or framework is applied on
live projects and feedback is gathered based on the difficulty found in applying it [95]. Due to the time requirements to implement SPI-MEF in a real industrial setting, dynamic validation was not considered as feasible within the scope of this thesis. Therefore, a static validation is implemented to validate the framework. The main purpose of the static validation is to gather expert opinion on SPI-MEF, in terms of its correctness, applicability, usability and usefulness in practice. The static validation was structured in six steps (planning, designing, piloting, conducting, analyzing and framework refinement based on analysis of the results) as depicted in the Figure 41. Details of each of these steps are described in the following sections.

7.1 Validation scoping

The purpose of the validation scoping step is to select the main parts of the framework that need to be validated. The scoping is required for an effective and efficient design of validation, i.e. covering all major aspects of the framework taking time constraints into consideration.

The main objective of the validation was threefold. Firstly, the objective was to validate the core concepts which are tied with different steps of the framework. Therefore, at first, the steps and core concepts that were deemed to be important for validation were identified. The steps that were chosen for validation are: “Gap Analysis of Evaluation Quality”, “Evaluation Scoping”, “Determination of measures”, “Selection of evaluation methods”, “Evaluation Scheduling”, and “Holistic View Creation”. The concepts that were chosen from these steps were: “Evaluation Opportunity Matrix”, “Evaluation Area”, “Primary and Complementary Measures”, “Derivation of Measures Using GQM”, “Confounding Factors”, “Time to Evaluate” and “Holistic evaluation”. Figure 42 illustrates the steps and corresponding concepts. The step “Gather SPI Initiative Information” was not chosen for validation because it comprises only the documentation of SPI initiative information that is relevant for the purpose of evaluation. The step “Establishment of Measurement Program” needs no validation since providing a guideline of how to establish a measurement program is not in the scope of the framework. In this step, SPI-MEF only suggests that organizations should decide how to adopt/incorporate the metrics identified for SPI evaluation in their existing measurement program. The steps “Conduct Validation” and “Evaluation Feedback Analysis” cannot be statically validated but could be validated practically by a dynamic validation in future work.

The second objective of the validation is to gather expert’s opinion regarding the applicability of the framework in terms of its usability, usefulness and capability of providing a systematic way for SPI evaluation. Therefore, a brief overview of the whole framework describing each stages and steps was also considered within the scope of validation so that questions regarding applicability of the framework can be answered by the validators.

The third objective of the validation is to elicit new aspects which were not considered or to identify issues which need to be addressed in the refinement of the framework. Therefore, it was planned to ask the validators for specific or overall suggestions to improve the framework.
Figure 42 illustrates the topics/concepts that were selected for validation. A detailed description of all the stages of SPI-MEF and their corresponding steps and concepts can be found in Table 15 in Section 5.5.

Each topic/concept that was validated has a specific aim. Table 39 summarizes the topics/concepts covered in the validation and the aim for validation.

<table>
<thead>
<tr>
<th>No</th>
<th>Topics/concepts for validation</th>
<th>Aims</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Evaluation area</td>
<td>To validate if the concepts of measurement levels and viewpoints are appropriate generalizations. Furthermore, to investigate if the proposed evaluation area is helpful to characterize the sources and destinations of information in improvement evaluation.</td>
</tr>
<tr>
<td>2</td>
<td>Primary and complementary measures</td>
<td>To validate the concept of primary and complementary measures in the context of improvement evaluation. Additionally, it is assessed if the metaphor of the &quot;Iron Triangle&quot; is useful to create a mindset which supports the elicitation of the appropriate success indicators.</td>
</tr>
<tr>
<td>3</td>
<td>Derivation of measures using GQM</td>
<td>To validate the proposed approach to derive the needed measures for improvement evaluation. Since the approach incorporates the previously discussed concepts and uses GQM as a supporting tool it is crucial to</td>
</tr>
</tbody>
</table>
validate to which extent it is useful in practice.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Confounding factors</td>
</tr>
<tr>
<td>5</td>
<td>Time to evaluate</td>
</tr>
<tr>
<td>6</td>
<td>Holistic Evaluation</td>
</tr>
<tr>
<td>7</td>
<td>Evaluation opportunity matrix</td>
</tr>
<tr>
<td>8</td>
<td>Applicability</td>
</tr>
<tr>
<td>9</td>
<td>Suggestions</td>
</tr>
</tbody>
</table>

### 7.2 Validation design

The validation design step consists of three sub-steps: planning, actual design and piloting the validation. These three steps are described in detail in the following sections.

#### 7.2.1 Planning

Before the actual design of the validation, an initial planning was conducted. It includes determining the subjects that will take part in the validation, contacting the subjects, choosing the methods to use for the validation, and scheduling the validation. It was decided to validate the framework by both academic researchers and industry practitioners so that both theoretical aspects and the practicality of the framework can be assessed. However, more emphasis was given to justify the applicability of the framework in practice. Therefore, it was decided to select the academic researchers having industry experience or working closely with industry.

The researchers’ expert opinion is gathered through an interview. Two forms of interviews are conducted: personal interview and telephone interview. The personal interview is conducted when it is possible to meet the researcher in person and a face-to-face interview can be arranged. A telephone interview is conducted if it is difficult to meet the researcher in person due to his geographical location. For industry practitioners, besides face-to-face interviews, a self-administered questionnaire is used to gather their expert opinion. The intention is to get as much feedback as possible from industry. Therefore, if it is difficult to arrange a face-to-face interview with an industry practitioner, the questionnaire is sent to the subject such that he can compile and send it back at a convenient time. Thereafter, 13 academic researchers and 11 practitioners from industry were selected and contacted. Finally, nine researchers and seven practitioners from the industry agreed to participate in the validation of the framework. All academic researchers agreed to provide around 45-60 minutes for the interview and industry practitioners were willing to schedule a 1½ hour meeting. The industry practitioners to who the self-administered questionnaire is sent, the required time commitment for its compilation was of lower priority. However, it was decided to make the questionnaire short and precise so that it does not take much effort
and time for its compilation. Based on the estimated time for each interview a schedule for meetings is prepared. For the 10 interviews (nine face-to-face interviews and one telephone interview) one full week is planned that includes the preparation, execution and transcription of the interviews. It was decided to do at most two interviews per day. Besides, the self-administered questionnaire is also sent at the same time to the remaining participants.

7.2.2 Design

The main challenge in the validation design was to develop strategies which transfer the authors’ ideas most effectively and efficiently to the experts in order to elicit an informed opinion. There were mainly two restrictions which had to be taken into consideration. First, after an initial inquiry of the expert candidates, it was clear that the allocated time for a participation was limited (approximately one hour, to show the order of magnitude). Second, mostly for the industry experts, a face-to-face meeting was impossible due to their geographical location.

As discussed before, two methods are used for the purpose of validation: interview (face-to-face and telephone) and self-administered questionnaire. Although the ways how each of these methods is conducted differ, the contents that need to be validated are the same. Therefore, in the design of interview and questionnaire it is tried to keep the content as similar as possible so that the results can easily be compared and combined. Figure 43 shows how the validation through expert opinion is designed for this study and how the results are combined for the purpose of analysis. The following sections describe the interview and questionnaire design in more detail.

Figure 43: Approach for validation through expert opinion
7.2.2.1 Interview design

The interview (personal interview and telephone interview) is designed in a similar way for both the academic researchers and industry practitioners following the steps mentioned in research methodology chapter (see Section 4.1.2.1). The selected interview type is structured interview. Therefore all the questions that are asked to the interviewees are prepared beforehand. To each interviewee the same questions with exactly the same order are presented. This ensures that answers can be reliably aggregated. At first the questions that need to be asked are identified and categorized as high, medium and low priority. Prioritization of the questions is important because the interview should be designed in a way such that it can be finished in the stipulated amount of time while considering all high priority questions.

To clarify the purpose of the interview and the contents that need to be validated, a distilled description (10 pages) of the frameworks' concepts is prepared and sent as preparatory documentation to the interviewees. Additionally, a short presentation is held at the beginning of each interview meeting to give an introduction to the framework and to give a refresher of its main concepts. The preparatory documentation also helps to minimize the time restriction problem as mentioned earlier by providing the required information and explanation beforehand to the interviewee.

Some general guidelines and a time plan of the interview are also prepared in the design of the interview. All the interviews are conducted by the same interviewer to maintain the consistency in the way the questions are presented and three note takers record the answers. Only the interviewer asks questions to the interviewee and the note takers interfere only if an answer could not be understood properly. The interview meeting starts with a short presentation of 10 minutes followed by the actual conduction in which the questions are asked. The note takers also keep track of the time and notify the interviewer after half of the allocated time has elapsed. All the notes taken at the time of the interview are transcribed into a document, and later, after all interviews are conducted, compiled into a single document. The recorded answers in this compiled document are analyzed to elicit refinements to the concepts and the presented preliminary framework.

The interview results are verified in terms of their generalizability, reliability (how consistent the results are) and validity (whether the interview investigates what is intended to be investigated). How the generalizability, reliability and validity issues are addressed is discussed more in detail in “Validity Threats” (Chapter 8). The next sections describe the design of interview questions in more detail.

The expert opinion elicitation questions are designed slightly differently for the interviews in academia and in industry, although most of the concepts are covered in both. Topic 6 is handled exclusively in the academia interviews since it was regarded as too difficult to judge for the industry interviewees and covers solely theoretical aspects. Furthermore, since the topics were ranked beforehand by their priority and Topic 6 was judged as of moderate priority, it was decided to drop it from the industry interview to allow more time for the remaining questions.

Since one interview is conducted through the telephone, the questions and the preparatory material is sent to the interviewee one week in advance in order to render the communication more efficient and to avoid misunderstandings. For the face-to-face meetings, only the preparatory material is sent to the interviewees in advance. Table 40
describes the topics (core concepts) and questions covered in both academic and industry interview.

Table 40: Academia / Industry interview questions

<table>
<thead>
<tr>
<th>No</th>
<th>Topic</th>
<th>Interview questions</th>
</tr>
</thead>
</table>
| 1  | Evaluation area | IW1.1: Do you think that the evaluation area represents clearly WHO is interested in the evaluation and WHICH measurable entities need to be considered?  
IW1.2: In your opinion, is the generalization of the viewpoints in Implementer, Coordinator and Sponsor appropriate? |
| 2  | Primary and complementary measures | IW2.1: Do you think that the consideration of primary and complementary measures increases the accuracy for improvement evaluation?  
IW2.2: Do you think that the derivation of complementary measures (based on the principle of the iron triangle) clarifies WHAT should be measured and WHY it should be measured? |
| 3  | Derivation of measures | IW3.1: Do you think that the proposed approach provides a clear and systematic method for the elicitation of measures of improvement evaluation?  
IW3.2: Is the proposed interface between SPI-MEF and GQM appropriate?  
IW3.3: To what extent do you consider the proposed approach useful and usable in practice? |
| 4  | Time to evaluate | IW4.1: Do you think that the determination of the appropriate time to evaluate using the concept of Lag Factor and Degradation Factor is sound?  
IW4.2: Do you think that periodic evaluation (to see the trend over time) is an effective way to gauge the outcome of SPI initiatives? |
| 5  | Holistic view | IW5.1: Do you think that the proposed model to construct the "Holistic View" is sound?  
IW5.2: Can you identify any improvement opportunity for the model? |
| 6  | Evaluation opportunity matrix (covered only in academic interview) | IW6.1: From your experience, is increasing accuracy and coverage the appropriate way to improve the quality of evaluation?  
IW6.2: What is, in your opinion, more important for a high quality evaluation (accuracy or coverage)? |
| 7  | Applicability | IW7.1: To what extent do you consider the idea that the evaluation of improvement is independent from the actual initiative as realistic?  
IW7.2: Do you think that the framework scales with increasing complexity of the SPI initiative?  
IW7.3: To what extent do you think SPI-MEF is
Suggestions

IW8.1: Can you name and explain any deficiency you have observed in the framework?

IW8.2: Can you name and explain the major benefits you have observed in the framework?

IW8.3: Do you have any suggestions to improve SPI-MEF in terms of applicability, realism or usefulness?

7.2.2.2 Questionnaire design

The self-administered questionnaire is designed following the steps mentioned in the research methodology (see Section 4.1.2.2). The steps that are followed in the questionnaire design are described briefly below:

i. **Determining the questions to be asked**: The questionnaire for industry practitioners covers the same topics as the industry interview (Topic 1-8 of Table 40 except Topic 6). This allows to combine the results of both industry interview and questionnaire. For each of the selected topics, the specific questions are formulated in a slightly different way than in the interview as described in the following step.

ii. **Selecting the question type/format for each question and specify wording**: For each of the concepts that need to be validated a closed ended question followed by a box to express the rationale behind the answer is prepared. The questions are formulated in a way such that the degree of agreement/disagreement can be expressed (using a Likert scale). It is assumed that such a strategy would be beneficial for both parties: the experts could more easily express their consent/discordance to a concept and the authors have an indication how to interpret the elaborated answer to the question, which was encouraged to state in a few sentences in the box given after each question. Besides, the wording of the questions is selected very carefully. Salant and Dillman described some common wording problems in questionnaire design in [114]. Vaguely worded questions, use of unconventional terminology, biasing questions, too demanding or too difficult questions, double negative questions, too much knowledge assumed on the part of the respondent, answer choices which are not mutually exclusive are typical pitfalls in formulating a good questionnaire. In the questionnaire design for interview experts, it is attempted to minimize these common problems as much as possible following the possible solutions mentioned by Salant and Dillman in [114].

iii. **Designing the question sequence and overall questionnaire layout**: The questionnaire document is structured in four different sections: industry expert background information, a brief introduction to the framework, key concepts validation questions and general validation questions (e.g. applicability of the framework in practice, general suggestions). Each key concept validation question is preceded by the aim for validation followed by a brief explanation of the concept itself. Definitions of terms are given along with the questions when it is required. Besides, figures are added to illustrate the concepts. The font size
and style are used consistently throughout the document to easily identify the sections, questions, answers, terms etc.

iv. Developing ancillary documents: In this step other supportive documents like a pre-notification email, a reminder email (to remind respondents to complete the questionnaire) and an appreciation email was prepared. A study of Yammarino indicates that repeated contacts in the form of preliminary notification and follow-ups are effective in increasing the response rates [115].

The topics covered in the questionnaire for industry experts are the same as the topics covered in the industry interview (see Table 40). However, Question IW7.1 and IW7.2 are replaced by question QT7.1 and QT7.2. This was decided because question IW7.1 and IW7.2 were found to be too demanding and hard to judge without a detailed explanation of how the framework actually works. On the other hand, in the face-to-face interview, these two questions are included since a short presentation of the complete framework is given at the beginning of the interview meeting.

Table 41 shows the covered concepts and the included questions in the questionnaire for industry experts. The full questionnaire that was sent to industry experts can be found in Appendix A.

<table>
<thead>
<tr>
<th>No</th>
<th>Topic</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Evaluation area</td>
<td>QT1.1: To what extent do you agree that the above schema (evaluation area) represents clearly WHO is interested in the evaluation and WHICH measurable entities need to be considered?</td>
</tr>
<tr>
<td>2</td>
<td>Primary and complementary measures</td>
<td>QT2.1: To what extent do you agree that the consideration of primary and complementary measures provides a more complete view on how you look at the effects of the SPI initiative? &lt;br&gt; QT2.2: To what extent do you agree that the derivation of complementary measures (based on the principle of the iron triangle) clarifies what should be measured and why it should be measured?</td>
</tr>
<tr>
<td>3</td>
<td>Confounding factors</td>
<td>QT3.1: To what extent do you agree that confounding factors can affect the validity of the evaluation results?</td>
</tr>
<tr>
<td>4</td>
<td>Derivation of measures</td>
<td>QT4.1: To what extent do you agree that the proposed approach provides a clear and systematic method for the elicitation of measures of improvement evaluation? &lt;br&gt; QT4.2: To what extent do you agree that the proposed approach is useful in practice?</td>
</tr>
<tr>
<td>5</td>
<td>Time to evaluate</td>
<td>QT5.1: Do you think that it is feasible for an organization to determine the value for the Lag Factor and Degradation Factor using their historical data and experience? &lt;br&gt; QT5.2: To what extent do you agree that the concept of Lag Factor and Degradation Factor can be used to determine the appropriate time to evaluate the outcome of SPI initiative?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
| 6 | Holistic view | **QT6.1:** To what extent do you agree with our approach to construct the holistic view?  
**QT6.2:** To what extent do you agree that consolidating the evaluation results into “one picture” (holistic view) gives benefits to the organization? |
| 7 | Applicability | **QT7.1:** To what extent do you agree that SPI-MEF is usable in practice?  
**QT7.2:** To what extent do you agree that SPI-MEF is useful (giving benefits) in practice?  
**QT7.3:** To what extent do you agree that SPI-MEF is adaptable and customizable to different company settings (size, core business, culture)? |
| 8 | Suggestions | **QT8.1:** Can you name and explain briefly any deficiency you have observed in the framework?  
**QT8.2:** Can you name and explain the major benefits you have observed in the framework?  
**QT8.3:** Do you have any suggestions to improve SPI-MEF? If yes, please state them. |

### 7.2.3 Piloting

Both the interview and the questionnaire are piloted beforehand. Regarding the interview three components are piloted: the interview preparatory documentation, the presentation (slides) summarizing the framework, and the conduction of interview. For the piloting two Software Engineering students were approached and asked for their participation. The purpose of piloting interview preparatory document is to verify its understandability and clarity in presenting the concepts of the framework. The presentation and interview conduction were piloted to determine their overall timing, completeness, clarity and understandability.

For the remote experts, a more elaborated summary of the framework with illustrative examples is prepared including the questionnaire covering the presented material. In [116] Kasuníc describes three purposes for piloting the questionnaire instrument: to expose the problems or weaknesses in the questions, questionnaire layout, and the process. The aim of piloting is to determine if the right questions are being asked, the questions are understandable, the answer choices are appropriate and no answer choices have been overlooked, and if there are any terms that are not understood. The questionnaire layout is evaluated to determine if there exist difficulties in navigating through the questionnaire, the font is easily readable, and if the conventions for font size, styles, and color are consistent and helpful to the respondent. The questionnaire document is appraised by a Software Engineering student to fulfill the aforementioned purposes.
7.3 Academic validation results

The validation was conducted with two different groups: academic researchers and industry practitioners. For academic researchers the personal interview and the telephone interview were the methods to elicit their expert opinion, while for industry practitioners the personal interview and the questionnaire are used. In this section the results of the academic validation are presented.

Nine researchers from academia participated in the academic validation. All but one interviewee are currently employed at the Blekinge Tekniska Högskola; nevertheless, they experienced education in various universities in Sweden, Germany, Australia and Turkey, which allows the assumption that their opinions and expertise are not streamlined. The time-frame of their academic experience ranges from three to thirty years, while their industrial experience ranges from two to eighteen years.

In Table 42 the background of the academic researchers is summarized. The information is anonymized for the purpose of confidentiality. Furthermore, as a convention for the remaining part of this document, female pronouns are used to refer to the interviewees, regardless their actual gender.

<table>
<thead>
<tr>
<th>No</th>
<th>Interviewee</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Interviewee A</td>
<td>She has eight years of academic experience with a major interest in Software Size Measurement and Estimation, Effort Estimation, Project Management, Software Process Improvement, Business Process Modeling and Requirements Elicitation. In her industrial career she worked as System Analyst, Project Coordinator and Consultant since 2001. She was involved in several CMMI Level 3 improvement initiatives as a consultant.</td>
</tr>
<tr>
<td>2</td>
<td>Interviewee B</td>
<td>Her research area is large scale software management, related to strategically managing large quantities of products and product requirements. Further research interests include software product management, software quality, economic issues in software development, and software process management.</td>
</tr>
<tr>
<td>3</td>
<td>Interviewee C</td>
<td>Her major research interests are Software Product Line Engineering (1.5 years) and Software Process Improvement, Agile and Lean Software Development and Software Measurement (1.5 years). She worked as a developer for service oriented applications and is currently embedded as a researcher in a major telecommunications company where she is part of a unit concerned with software process improvement and measurement. She was involved in SPI initiatives by driving an organization towards a lean software process using measurement influenced by lean principles and has experience in assessing the achieved benefits.</td>
</tr>
<tr>
<td>4</td>
<td>Interviewee D</td>
<td>She has three years of academic experience with the main interest in Value-Based Software Engineering. She worked for a total of four years in the software development industry, occupying roles as Support Responsible, Business Analyst and Developer. She gained experience in assessing the achieved benefits of software process improvements through her research work.</td>
</tr>
<tr>
<td>5</td>
<td>Interviewee E</td>
<td>She conducts empirical research in Software Verification &amp; Validation</td>
</tr>
</tbody>
</table>
and Search-Based Software Engineering since seven years and has three years of industry experience.

6 Interviewee F
[Int_F]
She has three years of academic experience with a major interest in Strategic Software Engineering and Software Product Management. Her industry experience includes four years of work as Software Developer and Team Leader, and two years of research collaboration with industry. She was involved in SPI initiatives by improving the verification and validation process (early verification) and by improving the strategic decision making process.

7 Interviewee G
[Int_G]
Her major research areas are Requirements Engineering and Software Architectures in which she is involved since ten years.

8 Interviewee H
[Int_H]
She has thirty years of academic experience with the main interest in Software Engineering (specifically architecture/reuse, process engineering, measurement, and empirical studies), Embedded Systems / Ambient Systems, Technology Transfer and Software Process Improvement. She has ten years of industry experience where she worked as a consultant for many companies. Her involvement in software process improvement spans from being a consultant, where she also conducted formal assessments, to the membership in the IEEE/SEI Process Achievement Committee, in which measuring the success and benefits of SPI was a part of her activities.

9 Interviewee I
[Int_I]
Her major research interests are Verification & Validation, Automated Software Engineering, Requirements Engineering, Psychological / Human / Social Aspects on Software Development and Search-Based Software Engineering in which she has been involved since twelve years. She worked for eighteen years as a software development and strategic consultant. Her involvement in SPI initiatives covers the optimization of verification and validation activities, the design and implementation of a SPI framework specifically adapted to V&V and consultancy activities. Formal assessments were part of her teaching activities and assessing the benefits of SPI, in terms of defect and effort reduction, were responsibilities in her industrial employment.

7.3.1 Evaluation area (Measurement levels and viewpoints)

Reference questions: IW1.1 and IW1.2

Aim: To validate if the concepts of measurement levels and viewpoints are appropriate generalizations. Furthermore, to investigate if the proposed evaluation area is helpful to characterize the sources and destinations of information in improvement evaluation.

In general, all the interviewees agreed that the consideration of viewpoints and measurement levels is important and appreciate the idea of considering these two aspects in evaluation scoping. Nevertheless, there were few suggestions for improvement. The term "Implementer" needs more clarification as it can include a wide range of roles starting from the people who are involved in the requirements engineering process to people who are involved in supply and packaging [INT_F]. Besides, it was found that sometimes it is hard to differentiate between different viewpoints. For instance, it may become hard to distinguish between the two viewpoints "Implementer" and "Coordinator" as both of these can be the ones who implement the improvement and can fall into the "Implementer" category [INT_H]. On the other hand, viewpoints should not be orthogonal or mutually exclusive to each other, as the same role can be part of different viewpoints [INT_C]. It was suggested to
investigate whether certain other viewpoints need to be incorporated, e.g. the customer as external viewpoint [INT_A], people involved in supporting the process [INT_A] and the roles which are not directly related with improvement initiative but could be interested to see the evaluation results [INT_I].

The term "measurement levels" was not clear and ambiguous to some extent [INT_E, INT_H and INT_I]. People could mix it with other terms used in this research area because level can also be associated with step-by-step capability improvement to attain a certain maturity level [INT_E]. Some other names like "measurement aspects" or "measurement artifacts" [INT_H] and “measurement areas” [INT_E] were suggested to use instead of "measurement level". The proposed ordering of the measurement levels is not always true for process, project and product [INT_C]. A process can span over many projects and, as such, last longer than the individual projects [INT_C]. There may be also an overlap in the measurement levels, in the sense that a measurable attribute cannot always be characterized as being part of one specific level alone [INT_C].

7.3.2 Primary and complementary measures

Reference questions: IW2.1 and IW2.2

Aim: To validate the concept of primary and complementary measures in the context of improvement evaluation. Additionally, it is assessed if the metaphor of the "Iron Triangle" is useful to create a mindset which supports the elicitation of the appropriate success indicators.

All interviewees agreed that the distinction between primary and complementary measures is a useful and important concept. Although at first glance the concept is quite obvious and may be already unconsciously practiced by companies [INT_D], the explicit consideration of complementary measures may help to reduce the temptation of optimizing the outcome of primary measures which are directly related to the improvement goal [INT_C]. By evaluating measures which are not directly related to the improvement goal, imbalances in the initiative can be detected [INT_D] and therefore evaluated more accurately. It is important to note that complementary measures should not only be seen as indicators of adverse effects but can also reveal additional benefits which were initially not associated with the improvement goal [INT_B]. Although the concept is understood, the term "complementary" may induce misunderstandings for industrial practitioners since it suggests that the measures are just considered additionally to the primary (and important) ones [INT_J]. Therefore companies could be reluctant to define complementary measures since collecting additional data is costly [INT_I].

It may be difficult to identify the appropriate complementary measures but the "Iron Triangle” metaphor provides support to some extent [INT_E] and limits the set of candidate measures [INT_F and INT_G]. Nevertheless, it must be made clear that the "Iron Triangle" is understood as a metaphor since the relationship between success indicators may not always be as trivial as suggested by the triangle [INT_J]. For the practitioner it is necessary to illustrate the metaphor with concrete examples to substantiate the idea [INT_A] and possibly provide guidelines for assessing the cost of measures to identify and select the most efficient ones [INT_F]. Furthermore, for practical use it should be considered to merge the success indicators from the different measurement levels into a single table [INT_E], since there is a potential overlap in which attributes they address (e.g. Project level (Schedule, Cost), Product level (Time-to-market, Cost)). The illustration as it was shown to the academic interviewees may confuse industrial practitioners [INT_C]. Additionally, it was
proposed to add functionality [INT_G] and company reputation [INT_H] as success indicators.

7.3.3 Derivation of measures using GQM

*Reference questions: IW3.1, IW3.2 and IW3.3*

*Aim:* To validate the proposed approach to derive the needed measures for improvement evaluation. Since the approach incorporates the previously discussed concepts and uses GQM as a supporting tool it is crucial to validate to which extent it is conceptually sound and useful in practice.

In general, the integration of the evaluation area and primary/complementary measurement concepts with the GQM approach was picked up positively and judged as a systematic and comprehensible method. This top-down approach supports the elicitation of the necessary subset of potentially useful measures and alleviates the identification of reusable measures [INT_C]. Since the method was illustrated with a very simplistic example, concerns arose about its applicability in industrial settings and its scalability to more complex situations [INT_F and INT_I]. Specifically, it may introduce additional costs to educate employees without giving a directly visible benefit to the company. Moreover, the complexity of applying this method is for implementers and coordinators considerably higher than for sponsors [INT_I]. A predefined pool from which goals, questions and metrics could be selected and appropriate guidelines would probably preferred by companies since the proposed method may be too abstract to be adopted [INT_I].

Scalability issues could be addressed by following an incremental approach, and iteratively expand the measurement goals and the involved stakeholders [INT_H]. In order to be for practical use, step-by-step guidelines are needed that clarify how to use the approach in a generic way [INT_G], e.g. it must be elaborated on how the concept of primary/complementary measures integrates with GQM [INT_C]. A first step in verifying the methods’ practicability could be done by conducting experiments in academic settings with students [INT_B].

The conceptual mapping between SPI-MEF and the facets of GQM is appropriate but the illustration needs to be explained and elaborated in more detail [INT_G, INT_E and INT_I]. Additionally, confounding factors should be mapped to target entities instead directly to the context facet in GQM [INT_E].

7.3.4 Time to evaluate

*Reference questions: IW4.1 and IW4.2*

*Aim:* To determine if the concepts of Lag Factor (LF) and Degradation Factor (DF) can help to identify the appropriate time for evaluation.

In general, all interviewees appreciated the idea of defining the time-bounds for improvement evaluation by a LF, DF and periodic evaluation and agreed that the idea is conceptually good. The concept of Lag Factor is quite intuitive but it could be difficult, if not impossible, to determine the Degradation Factor [INT_I]. DF is quite unclear when the initiative is introduced for the first time [INT_G]. However, it could be easy to come up with the values of LF and DF in some contexts, e.g. in domains where the product is purely composed of software [INT_D]. Companies that are involved in large systems and have a long product life cycle can have difficulties in defining the values for LF and DF [INT_D].
Determining the Lag Factor in practice could also be a problem, since it is quite dynamic in nature and very context dependent [INT_E]. It depends on the concrete project taken into consideration. Furthermore, the natures of projects are not consistent [INT_C]. In a company where many small changes/initiatives are introduced in parallel, it is difficult, if not impossible, to find the right time to evaluate and to decide LF for a particular initiative [INT_C]. However, it could be possible for a company to come up with the values for LF and DF by group work with the participating stakeholders in the initiative [INT_F]. It needs empirical evidences also from industry to determine LF and DF [INT_H].

It is important to consider the potential threat in determining the value of LF. Since LF is an estimate, it is potentially difficult to come up with the appropriate value and it could be wrong [INT_C]. Managers might underestimate LF as they want to see the benefits immediately and this can cause a problem [INT_D].

Since, it is not easy to come up with the values of LF and DF, unless a guideline is provided, it will be difficult to realize it in practice and organization may instead use their usual approach to determine the time to evaluate [INT_G]. Therefore, methods / guidelines / instructions to come up with the exact value of LF and DF are required [INT_E and INT_G] since they are still missing in the framework [INT_B].

The interviewees partially did not agree with the heuristic provided for LF and DF that their values increase with the measurement levels, that is, “\(LF_{Prc} < LF_{Ppj} < LF_{Ppd} < LF_{Org} < LF_{Ext}\)” and “\(DF_{Prc} < DF_{Ppj} < DF_{Ppd} < DF_{Org} < DF_{Ext}\)”. \(LF_{Prc}\) can be greater than \(LF_{Ppj}\) since a process may span over several projects [INT_C]. \(LF_{Prc}\) can even be longer than \(LF_{Ppd}\) since some process can span over several products [INT_G]. Similarly some projects may have a longer LF than product [INT_F] and some products can have a longer LF than organization [INT_F and INT_I].

The concept of periodic evaluation was picked up positively although there were concerns regarding the frequency of evaluation. The frequency of evaluation depends on the context of the initiative and what is being evaluated [INT_I]. Determining the optimum interval between successive evaluations is important in order to minimize the cost of evaluation. However, the interval may also change over time and therefore, the evaluation plan needs to be adapted over time to balance the costs and benefits of the evaluation [INT_E]. It can also be the case that the time to evaluate is constrained by the project or product life-cycle since there may exist bottlenecks between the transitions of development phases and backlogs of work may skew the evaluation results [INT_C]. Evaluating the improvement periodically and showing the results to the involved people can also improve the commitment to the initiative [INT_A].

7.3.5 Holistic evaluation

Reference questions: IW5.1 and IW5.2

Aim: To validate the model on which the "Holistic evaluation" is based and to identify potential deficiencies.

The idea to illustrate the improvement evaluation as it was proposed by the "Holistic evaluation" was generally appreciated. Since the model reduces the amount of information due to scale conversions, the target audience for the "Holistic evaluation" would be top-level managers rather than SPI coordinators which need more detailed information for decision making support [INT_H]. Moreover the combination of the three generic viewpoints by
using the arithmetic mean reduces the information content; a chart representing each viewpoint would be preferable [INT_G]. The Gain/Loss component seems to increase the complexity of the model, without providing additional information. It could be integrated in the Impact Rating component [INT_I].

Several interviewees remarked to pay attention on the measurement scales used in the components of the model and the respective admissible mathematical operations. Since the Impact Rating component has an ordinal scale, addition and multiplication are not allowed [INT_A]; investigating how this problem is addressed in the calculation of Function Points, where similar issues are encountered, could be helpful [INT_B]. Since, in general, the Impact Rating is based on comparison to a baseline it is not optimal to construct it basing on subjective assessments by the viewpoints, given that a subjective component is already represented by the Subjective Weight [INT_E]. It would be therefore advisable to express the improvement impact in an objective way, e.g. by percentages [INT_A and INT_C]. On the other hand, it was also noted that objective measures are difficult to collect in practice and therefore a subjective assessment may be permissible if concrete guidelines and checklists are given in order to increase the conformity of the ratings [INT_F].

Another important aspect to consider is that the viewpoint (or role) who suggested a measure is not always the appropriate one to assess the impact of the improvement [INT_C]. Furthermore, coordinating the holistic evaluation may be a problem since many stakeholders are involved, increasing the cost of the evaluation. Moreover, there may be cognitive aspects which need to be considered: people may report the same values even though there were small, but in their opinion, marginal changes; influence from management may skew the subjective rating (political influence). The validity of the data is a big challenge in every measurement framework and should be addressed as early as possible, i.e. before the evaluation [INT_I].

7.3.6 Evaluation opportunity matrix

Reference questions: IW6.1 and IW6.2

Aim: To evaluate if the usage of the opportunity ("2 x 2") matrix is the appropriate tool to guide the practitioner to a higher quality improvement evaluation.

Both the terms "opportunity" and "coverage" were considered as potentially misleading and intuitively not comprehensible in this context [INT_I], still the general concept was understood and judged by the interviewees as appropriate to represent the improvement opportunities of the evaluation. It was asserted that a "holistic" evaluation as suggested by the matrix is probably confined by the implementation cost [INT_D] and companies may strive for a "good enough" evaluation [INT_D]. Indeed, the cost factor, important to consider for practitioners, was seen as a crucial aspect still missing in the opportunity matrix [INT_G, INT_F, INT_E, INT_D and INT_I]. Furthermore it was noted that the improvement goal of the initiative should be taken into account in the matrix [INT_A].

Achieving simultaneously higher accuracy and better coverage would be the optimal solution since they are equally important [INT_A]. Clearly, such an ideal is constrained by the resources a company is willing to spend and the prospected gains [INT_D]. Achieving high accuracy first was considered as favorable for several reasons: first, raising the accuracy of the evaluation may decrease the temptation (and the possibility) to "cheat" on the measurements [INT_C]; second, a wider evaluation in terms of which measurable entities are
assessed, that is, increasing the coverage, increases the complexity of the evaluation by introducing new confounding factors [INT_E]; third, achieving first valid evaluation results is preferable, coverage can be addressed in later iterations [INT_H]; fourth, higher accuracy may be easier to achieve since less stakeholders are involved as in contrast to increasing the coverage [INT_J]. On the other hand, wider coverage is preferred over accuracy to see a more complete picture of the improvement benefits and problems can be identified more easily since the visibility of the change is taken to a higher level [INT_B, INT_F and INT_G].

7.3.7 Applicability

Reference questions: IW7.1, IW7.2 and IW7.3

Aim: To validate the general applicability of the previously discussed concepts.

All interviewees agreed that the approach to evaluate the improvement can be thought as independent from the concrete initiative. Hence, it is possible to define an evaluation framework which can be applied in different improvement contexts and SPI-MEF was assessed to be flexible and general enough to fulfill this purpose. Although its generic nature, it seems realistic and provides a systematic way for improvement evaluation [INT_E], without being too complex [INT_F]. Nevertheless, an actual implementation of the framework is needed in order to assess is applicability in realistic settings [INT_B]. Immature organizations should employ an iterative strategy and address the complexity by incrementally refining and enhancing the measurement and evaluation techniques [INT_A, INT_F]. The concept of LF and DF may not scale with the complexity of improvement initiatives, that is, in initiatives where small, in parallel changes are introduced [INT_C]. Similarly, confounding factors may be controllable in small (or even just artificial) settings and impossible to address in complex situations [INT_C], while, on the other hand, such a measurement and evaluation framework may even not be needed in small organizations [INT_C]. Traceability is an issue for multiple parallel improvement initiatives but explicitly showing that different initiatives overlap helps to interpret the evaluation results [INT_E and INT_G]. With more ambitious initiatives, the complexity for implementers and coordinators will rise considerably [INT_I]. Generally, the framework can be applied independently from the company settings; it was assessed that more mature organizations would gain the most benefits [INT_E], while it could be adopted more easily in small organizations [INT_F]. This opinion was not shared by [INT_I] who asserts that small companies may have difficulties to implement the framework.

7.3.8 Suggestions

Reference questions: IW8.1, IW8.2 and IW8.3

Aim: To elicit explicitly any suggestions for improvement of the framework and to identify the major deficiencies.

Almost all improvement opportunities and deficiencies were already mentioned in the previous sections as part of the statements given by the interviewees. Additionally it should be considered to represent the radar chart in the "Holistic View" differently, i.e. showing also the individual success indicators for each measurement level instead of combining them [INT_C]. Also, the description of the framework needs to be enriched with elaborated examples and suggestions of appropriate metrics in order to ease its adoption for
practitioners [INT_A, INT_E and INT_G]. A topic which is not discussed in the framework is how data can actually be collected and how human aspects [INT_I and INT_H] need to be considered in measurement and in (unobtrusive) data collection [INT_J]. A discussion on the quality of the collected data and how it can be improved is also missing [INT_H].

The most commonly mentioned benefits provided by the framework are its structured approach [INT_A, INT_H and INT_I], the concept to determine the appropriate time for evaluation [INT_A, INT_B and INT_I], the idea of the "Holistic View" [INT_B, INT_F and INT_G], and the concept of the evaluation area [INT_B, INT_C, INT_F and INT_G].

### 7.4 Industry validation results

The industry validation is conducted by means of an interview and a self-administered questionnaire which content is presented in detail in Appendix A. For the self-administered questionnaire, the questions are accompanied by material about the framework such that the respondent can compile the questionnaire in an informed way. The companies in which the interviewees and respondents are employed are located in Sweden, the U.S., Malaysia and Singapore.

The companies’ core businesses are Telecommunications, Electronic and Electrical Manufacturing, and Global Communication Solutions. The interviewees' and respondents’ industry experience ranges from five to eighteen years.

<table>
<thead>
<tr>
<th>No</th>
<th>Interviewee/Respondent</th>
<th>Industry experience (years)</th>
<th>Company size (*)</th>
<th>Business unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Interviewee J ** [INT_J]</td>
<td>18</td>
<td>Large</td>
<td>Multimedia</td>
</tr>
<tr>
<td>2</td>
<td>Interviewee K ** [INT_K]</td>
<td>12</td>
<td>Large</td>
<td>Product Development Excellence</td>
</tr>
<tr>
<td>3</td>
<td>Respondent L [INT_L]</td>
<td>5</td>
<td>Small</td>
<td>Engineering</td>
</tr>
<tr>
<td>4</td>
<td>Respondent M [INT_M]</td>
<td>5</td>
<td>Large</td>
<td>Research and Development Engineering</td>
</tr>
<tr>
<td>5</td>
<td>Respondent N [INT_N]</td>
<td>10</td>
<td>Large</td>
<td>Data Networks Department</td>
</tr>
<tr>
<td>6</td>
<td>Respondent O [INT_O]</td>
<td>12</td>
<td>Large</td>
<td>Enterprise Mobility Solutions</td>
</tr>
<tr>
<td>7</td>
<td>Respondent P [INT_P]</td>
<td>10</td>
<td>Large</td>
<td>Research and Development</td>
</tr>
</tbody>
</table>

Note (*) : The company size is according to the European Recommendation 2003/361/EC [113]

(**) : Interviewee J and K had participated in a face-to-face interview and others responded through the questionnaire

The results from both interview and self-administered questionnaire are combined and discussed in the following sections. Since the questions asked in the interview and questionnaire were similar (except IW7.1, IW7.2, QT7.1 and QT7.2), only the questions asked in the questionnaire are referred.
7.4.1 Evaluation area (Measurement levels and viewpoints)

Reference questions: QT1.1

Aim: To investigate if the proposed evaluation area is helpful to characterize the sources (measurement levels) and destinations (viewpoints) of information in improvement evaluation.

In general the measurement levels were judged as appropriate and reasonable to represent the measurable entities in a software organization and the evaluation area provides the division of responsibility and a clear picture of the overall organizational implementation [INT_P]. However it may be a challenge to measure the External level accurately [INT_O]. An area of improvement would be to include the sponsor viewpoints also for the Process and Project measurement levels in order to fortify the commitment from management and to ensure that the measurement at all levels is given the right focus [INT_M]. The sponsor roles at these levels should be occupied by section or project leads [INT_M]. Indeed, Process and Project level should have the major focus for the evaluation [INT_J]; the development team can be Sponsor or Controller since it needs the authority to control the improvement initiative at these levels [INT_J]. If the implementer / development team has the authority to evaluate the initiative, the motivation and momentum for the process change may improve since the feedback loop for change proposal, approval and implementation is shortened [INT_K].

It is important that the development team is trained and comfortable with the development process; therefore it is crucial for sponsors and process managers to assess adherence and compliance to the agreed process in order to ensure its effectiveness [INT_N]. It was also proposed to represent the evaluation area more hierarchically, e.g. a higher hierarchy level would be interested as a Sponsor, middle as Coordinator, and lower level as Implementer and each measurement level should have the respective interested roles as Sponsor, Coordinator and Implementer [INT_L].

7.4.2 Primary and complementary measures

Reference questions: QT2.1 and QT2.2

Aim: To validate the concept of primary and complementary measures in the context of improvement evaluation. Additionally, it is assessed if the metaphor of the "Iron Triangle" is useful to create a mindset which supports the elicitation of the appropriate success indicators.

Both types of measures are needed to provide a complete view of the improvement initiative [INT_O] and ensure that one can make a wise decision on whether the implementation was successful or not [INT_P]. Sponsors and Coordinators are often focused only on primary measures [INT_O]; nevertheless the overhead of additional measures, like time and impact to schedule, need to be considered to assess if the increased effort for collecting and evaluating the data is worthwhile [INT_O]. Ultimately, it should bring benefit to the team, product and organization [INT_N]. It may be difficult to elicit complementary measures at each measurement level since the categories (success indicators) are not formally defined and are often elicited by gut feeling [INT_J]. The definition of primary and complementary measures is similar to a cost-benefit analysis and is important to consider in every measurement [INT_L]. The categorization in primary and “secondary” measures represents the needed division between “needed” and “good to have” measures [INT_M].
The concept of the iron triangle as a metaphor to visualize and to create a mindset for primary and complementary measures was broadly accepted, although the practitioner has to adapt it to the organizations' needs [INT_K]. The improvement goal and therefore the necessary measures probably vary according to the measurement level [INT_K]. Aligning the complementary measurements with the iron triangle, which is based on project management, clarifies that the taken measurements assess the impact of the implementation on the overall project level [INT_P]. The relation between quality, schedule and cost is well expressed in the “golden” triangle, as one cannot (or with difficulty), improve one without complementing the others [INT_L]. On the other hand, it is dependent how “quality” is defined since addressing customer needs in a timely manner may be more important for certain application domains [INT_N]. It is important to make process change benefits as visible as possible, otherwise the implementers' productivity may be negatively affected and ultimately influence the quality of their work if they feel overloaded with unnecessary process activities [INT_O]. The elements of the iron triangle may also be cost, schedule and scope since in most organizations quality is a given element [INT_M]. The scope of improvement is the variable that can be changed while the quality to which the scope yields must be acceptable based on exit criteria [INT_M].

7.4.3 Confounding factors

Reference questions: QT3.1

Aim: To determine if the issue of confounding factors is recognized in the industrial practice of improvement evaluation.

It was commonly agreed that controlling confounding factors is vital for an accurate evaluation. However, even though the concept and description of confounding factors sounds theoretically correct, it seems impossible to control them in practice [INT_L]. Creating the awareness of their existence is a necessary step [INT_N] and understanding them properly will help to create appropriate baselines for improvement evaluation [INT_P]. The number of variables needs to be limited in order to isolate the visible changes of the process under evaluation [INT_O].

7.4.4 Derivation of measures using GQM

Reference questions: QT4.1 and QT4.2

Aim: To validate the proposed approach to derive the needed measures for improvement evaluation. Since the approach incorporates the previously discussed concepts and uses GQM as a supporting tool it is crucial to validate to which extent it is useful in practice.

Although the proposed approach provides a methodological way of eliciting measures for improvement evaluation [INT_M], it was objected that there is no consistency in the expertise level on different projects [INT_L]. E.g. the difference in a statistic from two projects does not reliably represent the rate of effectiveness / efficiency / performance since human factors will distort the indicated results [INT_L]. On the other hand, the framework was judged as helpful for the change initiator to obtain the required feedback and to define the measurement goals accordingly [INT_N]. Nevertheless, the practitioner needs enough background experience to use the method appropriately [INT_N]. Since it is difficult to define the correct measures in the beginning of an improvement initiative, measures are often determined by gut feeling and experience [INT_J]. Although the approach was deemed
as systematic and potentially useful in practice, it may imply considerable costs in terms of time and effort [INT_K]. Also, the framework should provide process tailoring as not all organizations need the same level of detail and smaller ones will have constraints in implementing the approach [INT_M]. Furthermore it was objected that the proposed approach assumes an existing process, that is, for an organization lacking a mature development process, the received information would not be meaningful enough to compile any statistics and derive suitable metrics [INT_N]. It was also observed that the proposed derivation of measures is redundant with the previously discussed concepts [INT_P] and, from a practical viewpoint, the procedure should not be followed for every single improvement initiative and a single point of documentation for reference should be sufficient [INT_P].

7.4.5 Time to evaluate

Reference questions: QT5.1, QT5.2 and QT5.3

Aim: To determine if the concepts of Lag Factor (LF) and Degradation Factor (DF) can help to identify the appropriate time for evaluation and if organizations have the capability to define their concrete values.

Generally, the concept of LF and DF was deemed useful and important, although it is unclear or at least difficult to determine practically concrete values for these factors by historical data [INT_J, INT_O, INT_N and INT_M]. If several improvement initiatives are running in parallel and effects are overlapping it may be a challenge to determine the appropriate LF and DF [INT_J]. The people who determine the factors need to be experienced practitioners and experts in the field of the improvement in order to achieve reliable results in the evaluation [INT_M and INT_L]. Furthermore, there should be an ample set of projects from where historical data can be obtained and the nature/discipline of the selected projects should be considered for the choice [INT_N]. LF and DF should be defined case by case, that is, in each improvement initiative historical data can be used as support to define the appropriate intervals but they have to be adapted individually [INT_M].

A regular or periodic evaluation of the improvement is definitely needed to gauge the effectiveness of the initiative and to respond accordingly to the feedback in order to avoid waste of resources on things that bring no value to the organization [INT_N]. Also, a given improvement initiative may become stale over time and may need to be revisited for enhancements or may even be replaced by another initiative after the evaluation [INT_O]. One of the classic dilemmas in an organization is to balance the effectiveness of a process and its efficiency. Therefore, should there be two initiatives running in parallel which are not in synergy with each other, the improvements will be hard to observe or realize without constant monitoring [INT_M]. It was objected that periodic evaluation would be appropriate at the process and project level, but not effective at the organizational or external level since at these levels more confounding factors contribute to the improvement results [INT_L]. Furthermore, it would be beneficial to determine LF and DF within the context of the GQM where the measures are elicited, since the evaluation intervals are dependent on the implemented measurements [INT_P].
7.4.6 Holistic evaluation

Reference questions: QT6.1 and QT6.2

Aim: To validate the model on which the "Holistic evaluation" is based and to identify its potential benefits.

The construction of the holistic view for the evaluation of the outcome of SPI initiative represents a good mix and balance as it takes into consideration various aspects [INT_N]. However, it is hard to say if it works in reality [INT_J]. In fact, the evaluation is not always done in a very structured way rather it is done mostly by so called "gut feeling" [INT_J] in reality. Therefore, this "gut feeling" matter needs to be incorporated [INT_J]. The approach of calculating "Subjective Value of Improvement (SVI)" and "Average of Value of Improvement (AVI)" using various components (gain/loss, subjective weight, impact rating) of the model seems okay [INT_L] on the individual measurement levels. Nevertheless, the calculation of SVI and AVI in each measurement level and combining them into one picture to create holistic view is not appropriate as lower levels are more like a subset to higher levels, rather than equally weight them as “one picture” [INT_L]. The holistic view represents a subjective evaluation of the effectiveness of a process [INT_M and INT_O]. It eliminates noisy data by subjective evaluation but this could lead to not representing the true picture of the situation [INT_N]. The evaluation result can be made more objective by quantifying the improvement benefits in a monetary perspective (i.e. how many man hours are saved as the effectiveness of a code inspection is increased) [INT_M].

In general, it was agreed that the holistic view is beneficial [INT_L, INT_M, INT_N, INT_O and INT_P] as it allows an organization to quickly view the benefits of a given SPI initiative into various measurement levels. Hence it enables an organization to relate the improvement to various angles rather than looking into just one good or bad result of a specific measurable entity [INT_M and INT_O]. Making a final conclusion based on consolidating results from many viewpoints ensures that the various aspects of the measurement and its impact have been considered [INT_P]. The construction of the holistic view using a radar chart is also a good method to present the health of the initiative [INT_N]. However, some training is required to properly understand and interpret the chart [INT_N].

7.4.7 Applicability

Reference questions: IW7.1, IW7.2, QT7.1, QT7.2 and QT7.3

Aim: To validate the applicability in terms of usefulness, usability, and adaptability of the previously discussed concepts.

It was generally agreed that SPI-MEF is usable in practice by all the respondents [INT_M, INT_N, INT_O and INT_P] except one who neither agreed nor disagreed [INT_L]. It was appreciated that the evaluation process proposed by SPI-MEF is well defined and takes into account the participation of the entire organization and defines measurable and attainable goals to track the effectiveness of the initiative throughout its implementation which is vital for any initiative to succeed [INT_N]. SPI-MEF provides a step by step guidance on how to implement and evaluate a software process improvement initiative [INT_P]. It is anticipated that, if SPI-MEF is implemented correctly, it will support the continuous improvement behavior in the organization by its methodological and quantitative approach [INT_M]. It is also useful for an organization to know the cost and benefit of
implementing certain initiatives \( \text{INT\_O} \). Since some of the practices proposed by SPI-MEF have already been used in practice by SEI Level 5 organizations, confidence on its usefulness in practice is given \( \text{INT\_N} \). However, the practicality of SPI-MEF will vary among organizations with different priorities, culture and size \( \text{INT\_M} \). Therefore, guidelines on how to tailor the framework in accordance with different needs of an organization should be provided to make it more usable \( \text{INT\_M} \). The framework will certainly be useful for an organization which never had a proper process to evaluate an improvement \( \text{INT\_P} \). One challenge involved in implementing it, is to get the buy-in from the organization to spend the time and resources on such an activity \( \text{INT\_O} \).

Most of the respondents considered SPI-MEF as a framework which can be adapted, customized and scaled to different company settings \( \text{INT\_L, INT\_M, INT\_N and INT\_O} \). However, the company culture is the key factor that would determine the acceptance of it at the initial stage \( \text{INT\_M} \). Then of course top management has to be convinced by the driving force of the change, otherwise an implementation of the framework is derailed before it is even started \( \text{INT\_M} \). It is also hard to judge the applicability of the framework in different company settings without really using it since the practical world and theoretical world are different \( \text{INT\_J} \). It is also perceived that it might not be suitable for companies with a small development team or projects that require quick prototyping \( \text{INT\_N} \). One of the respondent considered that the framework is rigid and apprehended that the framework is meant to be used as it is, since it never was discussed how to customize it according to different organizational settings \( \text{INT\_P} \).

7.4.8 Suggestions

Reference questions: QT8.1, QT8.2 and QT8.3

Aim: To elicit explicitly any suggestions for improvement of the framework and to identify the major deficiencies.

It was agreed that the framework as a whole is consistent and usable with the general industry practices \( \text{INT\_M} \) and there is no key deficiency which have been observed in the framework \( \text{INT\_M} \). Nevertheless some deficiencies were mentioned by the respondents. The framework is too theoretical and assumes consistency in the human nature \( \text{INT\_L} \). Customer participation needs to be included in the evaluation process to be more outward looking \( \text{INT\_N} \). It is not certain that Lag and Degradation Factor can be obtained and subjectivity of the results is a concern to ponder upon \( \text{INT\_O} \). The framework is targeting more on middle management level \( \text{INT\_K} \). Needs from lower levels of the organization are not covered properly \( \text{INT\_K} \). The framework failed to discuss on how to customize it according to the organizations maturity and needs \( \text{INT\_P} \).

Some of the benefits of the framework were explicitly mentioned by the respondents. The framework was considered as an excellent and structured way in doing evaluation \( \text{INT\_J} \). The framework is well balanced taking into consideration various aspects, and can provide a very good cost and benefits measurement \( \text{INT\_L} \). The proposed concepts of primary and complementary measures are keys in prioritizing what is important and what is urgent. A good practitioner can thus leverage from this idea by prioritizing the needs of project versus the need for continuous software process improvement \( \text{INT\_M} \). The derivation of measures involving the inputs from the entire organization using the GQM approach provides a systematic way to obtain the appropriate metrics to monitor and
evaluate process improvement \([INT\_N\ and\ INT\_O]\). Having such framework can give an organization a proper guidance on how to evaluate an SPI initiative \([INT\_P]\).

Most of the suggestions given by the respondents to further improve the framework were already mentioned in the previous sections along with the discussion of their feedback on different concepts. Additionally it was suggested to provide some examples in order to improve the applicability of the framework in the real world \([INT\_J]\). It was agreed that the evaluation of improvement is independent from the actual initiative; however, it needs to be tailored with the SPI initiative \([INT\_J]\). Guidelines on how to customize the framework and provisions of tailoring the guidelines and methodology according to the organizations maturity and needs are very important in determining the critical success factors for different organizations \([INT\_M]\). The establishment of a governance body in ensuring that all SPI initiatives are implemented in synergy is of utmost importance. Recommendations of how to do so would improve the totality of the SPI-MEF framework \([INT\_M]\).

7.5 Validation analysis and refinements

The results of academic validation and industry validation are compared and combined for analysis. Based on the analysis, necessary refinements were done on the final framework. However, refinements were done only for feedback which was found appropriate and which was expected to meet the intended purpose of the framework. The final version of the framework after all the refinements were incorporated is described in Chapter 5 and the summary of the first version of the framework is given in Appendix A. The following sections present the combined analysis of both academic and industry validation along with the refinements that are made in SPI-MEF.

7.5.1 Evaluation area (Measurement levels and viewpoints)

The concept of the evaluation area was validated by both academic researcher and industry practitioners. Both groups brought forward some issues, where obviously academic people mostly commented on theoretical aspects and industry peoples' comments were mostly focused on the applicability of the concept in practice. The measurement levels were adjudged appropriate by both groups. Nevertheless, there were concerns regarding the proposed ordering (Process \(\rightarrow\) Project \(\rightarrow\) Product \(\rightarrow\) Organization \(\rightarrow\) External) of the measurement levels. Actually, the ordering could be different if it is observed from the perspective of evaluation rather than measurement. In measurement, measures are collected from different entities while in the evaluation an assessment and analysis of the collected measures is conducted. For example, the effect of a change in the process would be visible first in process and therefore, measures can be collected from process immediately. However, the evaluation of this process change can take longer as this process change can span over several projects. In the authors' opinion it is important to highlight this differentiation, since it represents one of the arguments for the defined ordering of the measurement levels.

There were also a few suggestions to change the term "measurement levels" to other names like "measurement areas", "measurement aspects", or "measurement artifacts". After a discussion and investigating the feasibility of adopting these names, it was found that the proposed names actually fail to carry the intended meaning. There is some sort of ordering in the measurable entities as discussed in the Section 8.16 in [28] but the proposed names
cannot be related to explain this ordering. On the other hand, the word "level" in the term "measurement level" implies this ordering.

Some confusion arose regarding the roles in each viewpoint and in the interpretation of the categorization of the viewpoints. For instance, it was not clear that the same role can be part of different viewpoints (e.g. a project manager who has the viewpoint of a Coordinator in the Project level can also have the Implementer viewpoint in the Product level). This may be due to the fact that the short document provided to the interviewees/respondents as an introduction to the framework did not suffice to clarify all the considerations in this aspect. To reduce these potential confusions and the chances of misinterpretation, the discussion on each category was enhanced with examples in the final framework. Multiple instances of the evaluation area were illustrated with concrete examples of viewpoints; one along with the concept (Table 5), another with the running scenario (scenario in Section 5.3.1), and one in the elaborated scenario in Section 6.2.1.1.

It was also suggested to investigate whether certain roles can be incorporated in separate/new viewpoints since they are not covered by the suggested categorization. However, the other suggested roles were found to be irrelevant since the aim of the viewpoints is to define who is interested to see the result, not who is affected by the improvement.

7.5.2 Primary and complementary measures

The most outstanding finding in the discussion of this concept is the indication that the distinction between primary and complementary measures is probably well-known, although subconsciously, by both academics and industry people. Since the terms "primary" and "complementary" measurements were introduced by the framework, it was necessary to explain during the interview what is exactly meant by this categorization and, as a consequence, it was deemed as important to state their definition very precisely in the elaboration of the framework. As it was observed by one academic interviewee, the term "complementary" may induce misunderstandings, and indeed, an industry interviewee interpreted the measures as the "needed" (primary) and "good to have" (secondary) ones. Clearly, this is not the intended interpretation and several remedies were discussed to avoid this misinterpretation. As a result, a renaming of the terms was discarded, since any naming inherits ambiguities depending on the background of the reader. Therefore, in order to minimize the space for interpretation, the definition of the terms "primary" and "complementary" should be enhanced and the examples for measurement derivation should be elaborated with more detailed steps which illustrate the concept in more depth. Additionally, it should be made very explicit in the framework that "complementary" measures are not optional ("good to have"), but necessary for a complete evaluation. Therefore, in order to reduce the misunderstanding that could be induced by the term "primary" and "complementary", it was specifically highlighted in the final framework (Section 5.2.2.1) that complementary measures are not optional ("good to have") but as important as primary measures for a complete evaluation.

Although the principle of the "Iron triangle" was judged as appropriate by both academics and industry representatives, it may not provide enough support for practical use. As it was pointed out in the discussion, the "Iron triangle" can be used as a metaphor; however more concrete support for practitioners needs to be included in the framework. Therefore, to ease the application of the iron triangle metaphor that was introduced along
with the concept of "primary and complementary measures", a pool of metrics is provided in the final framework. Measures are grouped according to measurement level so that practitioners can select the needed success indicators and the corresponding metrics (see Appendix B). It is assumed that this enhancement of the framework will better illustrate the idea of the concept and support the user more concretely. However, it should be noted that the pool of metric has to be seen as reference and it should not be regarded as complete. Organization need to populate it along with the accomplishment of the evaluations.

7.5.3 Confounding factors

This concept was specifically asked to the industry interviewees in order to exhibit if they consider it as an important issue in the practical evaluation of SPI. Compared to the other questions, the input to this concept was rather thin, although positive. Indeed, it was deemed as a necessary step to creating awareness for confounding factors and considering confounding factors appropriately in the construction of baselines and practical ways to control them in an industrial setting are needed. In the final framework, a short description of typical confounding factors that need to be taken into consideration for evaluation planning or during the evaluation is included along with some tips to address them (Section 5.2.3.3).

7.5.4 Derivation of measures using GQM

The feedback to the questions regarding the proposed method was dichotomous: on one hand, the approach was judged as systematic and comprehensive, which indicates that, in principle, the method can be of practical use and provide appropriate support for practitioners. On the other hand, some interviewees experienced the approach as quite complex and time-consuming which implies caveats in its applicability in terms of training and education of employees, and in justifying the additional resources needed for its implementation. The concerns about the complexity of the method can be addressed by considering that the process of derivation of measures is an iterative one and is indeed scalable to more realistic settings than those which were shown in the example given in the interview material. Two important advantages of the approach compared to an ad-hoc definition of metrics which were not addressed in the interviews should be noted when discussing its efficiency. First, due to the nature of the GQM paradigm, traceability of the improvement goals to the elicited measures is given. This helps to justify the effort to collect and analyze the chosen metrics. Second, the documentation produced by the elicitation process represents a valuable asset for the organization which can be reused in future evaluations, i.e. the required effort should be accounted in a long-term perspective.

As noted briefly in the previous section, a pool of measures should be added to the framework. In combination with a guideline and concrete steps how to use SPI-MEFs’ method to derive measures, it is assumed that the usability and practicability is increased. Adding to the framework, as it was proposed by one interviewee, a palette of goals, questions and metrics on which the user can base his measurement program on, was regarded by the authors as inflexible and difficult to maintain. It would be more appropriate to define a step-by-step guide which leads the practitioner to formulate his own goals and questions and then provide a pool from which he can pick the needed metrics. Clearly, this implies more effort on the part of the user of the framework; however, this approach makes it flexible and applicable in a wider range of scenarios. Therefore, a step-by-step guideline is presented in the final framework (see Section 5.2.2) to increase the usability of the proposed method of measure derivation. Besides, some templates like the GQM Construction Table
Template, the Measurement Goal, Question(s) and Metrics Template were introduced to support the steps of derivation of measures.

7.5.5 Time to evaluate

Both the academic researchers and industry practitioners agreed that the concept of Lag Factor (LF) and Degradation Factor (DF), and periodic evaluation is conceptually right. However, there were concerns regarding its applicability in practice in terms of coming up with the right values for LF and DF. The main concerns were how to come up with these values in the first place when the initiative is new or when several improvement initiatives are running in parallel. DF was considered harder to define as compared to LF. However, DF is the key concept that helps to define the time-bounds for periodic evaluation and could also help to determine the optimum interval between successive evaluations which is important to minimize the cost of evaluation. Periodic evaluation is important to gauge the effectiveness of the initiative. Therefore, it is important to consider LF, DF and periodic evaluation all together without neglecting any of these factors. It was suggested to provide some guidelines on how to come up with the values of LF and DF. However, the framework needs to be applied in practice and empirical evidences are required before such guidelines can be provided. Otherwise such guidelines could be misleading. Therefore, at the beginning when the framework is introduced in an organization, experienced practitioners and experts in the field of process improvement could help to come up with these values. Thereafter, organizations can learn and improve their accuracy to determine LF and DF when they gain more and more empirical evidence for appropriate values of LF and DF.

There were some disagreements with the heuristic provided for LF and DF that their values increase with the measurement levels (“LF_Prc < LF_Ppj < LF_Prd < LF_Org < LF_Ext” and “DF_Prc < DF_Ppj < DF_Prd < DF_Org < DF_Ext”). For instance, sometimes LF_Prc can be larger than LF_Ppj or LF_Prd since process can span over several projects or products and thus the evaluation of the impact of improvement on process can take longer than project or product. Therefore, the heuristic provided for LF and DF is not always true and adding such a heuristic in the framework description could easily confuse/mislead the user of the framework. Therefore, the heuristic provided for LF and DF was discarded in the final framework. Note that the ordering of LF and DF was only discarded, not the ordering of the measurement levels as discussed in Section 7.5.1.

7.5.6 Holistic evaluation

The discussion with both academia and industry revealed some important characteristics of the "Holistic View" representation of the improvement and which strengths and weaknesses are inherent in this approach. It was confirmed that the target audience for the holistic representation resides in top-level management for which the reduction in details can be seen as an advantage. The tool is therefore less adequate as decision-support for the continuation or further refinement of an improvement initiative (this has to be done at a lower level where details are conserved), but rather expresses the "health" on the initiative and highlights if the expected benefits are not achieved. The subjective element ("gut feeling"), as it is integrated in the model, was judged both positively and negatively. Subjective ratings in improvement assessment are used in industry and therefore applicable in the "Holistic View". The contribution of the framework would therefore be the formalization of that process, and it can be enhanced by describing how the rating can be
homogenized between the different stakeholders which are involved in the evaluation. To make the subjective rating in the improvement assessment more homogenized and consistent among the different stakeholders who are involved in the evaluation, in the final framework it was emphasized to agree upon on some guidelines on how to perform this rating (see Section 5.3.3.1).

The discussion about the model and its components to construct the "Holistic View" was more detailed in the academic interviews. It revealed weaknesses in the model, e.g. the redundancy of the "Gain/Loss" component which just introduces additional complexity without providing more information and provided the insight that a single diagram, aggregating all viewpoints, may not be desirable since the information how the improvement is experienced by the different stakeholders is lost in that case. Therefore, in the final framework, in the “Holistic evaluation” concept, the “Gain/Loss” component was discarded (refer to SPI-MEF summary in Appendix A). In addition to the combined holistic view of all measurement levels, another holistic view presentation was added for all the viewpoints in each measurement level (see Section 5.3.3.3) so that the evaluation done by each individual viewpoint can also be presented. Furthermore, the various opinions about the used measurement scales in the model motivated the authors to conduct further research on this issue.

7.5.7 Evaluation opportunity matrix

Only academic researchers were asked for validating this concept since it was found too theoretical to ask to industry practitioners. Academic researchers were mainly asked to judge the appropriateness of the factors considered in two different axes (accuracy and coverage) of the evaluation opportunity matrix and whether they can see any factors which are left out. The cost of the evaluation was considered as a very important factor to be incorporated. The absence of cost consideration may lead organizations to opt for a good enough evaluation and devoid them from expending money to gain high accuracy and coverage to achieve a holistic evaluation. Therefore, the cost factor should be included in this matrix and a discussion on the relation between quality of evaluation and cost should be included in the concept. Therefore, additionally to the two dimensions (accuracy and coverage) used in the "Evaluation opportunity matrix", a discussion on the cost factor was added in the final framework (see Section 5.1.2). It was suggested to include the improvement goal in the matrix too. However, it is not included as it is considered that there exists no relation between the quality of the evaluation and the improvement goal, that is, the quality of the evaluation does not (and should not) vary with the change of the improvement goal.

Higher accuracy and better coverage is of course good to achieve. However, it may not be feasible for companies to achieve both simultaneously in the first place since resources may be constrained. Therefore, it is important to know which one is important to consider first. There was dichotomy in the answers of the interviewees on this issue. Some suggested considering accuracy first while others considered coverage as more important. However, their answers revealed that giving emphasis on accuracy first has some formidable advantages. Achieving accurate and valid results first can increase the confidence on the quality of evaluation which then can motivate to increase the coverage adding more complexity in the evaluation and investing more resources. If the intention of the evaluation is to see a more complete picture of the improvement benefits first and identifying the problem areas, then coverage should get more emphasis than accuracy. As a conclusion,
selecting first accuracy or coverage depends on the intention of the evaluation and preference of the company itself.

7.5.8 Applicability

Although the concepts of SPI-MEF were illustrated to the interviewees at a rather high abstraction level (simple examples were given where appropriate), it was agreed both from academic and industry side that the framework is applicable and can provide real benefits to an organization which implements it correctly. It can be assumed that several concepts were recognized from the interviewees and deemed as useful since they were formalized by analyzing the actual industry practices (done in the work presented in [28]). Since the framework does not make any assumptions on the to be evaluated improvement initiative, it was judged as flexible on one hand but on the other, it was also considered as too generic and additional effort is needed to adapt and customize it to the specific company settings. Defining guidelines how to implement concretely the individual steps may improve the framework to some extent, although they can't substitute completely the expertise and experience provided by the user. In the final framework, each step of the framework is illustrated and exemplified by a running scenario to illustrate the applicability of the framework in the real world.

7.5.9 Suggestions

Both academic researchers and industry practitioners highlighted some deficiencies and benefits of the framework and gave some general suggestions on the overall framework aside from the suggestions for each specific concept. One of the most common suggestions was to elaborate the framework descriptions with examples to improve the understandability of the framework. Therefore, the final framework was enhanced by adding more examples where it was deemed to be important. Some tips for the execution of the framework, which are regarded useful by the authors, were also added. Besides, a complete scenario of applying the framework was also included as a supplement to the framework to enhance the understandability and clarify the steps of applying the framework (see Chapter 6).

The missing discussion on data collection mechanisms and consideration of human aspects was considered as one of the deficiency of the framework. The consideration of these aspects is not the core focus of this framework as it is assumed that those issues are addressed in the implementation of the measurement program. Experiences and guidelines for successful measurement program implementation can be found in [105].

The framework was deemed to be too theoretical to industry practitioners, which was quite obvious/actually expected as the framework was not applied yet in real practice and not refined according to industry needs.
7.6 Summary

In essence, the approach proposed by SPI-MEF for SPI evaluation was taken very positively by both academic researchers and industry practitioners. Both groups agreed in general that SPI-MEF has the potential to provide a systematic way of evaluating SPI-MEF. However, there were several suggestions or issues brought forward to improve the framework in terms of increasing its applicability in practice. Some of these issues and suggestions were considered for the refinement of SPI-MEF, some others were considered inappropriate for refinement as they don’t meet the intended purpose of the framework and few others were not addressed as those were not in the scope of this thesis but are recommended for future work.

Table 44 summarizes the issues/suggestions that were considered for refinement from the validation results.

Table 44: Issues/suggestions considered for refinements

<table>
<thead>
<tr>
<th>No</th>
<th>Topics/concepts for validation</th>
<th>Issues/suggestions</th>
<th>Refinements</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Evaluation area (measurement levels and viewpoints)</td>
<td>The viewpoint categories and the evaluation area concept need to be enhanced by more concrete examples to reduce the chances of misinterpretations and confusions.</td>
<td>The explanation of each category of viewpoints and evaluation area was enhanced with examples in the final framework. Multiple instances of the evaluation area were illustrated with concrete examples of viewpoints (in Table 6, in the running scenario Section 5.2.1, and in the elaborated scenario in Section 6.2.1).</td>
</tr>
<tr>
<td>2</td>
<td>Primary and complementary measures</td>
<td>Mentioning explicitly the purpose of “Complementary measure” in order to reduce the space of interpretation.</td>
<td>The purpose of “Complementary measure” was specifically mentioned in the final framework (Section 5.2.2.1) highlighting that complementary measures are not optional (“good to have”) but as important as primary measures for a complete evaluation.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incorporating pool of commonly used metrics to ease the application of the “Iron Triangle” metaphor proposed in the framework.</td>
<td>A pool of metrics grouped according to measurement levels and different success indicators was included in the final framework (see Appendix B) to ease the application of the “Iron Triangle” metaphor.</td>
</tr>
<tr>
<td>3</td>
<td>Derivation of measures using GQM</td>
<td>Step-by-step guidelines to increase the usability of the proposed method for deriving measures using GQM.</td>
<td>To increase the usability of the proposed method for deriving measures using GQM, a step-by-step guideline is presented in the final framework (see Section 5.2.2.2). Besides, some templates like the GQM Construction Table template,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td><strong>4  Confounding factors</strong></td>
<td>Creating awareness and understanding of confounding factors.</td>
<td>A short description of typical confounding factors that need to be taken into consideration during the evaluation is included along with some tips to address them (Section 5.2.3.3).</td>
<td></td>
</tr>
<tr>
<td><strong>5  Time to evaluate</strong></td>
<td>Discarding the heuristic provided for Lag Factor and Degradation Factor that their values increase with the measurement levels.</td>
<td>It was revealed that the heuristic provided for LF and DF (“LF_{pre} &lt; LF_{pj} &lt; LF_{ptd} &lt; LF_{org} &lt; LF_{ext}” and “DF_{pre} &lt; DF_{pj} &lt; DF_{ptd} &lt; DF_{org} &lt; DF_{ext}”) are not always true and adding such a heuristic in the framework description could easily mislead the user of the framework. Therefore, the heuristic provided for LF and DF was discarded in the final framework.</td>
<td></td>
</tr>
<tr>
<td><strong>6  Holistic evaluation</strong></td>
<td>Discarding the &quot;Gain/Loss&quot; component from the model of holistic view creation (see Appendix A: C6).</td>
<td>From the model for holistic view creation, the &quot;Gain/Loss&quot; component was discarded in the final framework (Section 5.3.3.1) as it just introduces additional complexity without providing additional information.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>In addition to combined holistic view of all measurement levels, adding another holistic view presentation for all the viewpoints in each measurement level.</td>
<td>A new holistic view presentation was added for all the viewpoints in each measurement level (see Section 5.3.3.3) so that the improvement experienced by each viewpoint can also be preserved and presented.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Emphasizing to agree upon on certain guidelines for the “impact rating” component of the holistic view creation model which is subjective in nature.</td>
<td>In the final framework, it was emphasized to agree upon on some guidelines on how to perform “impact rating” (see Section 5.3.3.1) so that the subjective ratings of the different stakeholders are more homogenized and consistent. In addition, a concrete example of such guidelines is illustrated in the detailed scenario and added as supplement to the framework (see Table 36).</td>
<td></td>
</tr>
<tr>
<td><strong>7  Evaluation opportunity matrix</strong></td>
<td>Incorporating cost factor in the evaluation opportunity matrix and adding discussion on cost</td>
<td>A discussion on the cost factor was added in relation to the other two dimensions (accuracy and coverage)</td>
<td></td>
</tr>
</tbody>
</table>
involvement for increasing quality of the evaluation. in the "Evaluation opportunity matrix" in the final framework (see Section 5.1.2).

8 Applicability

Defining guidelines on how to implement concretely the individual steps of the framework.

In the final framework, each step of the framework is illustrated and exemplified by a running scenario to illustrate the applicability of the framework in the real world. The running scenario is added separately at the end of each section in the final framework enclosed within a textbox (see Chapter 5).

9 Suggestions

Elaborating the framework descriptions with more examples to improve the understandability of the framework.

The final framework was enhanced by adding more examples where necessary. Some tips, which are regarded useful by the authors, were also added. Besides, a complete scenario of applying the framework was also included as a supplement to framework to enhance the understandability and clarify the steps of applying the framework (see Chapter 6).

Table 45 summarizes the issues/suggestions that were not considered for refinement and the rational for not considering them in refinements.

Table 45: Issues/suggestions not considered for refinements and the rationales

<table>
<thead>
<tr>
<th>No</th>
<th>Topics/concepts for validation</th>
<th>Issues/suggestions</th>
<th>Rationales</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Evaluation area (measurement levels and viewpoints)</td>
<td>Changing the proposed ordering of measurement levels (Process → Project → Product → Organization → External).</td>
<td>In the authors’ opinion the proposed ordering is appropriate as it was intended to highlight the entities from where measurements are collected and not the evaluation of these entities. See Section 7.5.1 for more detailed discussion.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Changing the term &quot;measurement levels&quot; to any other terms like &quot;measurement areas&quot;, &quot;measurement aspects&quot;, or &quot;measurement artifacts&quot;.</td>
<td>There is some sort of ordering in the measurable entities as discussed in the Section 8.1.6 in [28] but the proposed names cannot be related to explain this ordering. On the other hand, the word &quot;level&quot; in the term &quot;measurement level&quot; implies an ordering.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Incorporating certain roles as separate/new viewpoints since</td>
<td>The other suggested roles (e.g. people involved in supporting the process)</td>
</tr>
</tbody>
</table>
they are not covered by the suggested viewpoint categorization.

were found to be irrelevant since the aim of the viewpoints is to define who is interested to see the result, not who is affected by the improvement (see Section 7.3.1 and 7.5.1 for more details).

Several remedies were discussed for this issue. Any naming inherits ambiguities depending on the background of the reader. Therefore, instead of renaming this term, the purpose of “complementary measures” was described (Section 5.2.2.1) explicitly in the final framework in order to minimize the space for interpretation.

Besides the issues and suggestions that were discussed in Table 44 and Table 45, there were few others issues/suggestions that couldn’t be implemented either for not being in the scope or for time limitation of the thesis. These issues/suggestions are described in Table 46.

Table 46: Issues/suggestions not addressed in the SPI-MEF

<table>
<thead>
<tr>
<th>No</th>
<th>Issues/suggestions</th>
<th>Description/ rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Primary and complementary measures</strong>&lt;br&gt;Discussion on data collection and consideration of human aspects.</td>
<td>A detailed description of establishing a measurement program including the data collection mechanisms and human aspects was considered as missing in the framework. Giving a complete description of establishing measurement program was not in the scope of the framework. To give the user of the framework an idea on how to use the framework in combination with a measurement framework, the interface between SPI-MEF and a well-known measurement framework was illustrated (Figure 17). Establishment of the measurement program can then be realized by using the guidelines of well-known measurement frameworks. However, in future, a brief discussion on collection of data and the consideration of human aspects can be added in the framework.</td>
</tr>
<tr>
<td>2</td>
<td>Guidelines for determining the values of Lag Factor (LF) and Degradation Factor (DF).</td>
<td>It was suggested to provide some guidelines on how to come up with the values of LF and DF. However, providing such guidelines without empirical evidence is not appropriate. In future, the framework can be applied in industry and such guidelines can be produced.</td>
</tr>
<tr>
<td>3</td>
<td>Integrating monetary perspective in the framework.</td>
<td>It was suggested to quantify the improvement benefits from a monetary perspective (i.e. how many man hours are saved as the effectiveness of a code inspection is increased). It was revealed that coming up with monetary values is not always possible. However, this issue can be further</td>
</tr>
</tbody>
</table>
investigated in future work which will help to conduct cost-benefit analysis of improvement initiatives more concretely.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>4</strong></td>
<td>Changing the measurement scale of the &quot;Impact Rating&quot; component of the holistic view creation model.</td>
</tr>
<tr>
<td></td>
<td>The &quot;Impact Rating&quot; component in the calculation of &quot;Subjective Value of Improvement&quot; in the holistic view creation model has ordinal scale which does not support addition or multiplication operations. In future work the &quot;Impact Rating&quot; component can be further investigated so that other measurement scales can be used to support different mathematical operations. Translating improvement outcome into monetary terms, as mentioned in previous point, could be one of the possibilities for further investigation in this regard.</td>
</tr>
<tr>
<td><strong>5</strong></td>
<td>Adding cost-benefit description of applying the framework.</td>
</tr>
<tr>
<td></td>
<td>It was deemed to be useful to know the cost and benefit of implementing such a framework. Though, a discussion on cost impact is discussed in the final framework (Section 5.1.2), a more elaborated discussion is needed to illustrate the actual cost and benefit of applying the complete framework.</td>
</tr>
<tr>
<td><strong>6</strong></td>
<td>Guidelines for tailoring the frameworks in different organizational settings.</td>
</tr>
<tr>
<td></td>
<td>To increase the customizability of the framework, guidelines on how to tailor the framework in different organizational settings can be added. This was considered as out of scope of this thesis work and recommended as one of the other potential future work.</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td>Consideration of company culture.</td>
</tr>
<tr>
<td></td>
<td>Consideration of company culture is vital to ensure the acceptance of the framework within the company. Therefore, in future work, the differences in company culture can be investigated to better understand the major cultural issues that need to be considered for the effective implementation of this framework.</td>
</tr>
</tbody>
</table>
8 VALIDITY THREATS

In this section, threats to the validity of this research are discussed. A major threat to static validation, as it was conducted in this thesis, is the uncertainty whether the to-be-validated idea or the way how the idea is presented is judged by the experts. This issue arises from the inherent nature of static validation in which the validation is dependent on both the ability of the researcher to build logical and sound arguments and on the ability of the interviewee to follow these presented arguments. Hence, it is difficult to assure that the presented idea is the subject of validation and not the way how the arguments are formulated and apprehended by the interviewee. Therefore, for a complete validation of the framework, a dynamic validation (see in the beginning of Chapter 7) would be required which would show how the framework would perform in real settings.

According to Wohlin et al. [117] threats to validity are categorized into four categories, namely conclusion, internal, construct, and external validity. The discussion about threats to validity of this research follows this categorization.

8.1 Internal validity

Threats to internal validity are concerned with the observed relationship between the treatment and the outcome, i.e. the external factors that can influence an independent variable with respect to the causal relationship with a dependent variable [117]. Three threats to internal validity were identified which are related to the gathering of expert opinion: instrumentation, maturation, and selection.

The instrumentation threat is caused by bad design of artifacts [117] that are used in the expert opinion elicitation. This can lead to misunderstandings regarding the discussed topic and weaken the results from gathered data. To minimize this threat, the preparatory document and the interview questions were piloted first with three Software Engineering students to test whether the artifacts are clear and understandable. Afterwards, a refinement of the preparatory document and the interview questions were made.

The maturation threat exists if the experts’ behavior changes during the elicitation process as the time passes [117]. This can distort the gathered results if the subjects acquire new knowledge during the process or become detached [117]. This threat is regarded as minor since the interviews with the academic experts were conducted during a meeting which lasted approximately one hour each. The written questionnaire was compiled and returned by all industry experts within two weeks; since no deadline to return the questionnaire was given to the subjects, the rather quick response indicates that they were committed to the task and had interest in providing useful information. Furthermore, the questionnaire was designed to present the needed information and the questions concisely and precisely such that it can be compiled within approximately one hour.

The selection threat is concerned with the varying human performance and potential biases introduced by the selected subjects for the investigation, e.g. higher motivation of volunteers may lead to better results [117]. As the presented profiles in Section 7.3 and 7.4 show, both academia researchers and industry practitioners have several years of experience in their respective fields. Obviously there are differences in expertise in the specific areas of interest but this was regarded rather as an advantage than a threat since a major goal of the
validation was to identify new, not yet considered, aspects for the measurement and evaluation of SPI.

8.2 External validity

Threats to external validity are factors that can influence the ability to generalize of the results to a wider scope than covered by the study [117]. One threat of external validity identified in this research is interaction of selection and treatment. This threat is caused by not having a representative sample of the population [117]. To address this threat, the selection of academia experts has also taken into consideration the industrial experience of the subjects. The industry experts selected in this research are employed in different companies with different core businesses from Europe, the United States and Asia.

8.3 Construct validity

Construct validity is concerned with the relationship between theory and the observed outcomes of the research, that is, with the ability to generalize its results to the theoretical construct which motivate the research [117]. In this research, three threats for this category were identified: mono-operation bias, evaluation apprehension, and experimenter expectancy.

Mono-operation bias is caused by considering only a single subject, independent variable or case and hence, the study may not fully represent the investigated theory. This threat is considered moderate since two groups with different background were considered in this study and for each group (academic and industry experts) more than one subject was involved.

The threat of evaluation apprehension is caused by the human tendency to look better while being evaluated [117]. This can distort the result of the study since the subjects perform better than in a regular, unobserved, situation. To tackle this issue, the experts were guaranteed their anonymity and that their answers and opinions will only be used by the researchers.

Another threat identified in this study is experimenter expectancy that is caused by the desire of the experimenter to achieve results as they were expected [117]. The interviewer can try to steer the experts to give answers similar to what the interviewer expected. This threat was addressed by formulating the interview questions and the questionnaire in a neutral way.

8.4 Conclusion validity

Threats to conclusion validity are concerned with factors that affect the ability to draw the correct conclusions from the conducted study [117]. Three threats were identified in this study that fall under this category: random heterogeneity of subjects, random irrelevancies in experimental setting, and searching for a certain results.

Random heterogeneity of subjects is a threat caused by a heterogeneous sample such that individual differences within the sample could affect the study’s result [117]. To minimize this threat, the experts were selected based on their competencies and knowledge in software engineering and software process improvement.
Random irrelevancies in experimental setting are elements outside of the study setting which can disturb its conduction [117]. This threat is considered as minor since the interviews were conducted in an uninterrupted session and in a quiet environment. There were no discussions about the questions before the interview that could have influenced the interviewee’s answers.

Searching for results or “fishing” is the tendency of the researchers to search for a certain result or answer and ignore the inconvenient information [117]. To minimize this threat, all answers from the experts, whether they were positive or negative, were recorded and analyzed regardless the researchers’ expected outcome.
9 CONCLUSION

The aim of this thesis project is to integrate the measurement and evaluation model proposed by [28] in a usable and useful framework which provides SPI practitioners a systematic methodology to evaluate software process improvement initiatives. This is achieved by:

- Analyzing the model and its concepts in order to understand how the proposed ideas can be implemented in a structured framework.
- Formulating an intermediate version of the framework that implements the concepts and addresses the issues of improvement measurement and evaluation raised in [28].
- Statically validating this preliminary framework by gathering and analyzing opinions from academic and industry experts.

The contribution of this thesis is given by the elaboration and static validation of the framework. The following section provides an overview of how the initially stipulated research questions are answered in the course of this thesis while Section 9.2 proposes possible future work.

9.1 Research questions revisited

RQ1: How can the evaluation model in [28] be implemented in a measurement and evaluation framework of SPI initiatives?

A conceptual analysis of the model proposed in [28] is conducted and implemented in a structured framework. Table 47 recapitulates the issues for improvement measurement and evaluation raised in [28] and shows how and if they are addressed by SPI-MEF.

Table 47: Issues and implemented solutions

<table>
<thead>
<tr>
<th>No</th>
<th>Issue</th>
<th>Description (adapted from [28])</th>
<th>Addressed by</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>Heterogeneity of SPI initiatives</td>
<td>The spectrum of SPI initiatives ranges from the application of tools for improving the development, the change or introduction of new processes, to the implementation of organization-wide efforts to increase the development capability as a whole.</td>
<td>SPI-MEF is designed to make no implicit assumptions on the to-be-evaluated SPI initiative. This high-level nature of the framework provides flexibility but imposes also tailoring effort to practitioners.</td>
</tr>
<tr>
<td>II</td>
<td>Partial evaluation</td>
<td>The outcome of SPI initiatives is predominately assessed by evaluating measures which are collected at the project level. This implies that the improvement can be evaluated only partially, neglecting effects which are...</td>
<td>Both the “Gap analysis of evaluation quality” step (Section 5.1.2) and the evaluation area concept, particularly the “Evaluation scoping” step (Section 5.2.1), support the practitioner to define to which extent the initiative can and should be...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>Limited visibility</td>
<td>This issue is a consequence of the previous one since by the limitation of the measurement to a certain level; the gathered information is targeted for a specific audience which may not cover all important stakeholders of an SPI initiative.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>“Evaluation scoping” (Section 5.2.1), “Conduct evaluation” (Section 5.3.2) and “Holistic view creation” (Section 5.3.3), which implements one of the approaches presented in [28], address this issue by prescribing the explicit elicitation of the evaluation audience and by proposing how to present individual evaluation results, and aggregated results which show the performance of the improvement as perceived by the involved stakeholders.</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>Evaluation approach</td>
<td>Due to the vast diversity of SPI initiatives, it is not surprising that the approaches for their evaluation vary analogously. The evaluation and analysis techniques are customized to the specific settings where the initiatives are embedded and, since there exist not even rudimentary guidelines for their implementation, it is assumed that the design and development of the evaluation methods require a considerable amount of effort.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>This issue is partially addressed by SPI-MEF. Currently no suggestions are given how to select the appropriate evaluation method. The categorization given in [28] provides numerous practical examples of specific instances of the four general evaluation methods: basic comparison, statistical analysis, survey, and cost-benefit analysis. Further research is needed to identify reliably strengths and weaknesses of each of these methods in order to provide strong suggestions.</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>Metric definition inconsistencies</td>
<td>Kaner et. al. [118] emphasized how important it is to define exactly the semantics of a metric and the pitfalls that arise if no common agreement exists on what the metric actually means, i.e. which attribute she actually represents.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>This is a fundamental problem of software measurement and goes beyond the scope of improvement evaluation. SPI-MEF addresses this to some extent by prescribing in “Determination of measures” (Section 5.2.2) to put the elicted metrics in context (see Metric Form template). This approach supports the common understanding of the metrics between the stakeholders which are involved in the evaluation.</td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>Unobserved effects</td>
<td>This issue originates from the observation that the measurement of quality is the main focus of improvement initiatives. This led to the idea that by focusing on the measurement of a single attribute, others may change unobservably and a later improvement</td>
<td></td>
</tr>
</tbody>
</table>
|   |   | SPI-MEF addresses this issue by showing how the concept of primary and complementary measures (Section 5.2.2.1) can be implemented in the concrete case. By interfacing with the GQM approach (Section 5.2.2.2) a structured and traceable method is given to elicit the required
| VII | Measuring quality | Quality lies in the eye of the beholder and therefore the assessment of improved quality has to consider the different views of quality [1]:  
- User view of quality as capability to fulfill the intended purpose.  
- Manufacturing view of quality as the degree to which the specification is followed.  
- Product view of quality as the underlying characteristics of a product.  
- Value-based view of quality as the amount a customer intends to pay for. |
| VIII | Evaluation validity | This issue is concerned with the validity of improvement evaluations which is threatened by confounding factors. |

set of metrics which are able to assess the impact of SPI initiatives. See also Appendix B which provides a pool of metrics to support the practitioner in choosing them.

This issue is not addressed by a specific part or concept in the framework but the implemented model considers this issue as stated in [28]: The manufacturing and product view of quality are considered by including the process, project and product level as data sources for the assessment of improvement. The user and value-based view of quality is addressed, on a higher level, by proposing two generic evaluation methods, survey and cost-benefit analysis. Note that the user view of quality is actually also considered by the product and external measurement levels, since the metrics which fall into these address customer satisfaction and the effects of the improvement on the entities external to the organization.

SPI-MEF addresses this issue by providing a list of potential confounding factors (Section 5.2.3.3). However, besides the advise to carefully select the to-be-compared entities, no strong suggestion can be given to use one of the other techniques (as proposed in [28]) to compensate confounding factors.

“Evaluation scheduling” (Section 5.2.4) implements concepts which are supposed to increase the evaluation validity.

Both areas, confounding factors and time to evaluate, need nevertheless future research to improve the usefulness of the framework.
**RQ2: To what extent is the model applied in the framework able to assist SPI evaluation in an industrial context?**

In order to answer this research question, the concepts implemented in the framework are presented to academic and industry experts in software process improvement. The framework is therefore statically validated from two different perspectives. First, the academic experts can relate their industry experience with the proposed core concepts and suggest refinements on the conceptual level of the framework. Second, the industry experts can judge and suggest improvements on the practical level of the framework. This strategy allows a more in-depth validation of SPI-MEF since both theoretical and practical aspects of the framework are assessed.

As presented in the validation design (Section 7.2) the questions are grouped according to the to-be-validated concepts of the intermediate framework. Table 48 recapitulates the major findings.

Table 48: Summary of the validation results

<table>
<thead>
<tr>
<th>SPI-MEF concept</th>
<th>Validation result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation area</td>
<td>One important insight for the authors was to make a clear distinction between two different aspects. The first one is “measurement” as the act of collecting values from measurable entities. If this is seen in isolation, the ordering of the measurement levels as proposed by SPI-MEF is reasonable. The second aspect is the evaluation of these collected values, which occurs at a later point in time. Here, the ordering of the measurement levels is not given anymore since, as noted by the interviewees, processes can span over several projects and/or products. The generalization of the viewpoints (Implementer, Coordinator, Sponsor) seems to be appropriate although it has to be made clear that a specific role (e.g. Tester, Project Manager or SPI coordinator) can occupy different viewpoints depending on the measurement level.</td>
</tr>
<tr>
<td>Primary and complementary measures</td>
<td>The concept, although the naming was disputed, was found very useful. Based on the answers from the industry it was clear that the distinction between, or better, the meaning of primary and complementary measures needs to be made very explicit in order to achieve the right interpretation by the practitioner. Regarding the “Iron Triangle” metaphor it was noted that it can support the practitioner to some extent, but more concrete examples are needed and a pool of metrics, grouped by the identified success indicators, can improve the usefulness of SPI-MEF.</td>
</tr>
<tr>
<td>Confounding factors</td>
<td>Although the concept of confounding factors is known to industry, the questionnaire results indicate that it is regarded rather theoretical and difficult to solve in practice. Creating the awareness of a problem is the first step and addressed in SPI-MEF.</td>
</tr>
<tr>
<td>Derivation of measures using GQM</td>
<td>The input for this concept was twofold: the comprehensiveness and structure provided by the approach was perceived as an advantage while its complexity was seen as a major disadvantage. Granted, the</td>
</tr>
</tbody>
</table>
examples given in the interview/questionnaire may seem time-consuming and tedious. On the other hand, the authors think that through the iterative approach to evaluation provided by SPI-MEF the complexity can be reduced and be manageable, even more as soon as tool support is given to execute this task. Bringing metrics into context, as prescribed by SPI-MEF, can be a valuable asset for organizations in a long-term perspective.

| Time to evaluate | Here, the major gap between theory and applicability was observed. Although the concept was regarded as important and correct in theory, in practice it may be difficult to define concrete values for Lag and Degradation Factor. Empirical data and experience are needed to establish reliable values; since the former is currently not given as of the knowledge of the authors, practitioners have to rely on experts for the respective SPI initiative to estimate Lag and Degradation Factor. |
| Holistic evaluation | This concept was perceived very positively although the theoretical aspects, which were more discussed in the academic interviews, needed some refinement. Since SPI-MEF implements only one of the methods mentioned in [28] to construct the holistic view, the feasibility of the other approaches should be further studied. The target audience for this mean of communication was identified as higher-level management since due to the aggregation of data, detailed information is lost. |
| Evaluation opportunity matrix | Cost was considered as a major aspect missing in the evaluation opportunity matrix whose dimensions are accuracy and coverage. Organizations have always to consider the resources they invest in activities and improving evaluation quality is constrained by the additional cost which adhere with the proposed strategies by SPI-MEF. In overall, there was a tendency to address accuracy of the evaluation first since accurate results provide the needed confidence on the evaluation quality with which a later increase of coverage can be justified. |

**RQ3: To what extent is the proposed framework applicable in practice, in terms of its flexibility and reusability in different industrial contexts?**

This research question was answered also within the static validation of the framework. Specific questions regarding the applicability of the framework were given to both academia and industry experts. Admittedly, judging the flexibility and reusability of the framework based on a pure description without being able to “execute” the framework in real settings requires a remarkable amount of imagination and expertise. Therefore, to answer this research question appropriately, a dynamic validation with the concrete application of SPI-MEF in several organizations implementing different SPI initiatives is required and proposed as future work.

However, the interpretation of the answers from the static validation allows the preliminary conclusion that the framework is indeed flexible and provides reuse potential in different industrial contexts. The major advantage of the framework, namely its
independence from a specific SPI initiative, was also seen as a major drawback by the industry experts. Since the framework does not (and cannot due to its generality) consider details of specific initiatives, customization to the context of the organization is needed. Customization guidelines should be provided in a new revision of the framework and should be considered for future work. In the meantime, the framework is accompanied by an elaborated scenario which simulates the required steps and provides initial support for practitioners.

**RQ4: What are the refinements to be incorporated into the proposed framework based on the findings from RQ2 and RQ3?**

The analysis of the input from experts gathered during the static validation provided an interesting list of potential refinements. As shown in Section 7.6, several proposals were not integrated in the final framework and the rationale for these decisions is given there. Furthermore, seven ideas for refinement were not considered, either because they were out of the scope of this thesis work or because the schedule would have not permitted an appropriate integration into the framework. The following refinements were implemented in the final version of SPI-MEF:

i. The viewpoint categories and the evaluation area concept were enhanced by more concrete examples to reduce the chances of misinterpretation.

ii. The purpose of "complementary measure" was clarified by highlighting that it is not optional ("good to have") and indeed as important as primary measures for a complete evaluation.

iii. A pool of commonly used metrics is provided to ease the application of "iron triangle" metaphor.

iv. For the derivation of measures using the GQM approach a step-by-step guideline is presented to increase the usability of this method.

v. A short description of typical confounding factors and some tips on how to address them is provided to enhance the awareness and understanding of confounding factors.

vi. A heuristic proposed for Lag Factor and Degradation Factor, that states that their values increases with the measurement levels (External > Organization > Product > Project > Process), was discarded.

vii. The "Gain/Loss" component from the model of the holistic view creation was discarded as it just introduces additional complexity without providing more information.

viii. An additional holistic view presentation, considering all viewpoints in each measurement level, was added to present the improvement experienced by each individual viewpoint.

ix. A guideline for "impact rating" is included to guide the different stakeholders to accomplish the subject rating in a more consistent way.

x. The cost factor is discussed along with the other components of the "Evaluation opportunity matrix" concept.

xi. Steps to implement the framework are illustrated and exemplified with a running scenario and tips for each step. Besides, a complete scenario of applying the framework is included as a supplement to clarify the steps in applying it.
9.2 Future work

This thesis has produced a framework for the systematic evaluation of SPI initiatives that has been statically validated by expert opinion originating from both academia and industry practitioners. However, the static validation done was only the first step to verify the applicability of the framework. There are several potential future works to follow up the results of this thesis and they include:

i. Dynamic validation of SPI-MEF
Due to the tightly bound schedule of this Master thesis, SPI-MEF was not validated by applying it on actual industry settings. Applying SPI-MEF on live projects is crucial to validate whether the framework is usable in actual practice and adaptable to different organizational environments. Feedback from the dynamic validation will then be gathered to further refine the framework.

ii. Further study on some concepts of SPI-MEF
Some concepts in SPI-MEF are described on a high abstraction level, especially concepts concerning the time to evaluate (Lag and Degradation Factor) and confounding factors. Further studies can be carried out on a methodology/technique to determine the appropriate values for Lag Factor and Degradation Factor in different organizational settings and for different classes of improvement initiatives. For confounding factors, a deeper investigation is needed to explore the causal relationship between the improvement initiative and its effect.

iii. Enhancement of SPI-MEF
SPI-MEF is created based upon the proposed evaluation model in [28] which is based solely on the findings from a systematic review. Aside findings from the proposed evaluation model, SPI-MEF could be enhanced by getting more input/requirements from practitioners in the field of SPI evaluations.

iv. Tool support
As shown in the framework description (Chapter 5) and in the elaborated scenario (Chapter 6), the evaluation approach proposed by SPI-MEF is systematic and follows prescribed steps. It would be therefore reasonable to implement a tool to support the practitioner in managing the produced artifacts and to make the evaluation process more efficient.

v. Investigate not addressed issues in the framework
As shown in Table 46 in Section 7.6, several issues were identified but not addressed in the framework. These issues have to be investigated in the future for possible solutions and integration in the presented framework.

vi. Cost estimation for framework application
The framework does currently not provide any method to estimate the incurred cost of its application and no assessment of reusability of the produced artifacts. Such work would increase the usefulness of the framework and give practitioners an additional tool to adapt the framework to their needs.
REFERENCES


C.G. Mayhall, Hospital Epidemiology and Infection Control (Hospital Epidemiology & Infection Control, Lippincott Williams & Wilkins, 2004.


APPENDIX A

QUESTIONNAIRE FOR VALIDATION BY INDUSTRY EXPERTS

The aim of this questionnaire is to validate a framework developed in the context of a Master Thesis research project. The proposed conceptual framework targets the evaluation of Software Process Improvement (SPI) initiatives and validation by industry experts is needed to assess its usefulness and usability. The questionnaire is divided into four parts:

(A) Industry Expert Background Information
The section contains some short questions to create the profile of the industry expert.

(B) Introduction to the Conceptual Framework
The two-page flyer in this section describes the aims of the framework and gives an overview of the framework as a whole.

(C) Key Concepts Validation Questions
This is the main part of the questionnaire that presents the key concepts of the framework together with the questions that require input from the expert.

(D) General Validation Questions
This is the final part of the questionnaire that presents questions to validate the framework as a whole.

(A) Industry Expert Background Information
The following questions are aimed to create the profile of the expert whose opinions are gathered in our questionnaires. The objective of the profile is to make a summary of the sample population used in our framework validation. Information of the individual industry expert will be kept anonymous.

(A1) Personal Information

Full name: 

Years of industry experience: 

Company: 

Location: 

(A2) Company Information
1. What is the size of your company (i.e. number of employees)? [Fill with X]
   [ ] Small = 1-49  [ ] Medium = 50-249  [ ] Large = 250 and above

2. What is the core business / domain of your company?
3. What business unit / department are you currently working in?

(A3) Software Process Improvement (SPI) Involvement Information

**Definition of terms:**

**SPI initiative:** Any kind of software process improvement activities ranging from framework (e.g. CMMI, Six-Sigma, SPICE or other frameworks), practices (e.g. software inspection, software reuse) or tools (e.g. software development tool, debugging tool, etc).

**Role in SPI initiative:** Involvement in SPI initiative either by participating (as a developer, tester etc.), coordinating (as a process lead, team lead, manager etc.) or other relevant roles.

Have you ever been involved with any SPI initiative? If yes, please answer the following questions. **[Fill with X]**

[ ] Yes  [ ] No

1. How long were you involved in software process improvement?

2. Please state your role / responsibility in the initiative?

3. Please specify what SPI initiative you have been involved in?

4. Do you have any experience in SPI formal assessments (e.g. CBA-IPI, SCAMPI, etc.)? **[Fill with X]**

[ ] Yes  [ ] No

5. Do you have any experience in the evaluation of software process improvement initiatives, in terms of assessing their actual benefits (e.g. Improved product quality, reduced time-to-market, increased ROI, etc)? **[Fill with X]**

[ ] Yes  [ ] No
(B) Introduction to the Conceptual Framework

The two-page flyer in the following pages describes the aims of the framework and gives an overview of the framework as a whole.

**SOFTWARE PROCESS IMPROVEMENT MEASUREMENT AND EVALUATION FRAMEWORK (SPI-MEF)**

**TODAY’S CHALLENGES IN SPI EVALUATIONS**

With the increasing importance of software in industry and as well as in our everyday life, the process of developing software has also gained significant importance. Many companies are investing large sums of money in improving their software processes through different Software Process Improvement (SPI) initiatives.

SPI initiatives are all software engineering methods or activities which are intended to improve the performance of the software process and these can be categorized into frameworks (e.g. CMMI, CMM, SPICE, QIP, Six Sigma, etc.), software engineering practices (e.g. inspections, test-driven development, etc.) or tools that support software engineering practices. After the undertaking of such SPI initiatives, it is of major interest for organizations to evaluate whether the investment for the improvement pays off.

The evaluation of SPI initiatives is a complex undertaking because there are various aspects that need to be taken into consideration.

Firstly, there is no common understanding on what needs to be measured for evaluating different success indicators (e.g. productivity, product quality, etc.) of improvement.

Secondly, the success of improvement initiatives is often primarily based on the evaluation of project outcomes without considering other resulting ramifications. Consequently, the evaluation of improvement initiatives is too narrow and does not provide enough insight to validate their efficiency and effectiveness.

Thirdly, it is important to provide appropriate visibility of the evaluation results and evidence of improvement to the participating stakeholders of the initiative.

**NEED FOR SPI EVALUATIONS**

The evaluation of SPI through formal assessments like CMMI, SPICE or other well-known assessments are concerned about determining the capability level of the organizational process by means of investigating its adherence to certain standards and identifying strengths and weaknesses in the process for suggesting areas for improvement.

The evaluation we are concerned about is different from these assessments. The concepts proposed in SPI-MEF present an evaluation which considers the improvement from different perspectives and, in sum, increases the visibility, and consequently alleviates the assessment of the achieved benefits.

Since the evaluation of the outcome of SPI initiatives is complex but also crucial to the organization, there is a need for a measurement and evaluation framework to evaluate the outcome of SPI initiatives. In the current state-of-the-art, there is no framework specifically designed to carry out the evaluation of SPI initiatives’ outcome. SPI-MEF tries to address this gap.

**SPI-MEF APPROACH**

- Enable organizations to justify the investment on SPI by providing a measurement and evaluation framework for the systematic assessment of SPI initiatives.
- Provide a balanced view of the impact of SPI initiatives for a holistic evaluation by considering the five measurable entities in a software organization (Process, Project, Product, Organization, and Environment).
- Increase the visibility of the results of the improvement for all stakeholders according to their specific needs.
- Reuse of same generic framework in evaluating a broad spectrum of SPI initiatives since the framework is independent of specific SPI initiative.
- Provide an evolutionary path (basic to advanced) for an organization to do evaluation of SPI.
- Create awareness for taking into consideration confounding factors that might affect the accuracy and validity of the evaluation.

**SPI-MEF OVERVIEW**

*Key Concepts* Implemented in *Nurturing* *Design* *Implementation* *Evolution* Realized by *Steps* Supported by *Support Artifacts*
(C) Key Concepts Validation Questions

This section presents the key concepts of the framework. Validation questions are attached in the end of each key concept.

(C1) Measurement Levels and Viewpoints (Evaluation Area)

Aim: To define the scope of evaluation by characterizing areas where data can be collected and by identifying the stakeholders who are interested to see the evaluation of the improvement.

The measurement levels (Process, Project, Product, Organization, and External) represent the measurable entities that can be assessed to evaluate the software process improvement. The following list shows examples of metrics for each measurement level:

i. **Process:** Defect removal efficiency (Defects found in code inspection activity per time unit)

ii. **Project:**Developer productivity (Lines of code per time unit)

iii. **Product:** Quality (Number of customer reported defects per product size unit)

iv. **Organization:** Return on investment (% ROI = (Benefits of improvement / Costs of improvement) x 100)

v. **External:** Externalities of the improvement (Customer business growth)

SPI-MEF defines three generic viewpoints that represent the stakeholders who are interested to see the results of the improvement:

i. The **Implementer** viewpoint represents all the roles that are dedicated to put the software development in general and the process improvement in particular into practice.

ii. The **Coordinator** viewpoint subsumes the roles which generally participate in software development and in a software process improvement initiative as coordination and control entities.

iii. The **Sponsor** viewpoint represents those roles that fund and motivate the improvement initiative and, in parallel, those who are interested in evaluating the improvement according to its costs and benefits.
The scope of the evaluation (evaluation area) is defined by considering both the measurement levels and viewpoints, which serves as a driver for determining the appropriate measures.

**Table 1: Example instance of an evaluation area**

<table>
<thead>
<tr>
<th>Measurement Levels</th>
<th>Viewpoints</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Implementer</td>
</tr>
<tr>
<td>Process</td>
<td>Development team</td>
</tr>
<tr>
<td>Project</td>
<td>Development team</td>
</tr>
<tr>
<td>Product</td>
<td>Project Managers</td>
</tr>
<tr>
<td>Organization</td>
<td>---</td>
</tr>
<tr>
<td>External</td>
<td>---</td>
</tr>
</tbody>
</table>

Q1: To what extent do you agree that the above schema (evaluation area) represents clearly WHO is interested in the evaluation and WHICH measurable entities need to be considered? [Fill with X]

[ ] Strongly agree
[ ] Agree
[ ] Neither agree nor disagree
[ ] Disagree
[ ] Strongly disagree

Please give a short rationale for your opinion.

(C2) **Primary and Complementary Measures**

**Aim:** To create a mindset that supports the elicitation of appropriate measures to evaluate the improvement.

- **Primary measurements** are the set of measurements that are used to assess if the improvement goal has been reached.
- **Complementary measurements** capture the effects of the process improvement that cannot be directly related to the improvement goal. They assess the potential side-effects caused by the improvement initiative.
The “Iron Triangle” metaphor is used to illustrate the relationship between primary and complementary measurements and helps to derive the needed measures. In project management, the iron triangle is used to visualize the challenge to optimize the three major attributes cost, quality and schedule contemporaneously.

“For example, one can try to optimize schedule and cost, but most probably quality will then be negatively affected. The same idea is applied to the example shown in the above table. Defect density is the primary measure since it is directly related to the improvement goal (increase product quality). Effort in man-hours and project cycle time are complementary measures which need to be monitored in order to make the improvement an overall success.”

The example is based on the project measurement level, but the principle can be applied on every measurement level.

Q2: To what extent do you agree that the consideration of primary and complementary measures provides a more complete view on how you look at the effects of the SPI initiative? [Fill with X]

[ ] Strongly agree
[ ] Agree
[ ] Neither agree nor disagree
[ ] Disagree
[ ] Strongly disagree

Please give a short rationale for your opinion.
Q3: To what extent do you agree that the derivation of complementary measures (based on the principle of the iron triangle) clarifies what should be measured and why it should be measured? [Fill with X]

[ ] Strongly agree
[ ] Agree
[ ] Neither agree nor disagree
[ ] Disagree
[ ] Strongly disagree

Please give a short rationale for your opinion.

(C3) Confounding Factors
Aim: To increase the confidence in the validity of the improvement evaluation.

Confounding factors represent a fundamental threat for the evaluation of a process improvement initiative if any kind of comparison is used to assess its effects. A comparison is said to be confounded if the observed difference between two values (the effect of a treatment) is not solely caused by the treatment alone but can be partially attributed to an external factor. The challenge in evaluating SPI initiatives is therefore to identify and possibly control confounding factors in order to achieve valid results. In this initial stage, SPI-MEF's aim is to create awareness of the potential threats caused by confounding factors and, since they are dependent on the context in which the improvement initiative is carried out, exemplify those which are most commonly encountered, e.g.:

- Project nature (new development, enhancement/maintenance)
- Development model (Waterfall, Iterative models, etc.)
- Product size, complexity and domain
- Employee factors (experience level, staff turnover)
- Technology related factors (programming language, tool support, etc.)

An often encountered scenario may be as follows:

"Code inspection is introduced as an improvement initiative in a project. The aim is to decrease the number of defects as early as possible in the development life-cycle and
therefore to decrease the effort in testing and to decrease the number of delivered defects. Assuming that code inspection is used for the first time in the project, one way to assess its effectiveness could be to compare the test effort with another project which was completed without code inspections. It is crucial to select for comparison a project with similar characteristics as the one where the code inspection was introduced: the selected project should be similar in terms of its complexity (e.g. implemented features, work products and size) and in the experience level of the test staff. Furthermore, one has to consider that the introduction of a new practice like code inspections may adhere with a learning curve of the employees who conduct the inspections. Therefore, the effectiveness of the practice should be assessed only with trained employees.”

Q4: To what extent do you agree that confounding factors can affect the validity of the evaluation results? [Fill with X]

[ ] Strongly agree
[ ] Agree
[ ] Neither agree nor disagree
[ ] Disagree
[ ] Strongly disagree

Please give a short rationale for your opinion.

(C4) Derivation of measures for improvement evaluation
Aim: To provide a guideline for the elicitation of the appropriate measures for improvement evaluation.

The previously introduced concepts provide additional information that facilitates the usage of the Goal-Question-Metric (GQM) paradigm. The GQM approach is a systematic way to tailor and integrate an organization’s objectives into measurement goals and refine them into
measurable values. It provides template for defining measurement goals and guidelines on top-down refinement of measurement goals into questions and then into metrics, and a bottom-up analysis and interpretation of the collected data. SPI-MEF provides an interface with the conceptual level of GQM to define the appropriate measurement goals for improvement evaluation.

By using the input from SPI-MEF, it is possible to define very specific measurement goals which facilitate the elicitation of the appropriate questions and right metrics to answer those questions.

The following simple example shows how the concepts proposed by SPI-MEF interface with the GQM approach.

1. **Information provided by SPI-MEF**

<table>
<thead>
<tr>
<th>Improvement Goal</th>
<th>SPI Initiative</th>
<th>Target Entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase product quality</td>
<td>Code inspections</td>
<td>Project and Product</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Measurement Level</th>
<th>Implementer</th>
<th>Coordinator</th>
<th>Sponsor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td>---</td>
<td>Project Manager, SPI Process Coordinator</td>
<td>---</td>
</tr>
<tr>
<td>Project</td>
<td>Inspection Team</td>
<td>Project Manager, SPI Process Coordinator</td>
<td>---</td>
</tr>
<tr>
<td>Product</td>
<td>Project Manager</td>
<td>Product Manager, QA Manager</td>
<td>Head of Division</td>
</tr>
<tr>
<td>Organization</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>External</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>
2. Interfaces between SPI-MEF and GQM

<table>
<thead>
<tr>
<th>Business Goal</th>
<th>Increase product quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object of Study</td>
<td>Code inspection process</td>
</tr>
<tr>
<td>Purpose</td>
<td>Evaluate</td>
</tr>
</tbody>
</table>

Table 2: Example of interfaces between SPI-MEF and GQM

<table>
<thead>
<tr>
<th>Measurement Levels</th>
<th>Focus</th>
<th>Point of View</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Process</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MG1</td>
<td>Effectiveness</td>
<td>SPI Process Coordinator</td>
<td>Project</td>
</tr>
<tr>
<td>MG2*</td>
<td>Efficiency</td>
<td>Project Manager</td>
<td>Project</td>
</tr>
<tr>
<td>MG3*</td>
<td>Process conformance</td>
<td>SPI Process Coordinator</td>
<td>Project</td>
</tr>
<tr>
<td>Project</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MG4</td>
<td>Work product defects</td>
<td>Project Manager</td>
<td>Project</td>
</tr>
<tr>
<td>MG5</td>
<td>Work product defects</td>
<td>Inspection Team</td>
<td>Project</td>
</tr>
<tr>
<td>MG6*</td>
<td>Schedule</td>
<td>Project Manager</td>
<td>Project</td>
</tr>
<tr>
<td>MG7*</td>
<td>Effort</td>
<td>Project Manager</td>
<td>Project</td>
</tr>
<tr>
<td>Product</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MG8</td>
<td>Pre-release defects</td>
<td>Project Manager</td>
<td>Product</td>
</tr>
<tr>
<td>MG9</td>
<td>Post-release defects</td>
<td>Product Manager</td>
<td>Product</td>
</tr>
<tr>
<td>MG10*</td>
<td>Time-to-Market</td>
<td>Product Manager</td>
<td>Product</td>
</tr>
<tr>
<td>MG11*</td>
<td>Cost</td>
<td>Product Manager</td>
<td>Product</td>
</tr>
<tr>
<td>MG12*</td>
<td>Sales revenue</td>
<td>Product Manager</td>
<td>Product</td>
</tr>
<tr>
<td>MG13</td>
<td>Sales revenue</td>
<td>Head of Division</td>
<td>Product</td>
</tr>
<tr>
<td>MG14*</td>
<td>Cost</td>
<td>Head of Division</td>
<td>Product</td>
</tr>
</tbody>
</table>

Note: Measurement Goals (MG) with * are used to elicit complementary measurements.

3. Elaboration of Measurement Goals and Questions

In the following paragraphs a subset of the measurement goals presented in Table 2 are elaborated in more detail.

MG1: Primary Measurement Goal
Evaluate the process of code inspection with respect to its effectiveness on pre-test defect detection in Project from the viewpoint of the SPI Process Coordinator.

Question 1: What is the effectiveness of the used code inspection method?
Metrics:
1. Percentage of defects found by inspection = The number of defects found during inspection, divided by the total of defects found in inspection, unit test, integration test, system test and acceptance test.
Example: 240 defects are found during inspection and the total number of defects found in the project is 370. The inspection effectiveness is therefore 65%.
2. The previous metric does not take into account the severity of the detected defects during inspection. The severity level of a defect indicates the potential business impact for the end user (business impact = effect on the end user * frequency of occurrence). Severity levels are commonly defined as critical, serious, medium and low. A histogram of the number of the defects found by code inspection shows the effectiveness of the inspections process in terms of defect severity.

**MG2: Complementary Measurement Goal**
Evaluate the process of code inspection with respect to its efficiency in Project from the viewpoint of the Project Manager.

**Question 1:** How many defects are found during the inspection activity?

**Metrics:**
1. The **defect finding rate** is the number of defects found during one inspection hour. If 12 defects are found in a meeting which lasts 1.65 hours, the defect finding rate is $12/1.65 = 7.3$. The average defect finding rate can be calculated by taking the arithmetic mean of the defect finding rate of all meetings in a project.

**Question 2:** What is the time invested in inspection meetings?

**Metrics:**
1. The **inspection rate** is the amount of code inspected in one hour of inspection time.
   Example: 450 lines of code are inspected in one meeting that lasts 1.65 hours. Inspection rate = $450/1.65 = 273$ lines of code per hour. The average inspection rate can be calculated by taking the arithmetic mean of the inspection rate of all meetings in a project.
2. The **inspection effort** is the amount of time required to inspect 1000 lines of code. If the inspection rate is 273 lines of code per hour, the inspection effort is $1000/273 = 3.6$ hours. The average inspection effort is 1000 divided by the average inspection rate.
3. The **defect finding effort** is the number of defects found per 1000 lines of code. If the defect finding rate is 7.3 defects per hour and the inspection rate is 273 lines of code per hour, the defect finding effort is $7.3 \times 1000/273 = 26.7$ defects per 1000 lines of code. The average defect finding effort can be calculated by taking the arithmetic mean of the defect finding effort of all meetings in a project.
4. Often inspections require a preceding preparation. The **preparation rate** can be calculated analogously to the inspection rate, using the preparation time of each inspector.

**MG3: Complementary Measurement Goal**
Evaluate the process of code inspection with respect to its conformance to the specified standard process in Project from the viewpoint of the SPI Process Coordinator.

**Question 1:** Were the code inspections conducted according to the schedule (planned vs. conducted)?
Metrics:

1. The **conducted inspection rate** is the number of inspection meetings which were held according to the schedule.
   Example: 8 code inspections were planned for the project, but only 6 were conducted. The conducted inspection rate is therefore $6/8*100 = 75\%$.

**Question 2:** Were all the allocated participants present during the inspections (planned vs. actual participants)?

Metrics:

1. The **participation rate** is the number of inspectors which were actually present in the meetings.
   Example: 3 participants were planned and 2 participants conducted the inspection. The participation rate is therefore $2/3*100 = 66\%$. The average participation rate can be calculated by taking the arithmetic mean of the participation rate of all meetings.

**Q5:** To what extent do you agree that the proposed approach provides a clear and systematic method for the elicitation of measures of improvement evaluation? *[Fill with X]*

[ ] Strongly agree  
[ ] Agree  
[ ] Neither agree nor disagree  
[ ] Disagree  
[ ] Strongly disagree

Please give a short rationale for your opinion.

**Q6:** To what extent do you agree that the proposed approach is useful in practice? *[Fill with X]*

[ ] Strongly agree  
[ ] Agree  
[ ] Neither agree nor disagree  
[ ] Disagree  
[ ] Strongly disagree

Please give a short rationale for your opinion.
(C5) Time to evaluate

**Aim:** To propose an approach that helps to identify the time bounds for improvement evaluation.

The determination of the appropriate time to evaluate the improvement is important; however it is difficult to suggest a general rule which is universally applicable. Therefore SPI-MEF introduces the concepts of **Lag Factor** and **Degradation Factor** which support the practitioner to determine the time bounds for an improvement evaluation. The Lag Factor is the minimum time required for the improvement initiative to propagate a visible impact on the measurable entities (measurement levels). The Degradation Factor, on the other hand, is the maximum time for which an evaluation result can be considered as valid, or, more specifically, can be considered as representing the actual status of the improvement.

The value of Lag Factor and Degradation Factor is defined by organization based on the type of SPI initiative taken and organizational context.

“The illustration above shows the concept of Lag Factor and Degradation Factor where the Lag Factor and Degradation Factor are defined as 2 months and 5 months respectively for a certain SPI initiative at Project level. E1 is the evaluation at first point of time (2 months after the introduction of SPI initiative) when the impact of SPI initiative is visible at Project level. As the degradation factor is 5 months, the evaluation E1 is no longer a valid representation of improvement anymore after the 7th month (five months after E1 is conducted). That is why organization needs to do evaluation periodically to get the current representation of improvement and also to see the trend of impact of SPI initiative. In this illustration, two periodic evaluations are conducted after E1, and they are shown as E2 (4th month) and E3 (6th month).”

The evaluation should be conducted within the interval range defined by the Lag Factor and Degradation Factor. The figure below exemplifies the concept on the Project, Product and Organization measurement levels.
“In the figure above, it is illustrated that the minimum time to propagate the visible impact of improvement (i.e. Lag Factor) is higher at the Organization level ($LF = 10$ months) than Product level ($LF = 6$ months), and it is higher at Product level than Project level ($LF = 3$ months). The same relation applies for Degradation Factor.”

The only heuristic currently provided by SPI-MEF to determine the Lag Factor and Degradation Factor is that their values increase with the measurement levels, that is, “$LFPrc < LFPpj < LF_{Prj} < LF_{Prc} < LF_{Org} < LF_{Ext}$” and “$DF_{Prc} < DF_{Prj} < DF_{Pjd} < DF_{Org} < DF_{Ext}$.” Driven by these two factors, SPI-MEF proposes a periodic improvement evaluation, that is, an evaluation indexed by time to see the trend of the improvement.

**Q7:** Do you think that it is feasible for an organization to determine the value for the Lag Factor and Degradation Factor using their historical data and experience?

[ ] Strongly agree
[ ] Agree
[ ] Neither agree nor disagree
[ ] Disagree
[ ] Strongly disagree
Please give a short rationale for your opinion.

Q9: To what extent do you agree that periodic evaluation (to see the trend over time) is an effective way to gauge the outcome of SPI initiatives? [Fill with X]

[  ] Strongly agree
[  ] Agree
[  ] Neither agree nor disagree
[  ] Disagree
[  ] Strongly disagree

Please give a short rationale for your opinion.

(C6) Holistic View

Aim: To provide a “one-view” representation of the overall improvement evaluation results.

The Holistic View combines the evaluation results of different measurement levels and viewpoints. The construction of Holistic View is an iterative process. The improvement initiative propagates its visible impact to Process level first then to Project in which the process is applied and eventually to the Products emerging from various projects and then to Organization and External measurement level. Therefore, the Holistic View creation starts from the Process level and then evaluation is done periodically in each level to incrementally create the Holistic View.

The following components are used to construct the model:

- **GL (Gain/Loss):** To normalize the different measurements, their actual values are mapped to Gain (+1), Neutral (0) and Loss (-1), relative to previous evaluations.
- **IR (Impact Rating):** Each viewpoint (that is, each stakeholder of the improvement initiative) rates the impact of the improvement initiative on the respective measurement according to the following scale:

<table>
<thead>
<tr>
<th>Impact assessment scale</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low impact</td>
<td>1</td>
</tr>
<tr>
<td>Considerably low impact</td>
<td>2</td>
</tr>
<tr>
<td>Moderate impact</td>
<td>3</td>
</tr>
<tr>
<td>Considerably high impact</td>
<td>4</td>
</tr>
<tr>
<td>High impact</td>
<td>5</td>
</tr>
</tbody>
</table>

- **SW (Subjective Weight):** Each viewpoint gives a weight of subjective importance to every measurement. The weights sum up to 1.
From these components, the **Subjective Value of Improvement (SVI)** for each viewpoint is calculated as follows:

\[
SVI_{vw} = \frac{\sum (GL_i \cdot SW_i \cdot IR_i)}{n}
\]

*vw: Sponsor / Coordinator / Implementer*

*i: Metric 1, Metric 2, \ldots, Metric n*

*n: Total number of metrics*

The table below shows an example on how SVI is calculated for measurements at the process level. The **Average Value of Improvement (AVI)** is the arithmetic mean of SVI.

**Table 3: Example of SVI calculation at the process level**

<table>
<thead>
<tr>
<th>Viewpoint</th>
<th>Metrics</th>
<th>Gain / Loss (^{GL})</th>
<th>Subjective Weight (^{SW})</th>
<th>Impact Rating (^{IR})</th>
<th>Subjective Value of Improvement (^{SVI})</th>
</tr>
</thead>
</table>
| **Sponsor** | M1      | 1                      | 0.4                         | 3                      | \[
\frac{[(1)(0.4)(3) + (-1)(0.6)(4)]}{2} = -0.60
\]   |
|            | M2      | -1                     | 0.6                         | 4                      |                                  |
| **Coordinator** | M2     | -1                     | 0.3                         | 5                      | \[
\frac{[-1)(0.3)(5) + (1)(0.2)(5) + (1)(0.5)(4)]}{3} = 0.50
\]   |
|            | M3      | 1                      | 0.2                         | 5                      |                                  |
|            | M4      | 1                      | 0.5                         | 4                      |                                  |
| **Implementer** | M3   | 1                      | 0.5                         | 4                      | \[
\frac{[(1)(0.5)(4) + (-1)(0.5)(0)]}{2} = 1.00
\]   |
|            | M5      | -1                     | 0.5                         | 0                      |                                  |

**Average Value of Improvement (AVI):** \(-0.50 + 0.20 + 0.75\) / 3 = **0.90**

By calculating the AVI for each evaluated measurement level, the overall improvement can be represented in a radar (Kiviat) diagram as shown in the below figure.
Q10: To what extent do you agree with our approach to construct the holistic view? *Fill with X*

[ ] Strongly agree
[ ] Agree
[ ] Neither agree nor disagree
[ ] Disagree
[ ] Strongly disagree

Please give a short rationale for your opinion.

Q11: To what extent do you agree that consolidating the evaluation results into “one picture” (holistic view) gives benefits to the organization? *Fill with X*

[ ] Strongly agree
[ ] Agree
[ ] Neither agree nor disagree
[ ] Disagree
[ ] Strongly disagree

Please give a short rationale for your opinion.

(D) General Validation Questions

The general validation questions are divided into applicability and suggestions for improvement.

(D1) Applicability

Q1: To what extent do you agree that SPI-MEF is usable in practice? *Fill with X*

[ ] Strongly agree
[ ] Agree
[ ] Neither agree nor disagree
[ ] Disagree
[ ] Strongly disagree

Please give a short rationale for your opinion.
Q2: To what extent do you agree that SPI-MEF is useful (giving benefits) in practice? [Fill with X]

[ ] Strongly agree
[ ] Agree
[ ] Neither agree nor disagree
[ ] Disagree
[ ] Strongly disagree

Please give a short rationale for your opinion.

Q3: To what extent do you agree that SPI-MEF is adaptable and customizable to different company settings (size, core business, culture)? [Fill with X]

[ ] Strongly agree
[ ] Agree
[ ] Neither agree nor disagree
[ ] Disagree
[ ] Strongly disagree

Please give a short rationale for your opinion.

(D2) Suggestions

Q4: Can you name and explain briefly any deficiency you have observed in the framework?


Q5: Can you name and explain the major benefits you have observed in the framework?


Q6: Do you have any suggestions to improve SPI-MEF? If yes, please state them.


APPENDIX B

LIST OF SUCCESS INDICATORS AND METRICS

The following sections (B1 – B5) show the list of success indicators and their corresponding metrics. The list is taken from the output of a systematic review in evaluating SPI initiatives found in [28].

B1: Process measures

<table>
<thead>
<tr>
<th>Success Indicator</th>
<th>Commonly Used Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Efficiency</strong></td>
<td></td>
</tr>
<tr>
<td>i) Rework efficiency = No. of defect / defect fixing effort</td>
<td></td>
</tr>
</tbody>
</table>
| ii) Defect removal efficiency, DRE = E/(E+D)  
  - E is the number of errors found before delivery  
  - D is the number of errors found after delivery |
| iii) Process efficiency (%) = No. of <Phase> category defects detected by <Phase> review / Total no. of <Phase> category defects  
  <Phase> := SRS, Design, Code Review and Unit Test |
| iv) Inspection efficiency = Number of real defects found per hour |
| v) Fault Handling Efficiency (FHE) = Number of incoming fault events / number of events recorded as trouble report |
| vi) Fault Finding Efficiency (FFE) = Number of incoming fault events not recorded as trouble report because reason is an internal fault / number of events recorded as trouble report |
| vii) Test efficiency = No. of bugs found up to and including system testing / No. of bugs found during and after testing |
| **Effectiveness** |                       |
| i) Total Defect Containment Effectiveness (TDCE) = Number of pre-release defects / (number of pre-release defects + number of post-release defects) |
| ii) Phase Containment Effectiveness for Phase i (PCEi) = Number of phase i errors / (number of phase i errors + number of phase i defects) |
| iii) Inspection Effectiveness = Identified real defects / the number of seeded defects (in inspection) |
| iv) Fault Handling Effectiveness (FHEff) = Number of incoming fault events not recorded as trouble report because there was no fault at all + number of incoming fault events not recorded as trouble report because duplication of an issue |
| v) Fault Finding Effectiveness (FFEff) = Number of trouble reports which were not approved for correction |
| vi) Defect detection effectiveness = Individual defect detection rate (IDDR) = Number of real defects reported by individual subjects |
| vii) Fault-slip-through (FST) = Number of defects not found when it was most cost-effective |
## B2: Project measures

<table>
<thead>
<tr>
<th>Success Indicator</th>
<th>Commonly Used Metrics</th>
</tr>
</thead>
</table>
| **Defects**       | i) Defect density, \( DD = \frac{\text{No. defects}}{\text{size}} \)  
|                   | - No. defects according to development phase  
|                   | - Size in LOC, FP, etc  
|                   | ii) Defect distribution (\%) = \frac{\text{No. of <Phase> category defects}}{\text{Total No. of defects}}  
|                   | <Phase> := Analysis, Requirement, Design, Code, Document, Others  
|                   | iii) Yield (% of defects found and fixed before compilation and testing)  
|                   | iv) Fault-slip through (FST) = \frac{\text{Total no. of faults in late verification}}{\text{total no. of fault in the project}}  
|                   | v) Number of defects found per <Phase>  
|                   | <Phase> := Analysis, Requirement, Design, Code Review, Unit Test, Others  
|                   | vi) No. of injected defect / <Phase>  
|                   | <Phase> := Analysis, Requirement, Design, Code Review, Unit Test, Others  
|                   | vii) Defect Injection Rate (DIR) of requirements, design, coding and testing activities = \frac{\text{Defects injected at the activity}}{\text{all defects of the project}}  
| **Cost**          | i) Cost / Size  
|                   | - Size in LOC, FP, etc  
|                   | ii) Cost of rework  
|                   | iii) Rework effort in person-hour, staff-hour, etc  
|                   | iv) Effort * cost per staff-hour  
|                   | v) Person-month, person-hour, staff-hour, etc  
|                   | vi) Rework index = \% project effort spent in rework  
|                   | vii) Avoidable Fault Cost (AFC) = Determined by FST and the average estimated cost of finding a fault in a specific phase of the software life-cycle  
| **Schedule**      | i) Project cycle time (days, months, etc)  
|                   | ii) Cycle time normalized by size  
|                   | iii) Project duration = Time of gating review that decides to release the project to the customer or market - time of gating review which kicks off the project  
|                   | iv) Relative integration time = (Integration time / overall project time) \%, time is calculated in months  
|                   | v) X factor = The amount of calendar time for the baseline project to develop a product divided by the cycle time for the new project  
| **Productivity**  | i) Size / effort  
|                   | - Size in LOC, FP, etc  
|                   | - Effort in person-month, person-hour, hour, etc  
|                   | ii) Effort / size  
|                   | - Effort in person-month, person-hour, hour, etc  
|                   | - Size in LOC, FP, etc  
|                   | iii) Number of tasks solved per time unit  
|                   | iv) Actual productivity = \frac{\text{No. of written code}}{\text{person hours}}  
|                   | v) Apparent productivity = \frac{\text{No. written code} + \text{no. of reused code}}{\text{person hours}}  

Estimation

<table>
<thead>
<tr>
<th>Schedule:</th>
<th>Rate of delivery = No. of projects delivered per year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Productivity = (Number of features in project / total effort in minutes) * (2,400 minutes / week)</td>
</tr>
</tbody>
</table>

| Effort:  | Effort variance = (Actual effort - planned effort) / planned effort |
|         | Project Visibility Index (PVI) = Planned person-months / actual person-months |
|         | Effort Estimation Accuracy (EEA) = Actual project effort / estimated project effort |
|         | 100 * (actual effort - estimated effort) / estimated effort |

| Cost:    | Cost performance index (CPI) = Earned value / actual cost |
|         | Variance in the planned vs. actual cost for projects |

| Size:    | Size variance = (Planned size - actual size) / planned size |
|         | Error rate of size estimate = (LOC estimated – real LOC developed) / real LOC developed. |
|         | 100 * (actual LOC - estimated LOC) / estimated LOC |

| Others:  | Productivity variance = Estimated vs. actual LOC / hour |
|         | Cumulative Earned Value vs. Cumulative Planned Value |
|         | Actual no. of defects removed per phase vs. estimated no. of defects removed per phase |

B3: Product measures

<table>
<thead>
<tr>
<th>Success Indicator</th>
<th>Commonly Used Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General:</strong></td>
<td></td>
</tr>
<tr>
<td>i)</td>
<td>No. of reported failures in operation</td>
</tr>
<tr>
<td>ii)</td>
<td>Post-release fault density = No. of faults / size</td>
</tr>
<tr>
<td>iii)</td>
<td>No. of User-reported deficiencies (URD)</td>
</tr>
<tr>
<td>iv)</td>
<td>New Open Problems (NOP) = total new post-release problems opened during the month</td>
</tr>
<tr>
<td>v)</td>
<td>Total Open Problems (TOP) = total number of post-release problems that remain open at the end of the month</td>
</tr>
<tr>
<td>vi)</td>
<td>Changed Ratio of customer request = (No. of actual changed request / no. of customer request) * 100(%)</td>
</tr>
<tr>
<td>vii)</td>
<td>No. of customer calls</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quality</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliability:</td>
<td>Maintainability:</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>i) Failure Rate (FR) = Number of failures / execution time</td>
<td>i) McCabe's Cyclomatic complexity</td>
</tr>
<tr>
<td>ii) Mean Time Between Failure (MTBF) = (Downtime – uptime) / no. of failures</td>
<td>ii) Maintainability product (MP) = Change scope * effort * 100</td>
</tr>
<tr>
<td>iii) Mean Time To Failure (MTTF)</td>
<td>iii) LOC/subroutine</td>
</tr>
<tr>
<td>iv) Defect density by Configuration Mgmt build (DDb)</td>
<td>iv) Stability index = Percent of the code that has changed in a week, number of developers that made changes in a week, average number of LOC per developer, average time spent by a developer on the project</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## B4: Organization measures

<table>
<thead>
<tr>
<th>Success Indicator</th>
<th>Commonly Used Metric</th>
</tr>
</thead>
</table>
| **Economics**     | i) Cost benefits ratio = Benefit / cost  
|                   | ii) ROI = Profit / Investment = (Throughput – Operation Expense) / Investment; Throughput = Value Output = Sales Price - Direct Costs  
|                   | iii) ROI = Cost of improved defect introduction - cost of action learning initiative  
|                   | iv) $R = \frac{I}{(P \times N)}$, where $R =$ number of months to regain the investment  
|                   | I, I = invested capital, P = price to correct an error, N = reduction in number of errors / month.  
|                   | v) Business value of software-process- improvement investment = value returned on each dollar invested  
|                   | vi) (Benefits - Cost) / Cost * 100%  
|                   | vii) Investment cost on project which has SPI initiative = Percentage of base staffing needed more for process improvement activities * no. of people * months * hours * cost/hour ($/hr)  
| **Employee**      | i) No. working hours per week  
|                   | ii) Employee turnover rate  
|                   | iii) No. of projects the employee works on in parallel  
|                   | iv) No. of expired vacation days  
|                   | v) No. of unscheduled work at weekends  
|                   | vi) No. of sickness days  
|                   | vii) Amount of bonuses  

Note: The usual way of measuring the employee satisfaction is through survey / questionnaires. The above are some metrics which can be used to infer the employee satisfaction.

| **Growth**        | i) Product sales revenue growth  
|                   | ii) Operating profit growth  
|                   | iii) Innovation = Throughput New Product / Throughput * 100%  
|                   | (Throughput = No. of implemented idea)  
|                   | iv) Margin of sales  

| **Communication** | i) Improved communication between customers and software engineers / developers  
|                   | ii) Improved communication between users and software engineers / developers  
|                   | iii) Improved co-operation between quality assurance and the software engineers / developers  
|                   | iv) Improved organization communication  

Note: The above are not really metrics and the usual way of measuring the employee communication is through survey / questionnaires.
## B5: External measures

<table>
<thead>
<tr>
<th>Customer</th>
<th>Commonly Used Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>i)</td>
<td>Operation productivity</td>
</tr>
<tr>
<td>ii)</td>
<td>Product quality</td>
</tr>
<tr>
<td>iii)</td>
<td>Process efficiency</td>
</tr>
<tr>
<td>iv)</td>
<td>Product sales revenue</td>
</tr>
<tr>
<td>v)</td>
<td>Net profit</td>
</tr>
<tr>
<td>vi)</td>
<td>Improved in business operations*</td>
</tr>
<tr>
<td>vii)</td>
<td>Improved in employee satisfaction*</td>
</tr>
</tbody>
</table>

Note: The above are some general success indicators that can be used to measure the impact of SPI on the customer side. In fact, all the other success indicators listed in the Process, Project, Product and Organization can be used depending on how and where the product is being used in the customer’s organization. The ones marked with * are not really metrics and can be measured using survey / questionnaires.

<table>
<thead>
<tr>
<th>Society</th>
<th>Commonly Used Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>i)</td>
<td>Environmental pollutions / hazards</td>
</tr>
<tr>
<td>ii)</td>
<td>Public safety</td>
</tr>
<tr>
<td>iii)</td>
<td>Public health</td>
</tr>
<tr>
<td>iv)</td>
<td>Job opportunities</td>
</tr>
<tr>
<td>v)</td>
<td>Quality of life</td>
</tr>
<tr>
<td>vi)</td>
<td>Other social, economic and political benefits.</td>
</tr>
</tbody>
</table>

Note: The above are some general areas that we can measure the impact of the SPI to the society. The exact metrics are depending on the specific product that is directly / indirectly used by the people.