CHALMERS
UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG

Development and Assessment of Key Performance Indicators to describe the Performance of Software Engineering Support Infrastructure

Master thesis in Software Engineering

Jakob Noetzel

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2018

Master Thesis 2018

Development and Assessment of Key Performance Indicators to describe the Performance of Software Engineering Support Infrastructure

Jakob Noetzel

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2018

Development and Assessment of Key Performance Indicators to describe the Performance of Software Engineering Support Infrastructure
JAKOB NOETZEL

Supervisor: Miroslaw Staron, Department of Computer Science and Engineering
Examiner: Regina Hebig, Department of Computer Science and Engineering

Master Thesis 2018
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Göteborg Sweden Telephone
+ 46 (0)31-772 1000

Development and Assessment of Key Performance Indicators to describe the Performance of Software Engineering Support Infrastructure
JAKOB NOETZEL
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

Background: Over the years, industries have lived through a strong and irreversible shift towards the globalization of software-intensive high-technology businesses. In order to provide software engineers with large amounts of various data when they need it throughout different time zones, high speed performance business application systems are needed. Nowadays, software-heavy organizations rely more and more on quantifying their information about their applications and processes in order to lower maintenance cost and gain competitive advantage.

Goal: The goal of this thesis is to develop, deploy, validate and evaluate a measurement system for monitoring an enterprise application's responsiveness to user operations, addressing the main research question of "*How can, in an industrial context, 'good' Key Performance Indicators (KPI's) be identified and developed?"*. To develop this measurement system, an established responsiveness KPI is used and known software measurement standards are applied in order to ensure its quality and significance for the collaborating company and further research.

Method: The thesis is conducted by making use of the action research method in collaboration with a manufacturer in the automotive industry in order to develop, deploy and assess the measurement system.

Results: Through thoroughly describing and mapping measures based on software measurement standards, information needs for 'good' KPI's could be identified and actions could be derived to develop those. A measurement system was developed based on a known development process and by making use of a known responsiveness KPI and relevant software measurement standards. After a validation and evaluation in an industrial setting using an established quality model, it proved to be applicable to the organization's specific problem and provide valuable information and insight about their performance measurement actions and possible improvements.

Conclusion: According to the stakeholders perception, using the measurement system provides the possibility of improving the performance measurement and software maintenance and therefore decreasing the waste of resources in an industrial setting. By being based on known software standards and derived models and processes it provides the possibility to increase the impact of the KPI's in the organization and optimize data organization. Further, through the theoretical and empirical validation, it increases its acceptance on scientific ground.

Keywords: Software Engineering, Performance Measurement, Software Measurement, Key Performance Indicators, Apdex, Measurement System, Quality Model

Acknowledgements

I would like to thank my supervisor, Miroslaw Staron, who showed great patience and provided valuable feedback and suggestions throughout the whole thesis project. I would also like to thank Robert Maglica, who always offered a helping hand and provided me with information and feedback crucial for the project's success. Finally, I would like to thank the collaborating company at which the thesis project was conducted.

Jakob Noetzel, Gothenburg, September 2018

Content

List of Tables

List of Figures

Abbreviations

Apdex	Application Performance Index
CN	China
FRT	Frustrating Response Time
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
KPI	Key Performance Indicator
MS	Measurement System
Op.	Operation
QME	Quality Measure Element
RT	Response Time
SE	Sweden
SRT	Satisfying Response Time
TRT	Tolerating Response Time

1 Introduction

Over the years industries have lived through a strong and irreversible shift towards the globalization of software-intensive high-technology businesses (Herbsleb, & Moitra, 2001). Software and physical products are developed offshore, developing teams are spread around the world and businesses turn into never sleeping 24 hour operating enterprises.

In order to provide software engineers with large amounts of various data when they need it throughout different time zones, high performance business application systems are needed. Poor response times of such systems have a negative impact on development time, resulting in delays, hence loss of money, and enormous user frustration (Lazar, Jones, & Shneiderman, 2006; Lazar, Jones, Hackley, & Shneiderman, 2005).

Therefore, it is inevitable for software engineers to be able to monitor and maintain software according to their requirements specified in standards like ISO/IEC 25000 and ISO/IEC/IEEE 15939 (Staron, & Meding, 2018). To measure the performance of said systems, theoretical frameworks are used to develop measurement systems and so called key performance indicators (KPI). KPI's play a crucial role in the transformation of raw data into decisions of how to act upon specific outcomes of measurements (Staron, Meding, Niesel, & Abran, 2016; Staron et al., 2018). Theses KPI's are wrapped in so called measurement systems that serve the purpose of collecting, analyzing, and presenting data for a specific information need (Staron, Meding, Karlsson, & Nilsson, 2010). Often though, large organizations collect large amounts of data that are often irrelevant to their actual information needs and end up wasting resources by creating bad measures. That issue creates the need to stop collecting more data and adding more measures, but rather taking actions in order to reduce the data to meaningful information and optimize data analysis (Sevcik, 2005).

In this context, the industrial partner in this thesis project, a big manufacturer in the automotive industry, has set up a performance measurement solution for one of their enterprise applications, denoted *Application A,* and would like to improve their situation by further reducing the large amounts of data they have collected into useful KPI's that provide long-term monitoring and help maintaining their system's high quality requirements.

The thesis work focuses on studying the current performance measurement solution with regard to its capabilities to track *Applications A's* responsiveness to user operations. It will be assessed whether the performance measurement solution provides relevant data that can be condensed into KPI's that satisfy the stakeholder's information need. Based on that assessment and previous research, a KPI measurement system is developed to address the main research question of *How can, in an industrial context, 'good' KPI's be identified and developed?*. The measurement system is developed by making use of an existing, established KPI for application responsiveness, the Apdex (Application Perfromance Index), and putting it into context of known software measurement standards. The measurement system is then validated and evaluated using internal, historical data and a known quality model to ensure its significance and relevance to the company and further research.

The Thesis is structured as follows. The remainings of *chapter 1* present the problem statement, purpose of the study and limitations and delimitations. *Chapter 2* provides the body of knowledge that serves as the theoretical foundation this thesis is building on. *Chapter 3* presents previous research in specific fields that is relevant and used in this thesis to achieve the desired results. *Chapter 4* includes the research questions and describes the chosen research method. *Chapter 5* presents the results of this thesis work. *Chapter 6* provides a discussion in which the research questions are answered with respect to the results of this thesis. Finally, *chapter 7* closes with a conclusion.

Problem Statement and Purpose of the Study

The problem addressed in this thesis is the issue of how to improve the performance measurement of globally distributed enterprise applications, with respect to their responsiveness to user operations, in order to decrease the waste of resources and therefore increase productivity. The collaborating company has set up an automated performance measurement tool in order to track the applications performance. However, the collaborating company needs to organize their large amounts of data and optimize its collection, analysis and management towards their organization's information needs. It is not known, how relevant and suitable the collected data is for developing a useful KPI measurement system that satisfy the company's information needs of achieving long-term monitoring of *Application A's* responsiveness.

In this context the intention of this thesis is to develop a KPI measurement system, based on a solid theoretical framework and the assessment of the current performance measurement solution at the company, that tracks *Application A's* responsiveness and optimizes the company's data organization. For this, an existing and established responsiveness KPI is adopted and known software engineering standards are applied to it. Further, the aim is to validate and evaluate the measurement system to prove its usefulness and therefore lay the foundation for its integration into the organization's business operations and strategy.

Limitations and Delimitations

Since the thesis was conducted with one specific industrial partner, the generalization might be limited due to the company's specific setting.

One limitation is the limitation of time. Since this is a university project in collaboration with an industrial partner, the time the researcher was active within the organization was limited and some aspects of the thesis, for example the empirical validation of the KPI's, would have been more detailed if there was more time available.

Another limitation is that only research from the fields of traditional IT infrastructure was reviewed and applied. A broader literature review, for example including performance measurement of cloud computing systems, might have brought up additional insights. But also, due to time constraints of this thesis, these aspect were excluded.

2 Background

In this chapter, the concepts of performance and software measurement are shortly introduced because they define the domain of this thesis. Further, known and widely adopted software measurement standards are presented as they serve the purpose of building the foundation for the theory being applied in this thesis to develop the measurement system. Moreover, the notion of key performance indicators is introduced to create a link from the software measurement standards to the organizational context. Lastly, the concept of measuring responsiveness is presented since it the measurement system monitors an application's responsiveness to user operations.

Performance Measurement

Organizations that are better and faster informed through optimal measurement and information systems have a competitive advantage in the economic market (Schott, 1981). Schott (1981) claims that the organization will be able to analyse on a broader basis with better accuracy, think more complex, allocate resources more specifically and be reactive faster. Nowadays, software-heavy companies rely on quantitative data like never before in order to stay informed about their software's quality, improve their decision making process and overall productivity (Staron, Meding, Niesel, & Abran, 2016). Therefore, performance measurement has established itself as a crucial part in most companies' business strategy and daily operations. There are several definitions of performance measurement, but in general, it describes the process of collecting, analyzing and reporting information about an element within a company, for example a process, a system or an individual (Upadhaya, Munir, & Blount, 2014). It also defined as the process of quantifying the effectiveness and efficiency of an element (Neely, Gregory, & Platts, 1995).

Companies constantly need to improve their performance measurement as their operations, processes and software continuously evolve. The challenge then is to define the right information needs and understand what "good" KPI's actually are (Staron et al., 2016). Often, many organizations start by collecting large amounts of data about anything that they can possibly measure. Performance measurement then becomes a challenge due to the difficulty of identifying relevant data and reduce it to meaningful information (Sevcik, 2005). Key performance indicators and measurement systems help solving these problems and therefore act as a highly effective instrument for a company's productivity and success (Sevcik, 2005; Staron et al., 2016; Staron et al, 2018). These terms are further described in following subchapters. Moreover, measurement standards, software measurement standards in particular, provide a solid framework for practitioners to successfully describe, develop and implement measures and indicators into their business. These standards are further explained in following chapters and subchapters as well.

The collaborating company has set up performance measurement solutions for one of their enterprise applications and wants to improve their situation by further reducing the large amounts of data they have collected into useful KPI's that provide long-term monitoring and help maintaining their system's high quality requirements regarding performance, responsiveness in particular.

Software Measurement

Software measurement as a process is well established in the fields of software engineering. Though, the metrology of it behind the process is not very well known, which often results in poor measurement programs or systems (Staron et al, 2018). Software-heavy organizations rely more and more on quantifying their performance to support their software maintenance and decision making process (Staron, 2012). Thus, there is a need for frameworks that help defining crucial elements to software measurement and developing good KPI's. In software engineering, standards like the ISO/IEC 25000 series and ISO/IEC/IEEE 15939 provide frameworks for how to properly put software measurement to action by defining standards for the definition of measures, indicators and measurement systems, the process of developing these indicators and how to monitor, analyze, evaluate and report them. In this thesis, a measurement system is develop by making use of a known responsiveness KPI, the Apdex. But these software measurement standards are applied to the development process to ensure proper usage and documentation of definitions, the development process and evaluation. The relationship between these two standards is illustrated in *figure 2.1*.

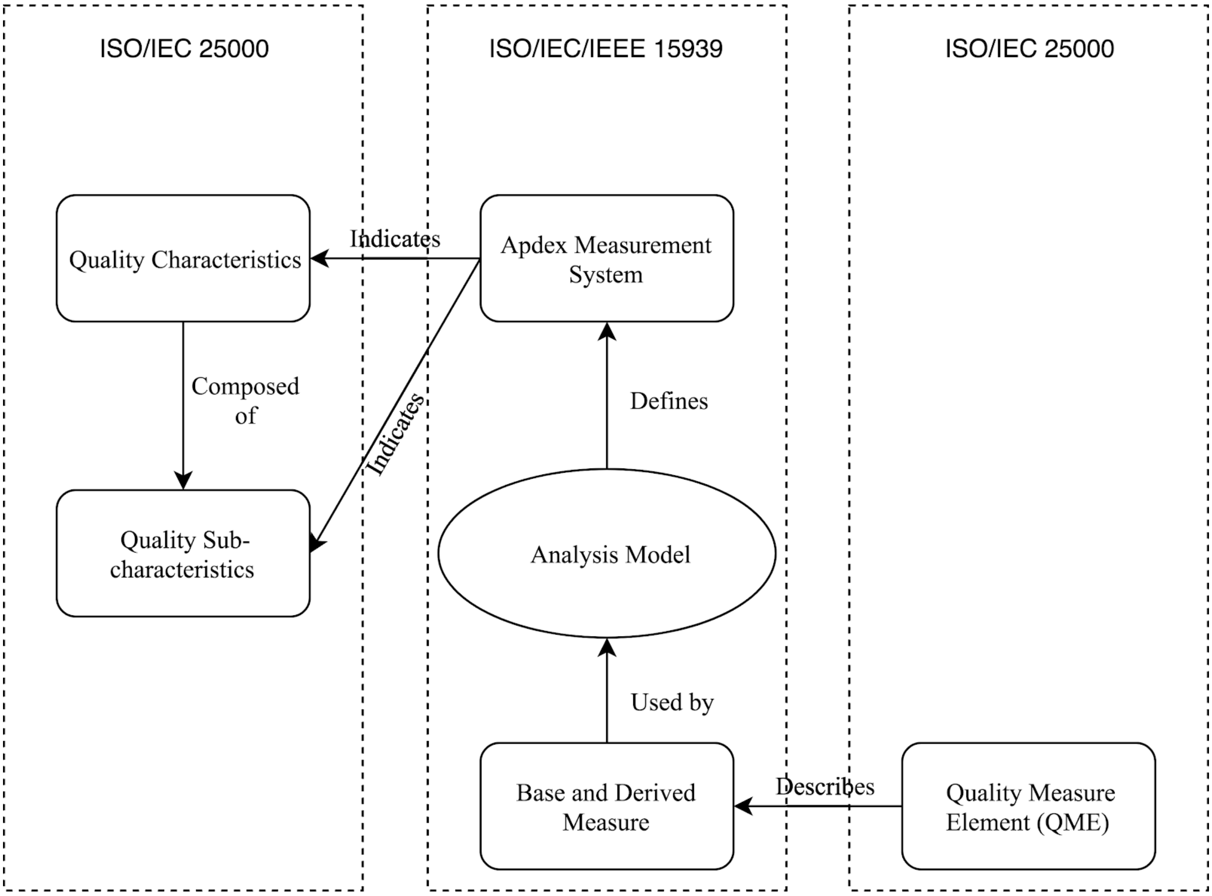

Figure 2.1: Dependencies between ISO/IEC 25000 and ISO/IEC/IEEE 15939

While the ISO/IEC/IEEE 15939 standard defines base measures and derived measures for the development of indicators, the ISO/IEC 25000 standards use these in their definition of measures. Further, the ISO/IEC 25000 standards provide quality attributes, which are used by the ISO/IEC/IEEE 15939 standard to describe the quality of its measures and indicators. On the other hand, ISO/IEC 25000 standards also use the definitions from ISO/IEC/IEEE 15939 to define quality measurement.

The Apdex measurement system in this thesis is developed by applying these standards to the case at hand. Firstly, by describing and mapping existing measures. Then, by making use of a development process based on ISO/IEC/IEEE 15939, as described in *chapter 3,* to develop and validate the measurement system. And lastly, by evaluating the quality of the KPI's of the system based on quality characteristics provided by these standards.

Software Metrology

Before the two standards from *figure 2.1* can be explained, the fundamental standard that is used in almost every engineering field needs to be presented - the vocabulary in metrology (VIM) (International Bureau of Weights & Measures, 1984).

Metrology is understood as the science of measurement (Staron et al., 2018). It describes a number of rules, guidelines and concepts that form the basis for creating numbers from real life circumstances - this process is known as measurement. The metrology provides a number of methods and principles that serve as the basis for the standards used in this thesis. Among others, relevant tools to this thesis described by Staron et al. (2018) that influence the measurement process are:

- Measurand - Describes a quantity that will be measured. It can also be called the measured property of an entity,
- measuring instrument - Describes the instrument that is used to collect relevant measurements
- measuring system - Describes a collection of instruments and/or devices that together provide the measured quantities of an entity.

These principles have significant importance when developing a measurement program or system that should support an organization's performance measurement and decision making (Staron et al., 2012). Since this standard is very generic, more specific standards need to be presented in order to lay a proper foundation for the thesis.

ISO/IEC/IEEE 15939

In the previous subchapter, concepts from the VIM that are relevant for this thesis were outlined. The ISO/IEC/IEEE 15939 is a complementary standard that is derived from the VIM specifically for the software engineering field. It describes a framework that explains a measurement process and how it is to be used in an industrial setting (Staron et al., 2018). Although this standard is not the only one to describe the measurement process, it is used due to its alignment to other standards used in this thesis and the wide adoption within the industry.

The ISO/IEC/IEEE 15939 standard specifies a process that serves the purpose of defining, collecting and analyzing data in software projects or organizations (Staron et al., 2010; ISO, 2017). The main goal of the standard is the definition of the information product, which is defined as a collection of indicators and their derived interpretations that (optimally) satisfy a stakeholder's information need (ISO, 2017). The information need is understood as a desired insight concerning various objectives (i.e. goals, problems or risks) within a company that is necessary for a stakeholder in order to address these objectives. Said objectives consist for example of projects, software or hardware. In addition to aforementioned definitions, there a number of other definitions which the ISO/IEC/IEEE 15939 standard provides that are crucial for the development of "good" KPI's in this thesis. They are:

- Base Measure (BM) - Described as a measure that is defined as an attribute and its method to quantify it.
- Derived Measure (DM) - Described as a measure that results from a function of two or more BM.
- Indicator - Described as a measure that allows interpretation and evaluation based upon the application of an analysis model to base and/or derived measures.
- Decision Criteria - Described as thresholds that indicate whether actions have to be taken based on the indicator's status.
- Attribute - Described as a characteristic of a measured entity. It can be expressed quantitatively or qualitatively.
- Information product - Described as a collection of indicators and their interpretations which describe the stakeholder's information need.
- Measurement method - Described as a Chain of actions taken in order to quantify an attribute based on a specific scale.
- Measurement function - Described as a function that combines BM's to a DM.
- Entity - Measured object that is described by its attributes.

- Measurement process - Described as the process of implementing, planning, conducting and analyzing software measurement within an organization and its projects.
- Measurement instrument - Described as the device or devices used to collect BM's
- Stakeholder - A person that is interested in the indicator and holds the information need. Further, the person that has the ability to act upon the indicator's status.

The most crucial part of a measurement system is the satisfaction of the stakeholder's information need (Staron et al. 2010). Example stakeholders are project managers, applications managers, test engineers, architects or product manager (Kilpi, 2001; Gopal, Mukhopadhyay, & Krishnan, 2005). The indicator presents information in form of a numerical value including its interpretation, which is based on an analysis model (Staron et al. 2010, ISO, 2017). This analysis model consists of defined decision criteria that describe the indicator's status. Usually, and also in this thesis, practitioners in the field of software engineering make us of a traffic light metaphor in order to determine decision criteria for the analysis model (Kuzniarz, & Staron, 2003; Pandazo, Shollo, Staron, & Meding, W., 2010; Staron, Meding, Hansson, Höglund, Niesel, & Bergmann, 2015). BM and DM provide the relevant information for creating the indicator.

These definitions and the process described by the ISO/IEC/IEEE 15939 standard are summarized in the so called measurement system information model which is presented in *figure 2.2*. The figure shows the graphical illustration of the development process, from raw data to the actual indicator that provides interpretation and a status upon which the designated stakeholder can take actions.

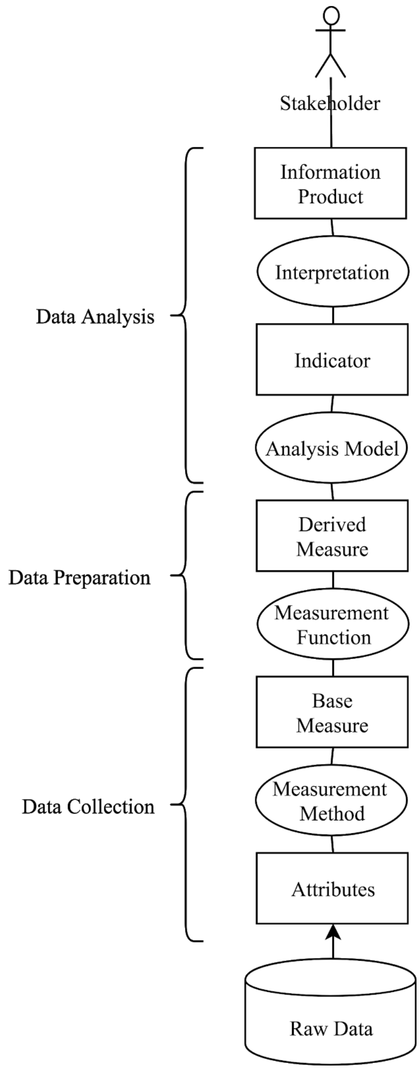

Figure 2.2: Measurement Information Model based on ISO/IEC/IEEE 15939

ISO (2017) provides definitions for each of these elements in the information model, which are categorized into three categories; data collection, data preparation and data analysis. These categories play an important role when it comes to evaluating the process and the information that is being processed, as described by Abran (2010) and Staron et al. (2016). Based on aforementioned definitions and the process presented in *figure 2.2*, a number of criteria is defined that determine the information's quality. These are:

- Certainty of the information product
- The information product is clearly understood by all parties
- There is proof that the information product fits its purpose

- The information product satisfies the organization's assumptions (i.e. business goals / strategy)
- The measurement procedure is correct, documented and accurate
- The measurement method is iterable and can be repeated
- The measurement method can be reproduced

This set of criteria serves as one basis for the quality model that is used in this thesis and presented in *chapter 3.*

ISO/IEC 25000

The ISO/IEC 25000 series is a set of standards that spans its focus from defining quality requirements, over quality models to quality measure elements (QME) (ISO, 2012; Staron et al., 2016; Staron et al., 2018). In this thesis, the focus of the ISO/IEC 25000 series lays on the definition on QME's and the measurement of data quality.

The series was a result of the fast paced development of the IT industry and hence the need of modifying existing quality standards (St-Louis, & Suryn, 2012). It improves the ISO/IEC 9126 standard by showing better alignment with the theory of metrology and its vocabulary (Abran, Al-Qutaish, Desharnais, & Habra, 2005). Furthermore, one of its main goals was the better coordination with the ISO/IEC/IEEE 15939 standard in terms of describing base and derived measures and indicators. Because the ISO/IEC/IEEE 15939 standard mainly focuses on the process of measurement, the ISO/IEC 25000 series is an important complementation when it comes to defining quality measures and evaluation data quality (Staron et al. 2016).

When it comes to describing measures, the ISO/IEC 25000 series provides the definition of quality measure elements (QME's). QME's are described as measures that are defined as properties including their measurement method of quantifying it, with the option of a mathematical function to transform the measure (ISO, 2012). QME's serve the purpose of concisely documenting measures in a format that is align with the theory of metrology and other standards. They provide basic information about the measure, such as its category, name, scale and a more detailed description. To illustrate how measures are described using the definition of QME, the following table uses one of the relevant measures in this thesis, the user operation response time, and describes it based on the ISO/IEC 25000 standard.

Table 2.1: Definition of the response time using the format of QME from ISO/IEC 25000

QME	Description
QME category	Time Duration
QME name	User Operation Response Time
QME ID	RT
Detail	User Operation Response Time measures the elapsed time between an executed user operation and the system's response to it
Input	Performance Measurement Tool
Documentation	RT defines the responsiveness of the application to a user operation
Measurement Scale	Ordinal
Measurement Focus	External

The thesis is using this format to describe the measures at the collaborating company and the relevant measures for the development of the Apdex measurement system. The collection of all measures described using this format can be found in *Appendix B*.

Another relevant part of the ISO/IEC 25000 series is a new set of criteria for evaluation data quality it provides that serves also as basis for the quality model used in this thesis. These criteria are listed below:

- **Accuracy**: to which degree the data for an entity stores are correct
- **Completeness**: to which degree data covers all specified user objectives
- **Consistency**: to which degree the data is described in a concise way
- **Credibility**: to which degree the data is trustworthy and represents what it is supposed to represent
- **Currentness**: to which degree the data is timely and up-to-date
- **Accessibility**: to which degree the data can be used by users with different characteristics to achieve a defined goal
- **Compliance**: to which degree the data conforms to defined standards
- **Confidentiality**: to which degree the data is protected from unauthorized access

- **Efficiency**: to which degree the data processes (i.e. storage, filtering, sharing) can be applied to the data
- **Precision**: to which degree random errors are described in the data
- **Traceability**: to which degree the data can be traced back to its original source
- **Understandability**: to which degree the data is understood by users with different characteristics
- **Availability**: to which degree the data is accessible and operationable then it is required to be used
- **Portability**: to which degree the data can be transferred from one application to another
- **Recoverability**: to which degree the data, after a failure or an error, can be recovered and brought back to its original state

As mentioned before, the ISO/IEC 25000 series complements the ISO/IEC/IEEE 15939 standard in terms of evaluation measures and data quality. Therefore, the list above is seen as an extension to the set of criteria provided by the ISO/IEC/IEEE 15939 (Staron et al., 2016).

Key Performance Indicators

In the previous subchapter, an indicator, based on ISO (2017), is defined as a measure resulting from the application of an analysis model to base measures and/or derived measures. This definition is very generic and standardized, while it focuses mainly on the actual process from raw data to the indicator and its interpretation based on a defined analysis model (Staron et al., 2016). The definition does not take the organizational context into consideration, which crucial for this thesis work. Therefore, the notion of the key performance indicator (KPI) is introduced.

Key performance indicators are defined as measures that focus on those organizational areas that are fundamental for its current and future success (Parmenter, 2015). They are also defined as business measures that can be customized towards an organization’s business operations and show the status and trend of such (Kaplan, & Norton, 1996). Parmenter (2015) describes seven main characteristics for a KPI that are taking the organizational context into account. These are:

- A KPI is a nonfinancial measure (not expressed in currency)
- A KPI is measured regularly (daily, monthly, every 3 months, etc.)
- A KPI is actionable and can be acted upon by the management (i.e. meetings can be called in upon the status of the KPI)
- A KPI is understandable and relevant employees know what is says and how to improve it (stakeholder knows how to fix it when problems occur)
- A KPI is tied to a team in order to create responsibility and accountability (resources are allocated for the management of the KPI)
- A KPI has influence on success factors of an organization (i.e. reduces maintenance cost)
- A KPI has been tested to encourage the right actions and does not lead to false behaviour within the organization (i.e. has been tested after a software update and proves to show the right indication whether it increased or decreased performance)

These characteristics and the KPI's notion in general are rooted in a theoretical framework different from the ISO 15939 standard, namely the balanced scorecard (Kaplan et al., 1996; Parmenter, 2015). The balanced scorecard described by Kaplan et al. (1996) provides a way to measure businesses' performance based on four perspectives, which are financial, customer, internal process and learning and growth.

As recognized in the aforementioned definitions and the KPI's characteristics, compared to the more generic indicator defined by ISO (2017), KPI's are strongly characterized by their connection to an organization's business goals or strategy. Therefore, this thesis focuses on the KPI's in this sense, because the developed measurement system needs to put its values into an organizational context. As for the assessment and development parts in this thesis, the term indicator is used due to the definition by ISO (2017). The resulting values are denoted as KPI's, because they are validated, tested and evaluated in an organizational setting.

Measuring Responsiveness

Since the purpose of the developed measurement system in this thesis is to track an application's responsiveness to user operations, the concept of responsiveness is introduced.

In computer science, responsiveness as a concept is a criteria of the overlaying principle of system robustness and describes a system's ability to execute specific tasks within a given time frame (Weik, 2000). In particular for the context of this thesis, it describes the elapsed time between a user's request to the system and its response. Responsiveness is considered to be one of the most crucial aspects of system robustness and usability since issues can lead to a major decrease in productivity and hence loss of money due to delays and user frustration (Lazar et al., 2006; Lazar et al., 2005; Nielsen, 1994).

To measure responsiveness, there was a need to determine what good responsiveness actually means. Research has been done to provide a framework to answer the question of how fast a system is supposed to respond based on the human's perception (Nielsen, 1994; Ramsay, Barbasi, & Preece, 1998; Bhatti, Bouch, & Kuchinsky, 2000). This research forms the body of knowledge for the KPI Application Performance Index (Apdex) that is used in in this thesis to develop the measurement system for the collaborating company. The research and the Apdex itself are further introduced in the following chapter.

3 Related Work

This section presents the related work in research that has been done prior to this thesis and was applied during the thesis work to achieve the results presented in *chapter 5*. the chapter starts off with the introduction of an existing and established KPI that has been used to develop the measurement system for the collaborating company. Afterwards, the development process and quality model that have been applied in this thesis will be introduced.

The Application Performance Index (Apdex)

The Application Performance Index (Apdex) is the result of a collaboration of companies within different industries with the goal to develop a KPI that successfully reflects an enterprise application's responsiveness to user operations (Sevcik, 2005). It Converts many response time measurements into one number on a uniform scale of 0 to 1 based on two Thresholds (T and F), of which only one has to be defined, and, resulting from these 2 thresholds, 3 performance zones of application responsiveness (satisfying, tolerating, frustrating). Since the application's performance directly influences the user's and therefore company's productivity, it is crucial to have a proper KPI in place that indicates the status of one's applications speed (Apdex Alliance Inc., 2007). But only measuring response times is not sufficient to provide a good KPI, the reporting of the KPI is equally important. Therefore the Apdex Alliance Inc. has developed the Apdex and defined a methodology for measuring and reporting it. It is seen as an open standard with the intention of becoming a widely used KPI and being objective of a continuous, collaborative improvement (Sevcik, 2005).

The Apdex Alliance Inc. (2007) defines a number of objectives for the Apdex, which are listed below:

- Providing a condensed summary of an application's responsiveness
- Providing an easy to understand value
- Providing a value in a universal, fixed scale from 0-1

- Providing a prompt interpretation of an application's responsiveness with 0 being the worst possible responsiveness and 1 being the best possible
- Provide the potential of making several business applications comparable through a universal reporting system

These objectives show great alignment with the information needs expressed by the stakeholders at the collaborating company. This is the main reason why this KPI has been chosen to be used in the development of the measurement system. Another reason is the ability of the Apdex to condense a high number of response times on a user-task level into a single number expressing a specific application's responsiveness. A task is here defined as a single interaction between a user and the system under investigation (Sevcik, 2005). This fits very well into the collaborating company's context because it has set an automated tool that is measuring hundreds of thousands of response times.

The process, defined by the Apdex Alliance Inc (2007), of how to calculate Apdex starts off with defining a report group. Before anything can be measured, boundaries of what exactly is being measured need to be set up. That is why a report group is defined. A report group defines a set of elements that describe the entity under investigation. Four elements need to be defined, which are:

- Type: The type describes the task or task chain that is being measured (i.e. a specific user operation)
- Application: The application that is being measured needs to be defined
- User Group: The user group of that application needs to be defined (i.e. based on geographical location)
- Time Period: The period of time the Apdex is being measured for needs to be defined (i.e. for 3 months, 6 months or daily, weekly, etc.)

The formula for calculating the Apdex is based on aforementioned three performance zones. These are:

- Satisfying: The application is fully responsive. The response time is below the defined threshold T and is not impeding the user's productivity
- Tolerating: The application is lagging performance in terms of responsiveness. The response time is greater than the defined threshold T but lower than the second threshold F
- Frustrating: The application shows unacceptable performance with response times greater than the limiting threshold F

As mentioned before, these two thresholds define the three performance zones which serve as the foundation for the Apdex formula. Defining the thresholds is one of the most crucial parts of the process, because it determines the credibility, believability and usefulness of the resulting Apdex KPI (Sevcik, 2005). Poorly defined thresholds result in bad KPI's that may encourage dysfunctional behaviour within the company.

To define the thresholds T and F, the Apdex Alliance Inc. (2007) has used research from known researchers within the fields of system usability. Nielsen (1994) explains that reasonable fast application operations are supposed to respond between 2 and 10 seconds. Ramsay et al. (1998) has conducted observations concerning loading times for web-pages and found that when a web-page loaded in 10 seconds, users would get much less interested in the content when it took more than 41 seconds to load. Another example are Bhatti's et al. (2000) controlled conducted computer experiments with users. Letting them setting up a new computer, results showed "good" to "bad" ratings in the range of 10 to 39 seconds. And also NetForecast, a company participating in the Apdex Alliance Inc., has done observations of user-system interaction in various business sectors and also found a relationship between the thresholds T and F with ratios of 3:1 and 4:1.

These findings and their alignment when it comes to the relationship between T and F were used as the foundation for the Apdex Alliance Inc. (2007) to define the threshold F as a function of T, namely:

$$F = 4T$$

When it comes to the definition of T, aforementioned research gives an indication of what “good” responsiveness is. Nevertheless, the individual context within the organization needs to be considered as well, i.e. the system’s capabilities and the organization’s goals. The KPI would turn out useless, if T was just strictly defined based purely on a theoretical framework, without taking the organizational context into account. That is why the research was used to analyse the situation at the collaborating company together with the stakeholder. Like this, reasonable thresholds could be defined, based on the theoretical framing and the organizational context. These thresholds were used to calculate the Apdex KPI’s for the user operations under investigation and are presented in *chapter 5* and *Appendix C*.

With two thresholds, the three performance zones are defined which lay the foundation for the Apdex formula presented below:

$$Apdex(T) = \frac{Satisfying + \frac{Tolerating}{2}}{Total\ number\ of\ Response\ Time\ Measurements}$$

The result is a sum of the ratios from the satisfying response times and tolerating response times divided by the total number of response times. The weighting of each performance zone becomes very clear. Satisfying response times are weighted with the value 1, tolerating with 0.5 and frustrating with 0. The formula in this thesis needed to be slightly changed in order to ensure full theoretical validity. More to this change is described in *chapter 5.*

As mentioned before, the reporting of the KPI is equally important to the actual measurement. As also stressed in ISO (2017), an analysis model is used by the Apdex Alliance Inc. (2017) to provide the KPI with the ability of interpretation and evaluation. The following *table 3.1* shows the intervals and their interpretation based on the Apdex Alliance Inc..

Table 3.1: Apdex Reporting Rules based on Apdex Alliance Inc.

Apdex Value Range	Rating / Interpretation	Colour Declaration
0.94 - 1.00	Excellent	Determined by practitioner
0.85 - 0.93	Good	Determined by practitioner
0.70 - 0.84	Fair	Determined by practitioner
0.50 - 0.69	Poor	Determined by practitioner
0.00 - 0.49	Unacceptable	Determined by practitioner

The Apdex Alliance Inc. (2017) is making use of a colour declaration system like the ISO (2017) standard, but leaves it open to the practitioner to decide on a customized colour declaration.

When Evaluating the reporting rules together with stakeholders at the collaborating company, the rules were seen as reasonable and were adopted as the analysis model for the Apdex. The rules were further simplified and a traffic light system was agreed on, which resulted in a refined analysis model presented in *table 3.2*.

Table 3.2: Refined Analysis Model

Apdex Value Range	Rating / Interpretation	Colour Declaration
0.85 - 1.00	Good	
0.70 - 0.84	Fair	
0.00 - 0.69	Poor	

This process shows how the Apdex Alliance Inc. is defining their process for creating the Apdex in an industrial setting. Since this thesis is basing the assessment, development and evaluation of the measurement system on several ISO standards, the Apdex will be adopted but put into the ISO context, which will result in differences of documentation format and reporting formats. For example for the Apdex' specification, while the Apdex Alliance Inc. has no defined format to present the KPI, this thesis applies previously presented software measurement standards to provide a detailed and concise specification for the Apdex measurement system, presented in *table 5.2,* along with a graphical specification illustrated in *figure 5.2* for better understanding.

Measurement System Development Process

Staron et al. (2010) describe a measurement system (MS) development process that is followed in this thesis. The reason to choose this process is because its deep connection to the ISO 15939 standard presented in this thesis. Furthermore, its purpose is align with the needs the stakeholders of the collaborating company expressed, namely the support of their decision making and reusability of their measures (Staron et al., 2010). Also, since ISO standards are widely adopted throughout the automotive industry, the decision to follow this process supports its acceptability.

Based on the definitions and the measurement information model presented in *chapter 2*, Staron et al. (2010) derive a process of 18 steps that is illustrated in *figure 3.1*. The steps range from step 1: Elicit information need from stakeholder to step 18: Deploy Information Product.

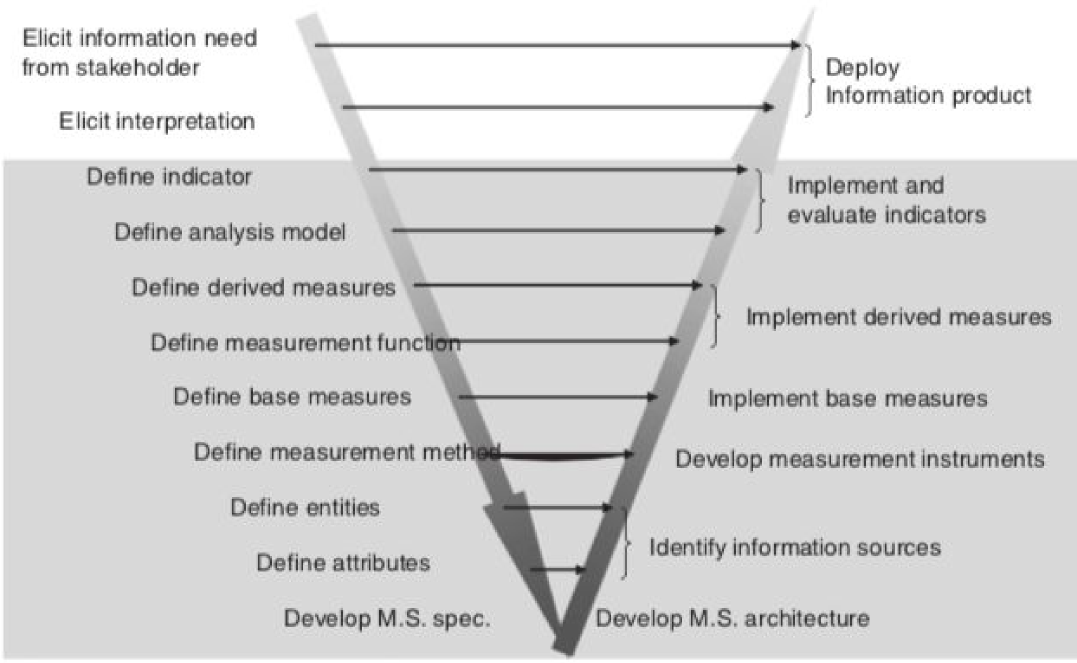

Figure 3.1: Measurement System Development Process as adopted from Staron et al. (2010)

With the shaded area, Staron et al. (2010) indicate that these steps can be conducted iteratively. Generally, these steps are conducted by the developer (steps 1-18) together with the stakeholder at the collaborating company (steps 1-10 and 17-18). For each step, Staron et al. (2010) provide questions the developer should ask himself when finishing this step in the process. These questions of each step result in deliverables and artifacts in the end that the developer provides for the stakeholder. They are based on the definitions by ISO (2017) and are listed in *table 3.3* below.

Table 3.3: Deliverables and Artifacts from the MS Development Process

Deliverable	Artifact
Information Product	Measurement System Specification
Indicators	Measurement System Architecture Specification
Derived Measures	
Base Measures	
Measurement Instrument	

As suggested by Staron et al. (2010), the measurement system in this thesis was developed using Microsoft Excel. This is due to the wide adoption within the collaborating company and therefore resulting increased understandability.

When a MS is developed, Staron et al. (2010) emphasizes the validation of the system, theoretically as well as empirically. For the theoretical validation, this thesis is using the theoretical framework of Kitchenham, Pfleeger, & Fenton (1995). The reason for that is, because the validation framework described by Kitchenham et al. (1995) allows to validate not only single attributes and measures but also derived measures (indirect measures), the measurement instrument and protocol. Further, the authors involve several other researchers work to discuss differences in their ways of validating, which gives a deeper understanding of the process of software measurement validation.

For the empirical validation, to verify whether the KPI measures what it actually is supposed to measure, Staron et al. (2010) suggest a pragmatic approach. First, the MS should be developed and deployed. Then, the MS is supposed to be observed over a period of time and assessed together with the stakeholder with regard to a comparison with the current measurement solution without the new MS. Finally, when the stakeholder's view is align with the indicator's values, the measures are considered to be empirically valid.

Although Staron et al. (2010) suggest to empirically validate the indicator with up-to-date data, because it is said to be more align with the up-to-date views of the stakeholder. They further state that despite considering historical data often being biased, it can still be sufficient to empirically validate a measurement system. In their work, Starong et al. (2010) develop a measurement system and validate it with current data. This thesis is making use of historical data due to the expression by the stakeholders that the way the company was measuring at that time is exactly the same as right now, which makes the historical data just as relevant as the current data. The results of the validation can be found in *chapter 5* and *Appendix C.*

Key Performance Indicator Quality Model

In theory, the use of Key performance indicators is very clear (Staron et al., 2016). Nevertheless, companies often struggle with putting "good" KPI's to work and linking them to their business strategy. Being able to create high quality KPI's requires knowledge and experience within an organization about their structures, goals, strategy and data sets (Staron et al., 2016). In order to answer the question whether an organization has created "good" KPI's, quality models are used to determine their quality based on different aspects.

Staron et al. (2016) developed a quality model that directly addresses the question of what characterizes a "good" KPI, with defining the KPI as a measure that is being actionable and supportive towards an organization's business goals. The quality model gathers quality attributes which not only evaluate the measures and the KPI itself, but also the measurement procedure from raw data to the actual KPI, as well as how well it fits into the organizational context. Like this, Staron et al. (2016) provide the ability for organizations to understand how well their KPI's are described, linked to their business strategy and how impactful they are for the company. Since this quality model has been developed during a collaboration with an industry partner, it fits well for the application in this thesis. Also, because the quality model is based on the ISO standards presented in previous chapters, it is well suited for the purpose of this study.

The quality model uses the ISO/IEC/IEEE 15939 and the ISO/IEC 25000 series as a theoretical foundation for the quality. Additionally, Staron et al. (2016) include the organizational change adoption theory and information quality to the foundation for the quality model in order to provide more attributes for the organizational context. these concepts are further explained below.

As presented in *chapter 2*, the ISO/IEC/IEEE 15939 standard provides a number of criteria for the quality of base and derived measures, measurement procedure and indicators in three categories (data analysis, data preparation and data collection). The lack of focus on the organizational context resulted in Staron et al. (2016) adopting a refined model described by Abran (2010) of the model presented in *figure 2.2*, in order to include the perspectives of the categories standard reference model and organizational reference context. Ultimately, these five categories are defined for the quality attributes to be grouped in.

When it comes to the ISO/IEC 25000 series, c*hapter 2* already presented a list of new criteria evaluating the quality of data. Further, Staron et al. (2016) complement this list with a set of quality attributes which were included in older versions of quality models from ISO standards.

To determine the quality of the effectiveness of a KPI, Staron et al. (2016) derive attributes from the organizational change adoption theory described by Goodman (1993) and Goodman, Bazerman, & Conlon (1980). The theory basically determines how well change is adopted in a company and how impactful it is. It gives the practitioners the ability to understand how a “good” KPI is put to work in an organization (Staron et al., 2016).

Regarding information quality, the ISO/IEC/IEEE 15939 already provides a number of criteria to determine whether the information from the measurement system is trustworthy. Additionally though, Staron et al. (2016) make use of another framework, because it provides an extensive set of 15 criteria to evaluate information quality in more detail. The framework used is the AIMQ framework described by Lee, Strong, Kahn, & Wang (2002).

In addition to the already gathered attributes from theory, Staron et al. (2016) complemented the list with 3 attributes that were derived from direct discussions with industry partners. These are transparency, actionability and traceability. All together, the quality model consists of 59 quality attributes, grouped in five categories. All attributes of the quality model that are used to evaluate the Apdex KPI in this thesis, their descriptions and category assignments are presented in *Appendix D*.

4 Research Design

This chapter introduces the research questions of the thesis and the research method that was chosen to answer these.

Research Questions

The aim of this thesis is to improve current performance measurement solution at the collaborating company and develop a KPI measurement system in order to satisfy the stakeholders information needs, namely monitoring *Application A's* responsiveness to user operations. Based on this aim, the following research questions and their sub questions will be answered in this thesis:

RQ1: *How can, in an industrial context, 'good' KPI's be identified and developed?*

In order to identify potential 'good' KPI's, existing measures within the organization need to be assessed, based on a theoretical framework, and opposed to the stakeholders information needs to identify gaps. These results can be used to identify actions to develop 'good' KPI's. For these actions, existing performance measurement standards and approaches are needed. That leads to the first sub question:

RQ1.1: *How can existing performance measurement standards and approaches be applied to develop 'good' KPI's?*

To address this question, an existing and established KPI is implemented into the company and known software measurement standards are applied to it to develop a measurement system that successfully monitors *Application A's* responsiveness. To prove its usefulness and and significance to the company and industry, the quality of the KPI needs to be determined. That leads to the second sub question:

RQ1.2: *How 'good' are the developed KPI's and how well do they fit into the organizational context?*

In order to determine whether the developed KPI’s is ‘good’ and how well it fits into the organizational context, the KPI’s are validated with internal, historical data and evaluated by applying an established, on performance measurement standards based quality model.

The answer to RQ1 and its sub questions will provide a KPI measurement system that includes ‘good’ KPI’s that are able to monitor *Application A’s* responsiveness and improve the collaborating organization’s performance measurement.

Action Research

To answer the research questions, action research as it is described by Susman and Evered (1978) was conducted. The research method includes a five-phase cyclical approach, which is illustrated in *figure 4.1*. It is also required as a first step to establish a client-server infrastructure or research environment. Each of these phases, in context to the research questions, are further explained below in the *Workflow* subchapter.

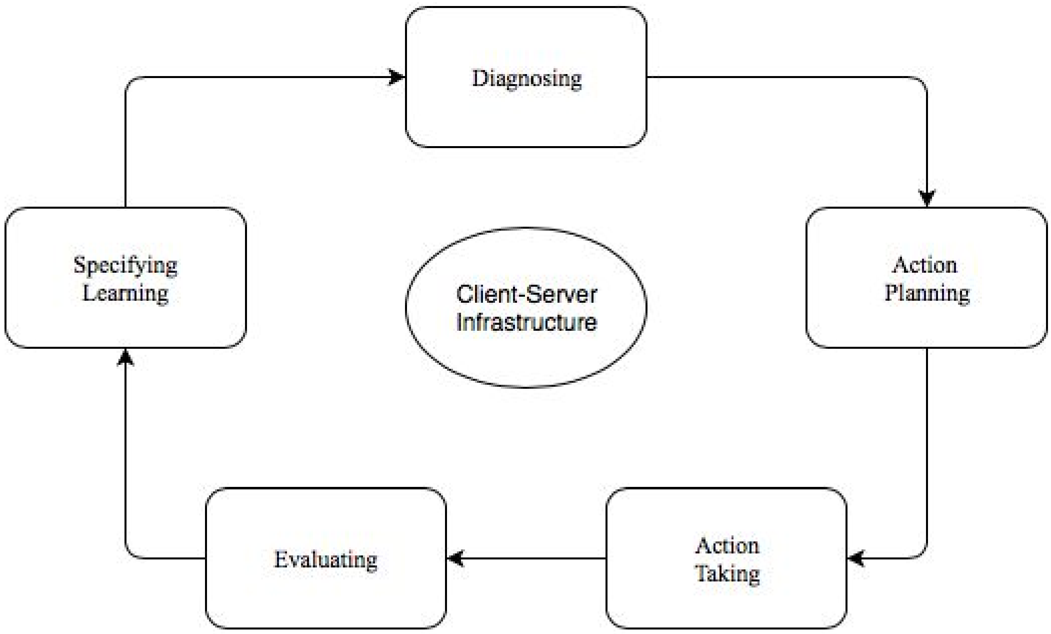

Figure 4.1: The Action Research Structural Cycle

Action research provides all stakeholders with highly relevant research results, due to the fact that it is based on a practical approach, with the goal of immediately solving a problem at hand while informing theory (Baskerville, 1999). Moreover, as this cycle continues, despite the outcomes being successful or not, further knowledge is created regarding the organization and the validity of the related theoretical framework.

Of significant importance is also the domain in which the action research method is used. Baskerville and Wood-Harper (1996) characterize the ideal domain by the active involvement of the researcher in order to create benefits for both parties, the applicability of the newly obtained theoretical knowledge onto the problem and the project being a process of linking theory to practice.

This approach fits well into the setting at the collaborating company, because it is a unique setting involving the interaction of several human subjects with different expertise. Both researchers and practitioners are involved in the problem solving process. Further, due to the immediate availability of history data, the applicability and evaluation of obtained findings and knowledge is ensured.

Workflow

Based on the aforementioned theory, the adopted workflor for this thesis is shown in *figure 4.2*.

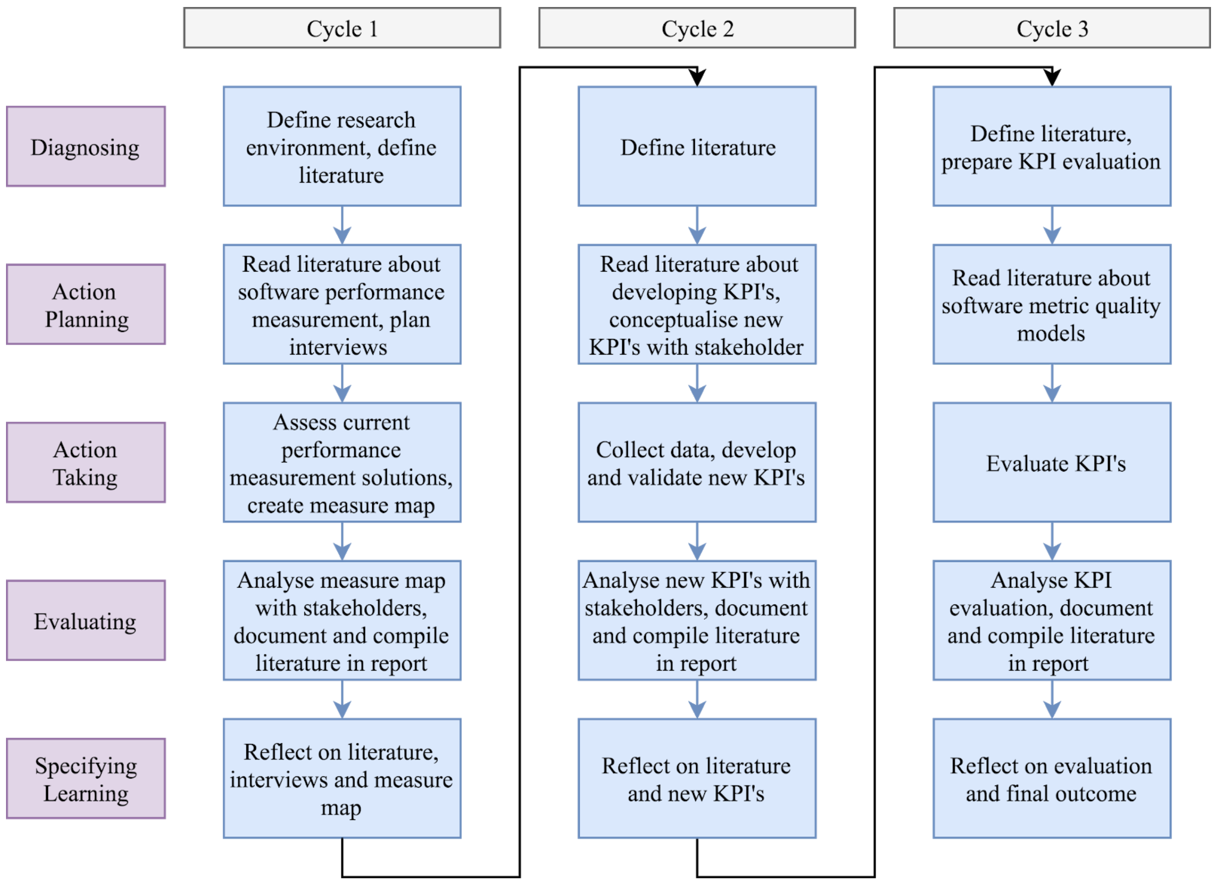

Figure 4.2: Adopted Workflow for the Action Research in this thesis

The workflow is divided into 3 cycles within the conducted action research. The first cycle provides a first insight into the improvement the industrial partner is seeking and provides a picture of the current actions that are taken in order to measure their applications performance. Furthermore, a map was created to identify a gap between the current state and a potentially desired future state. The second cycle is conducted to develop a KPI measurement system in order to close the gap and provide the stakeholders with KPI's that monitor *Application A's* responsiveness. The third and final cycle serves the purpose of evaluating the KPI's quality and usefulness with respect to scientific literature and the stakeholders needs.

The rest of the subchapter serves the purpose of further explaining the phases of the action research conducted throughout the three cycles in this thesis.

Client-Server Infrastructure

This subchapter describes the client-server infrastructure that was set between the host organization and the author of this thesis.

Before the first cycle, an agreement was set which clarified the entrance and exit of the researcher, the boundaries of the domain the stakeholders are working in and also responsibilities from the host organization and the researcher to one another. Baskerville (1997) stresses the importance of a client-server infrastructure as a prior step before the actual action research iterations, because it also specifies the authority under which the stakeholders take their actions. These conditions and also the requirement of a collaborative approach between researcher and practitioners are key aspects to a functioning action research client-server infrastructure described by Baskerville (1997).

After clarifying the entrance and exit for the researcher of this thesis, the boundaries of the domain for the thesis were set. The domain included the application under investigation, *Application A,* for which a performance measurement tool was set up in order to track its specific performance. *Application A* is a large product lifecycle management tool, which is crucial for the company's daily operations and therefore productivity. The domain and the performance tool's architecture is presented in a simple form in *figure 4.3* and is further described below.

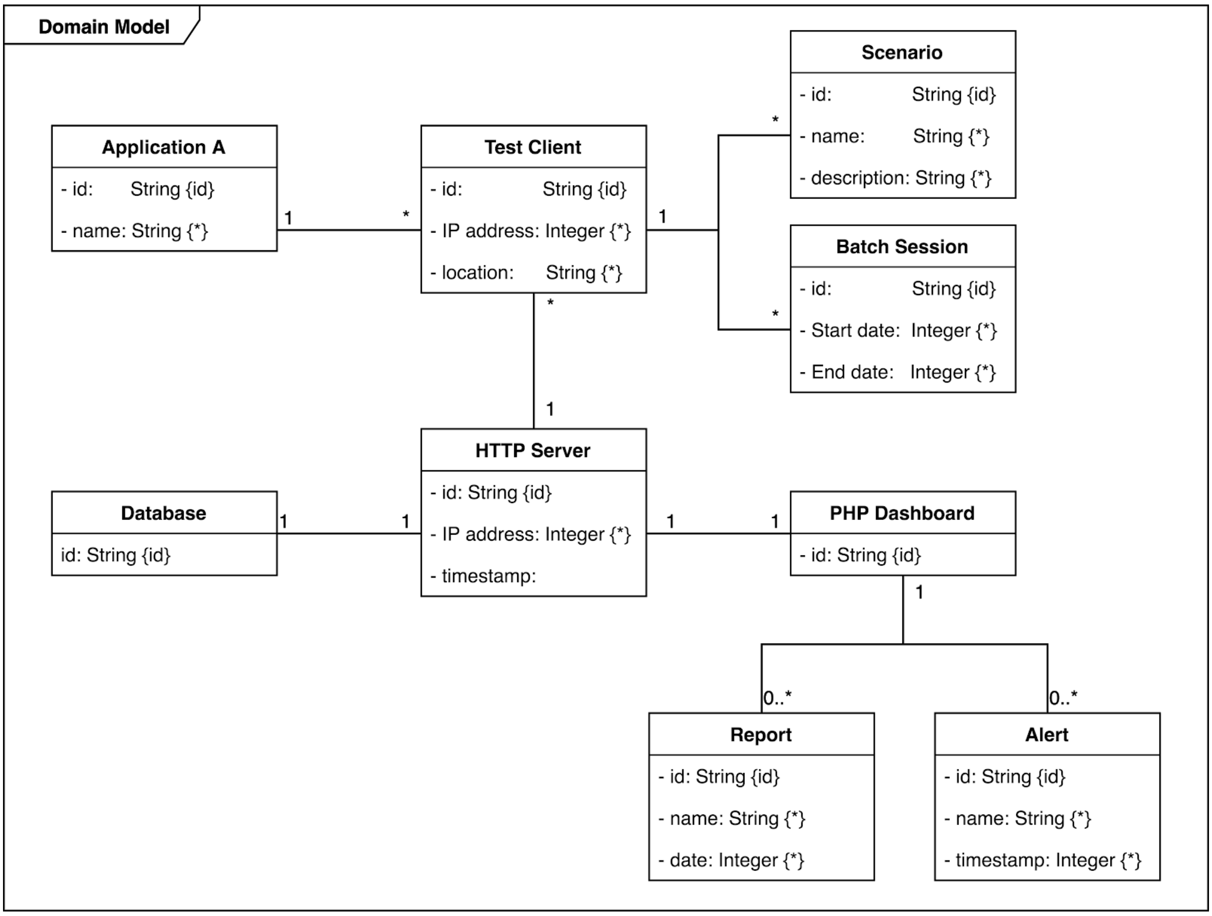

Figure 4.3: *Application A* Domain Model & Performance Tool Architecture

The measure probes of the measurement tool are installed as an application plugin on the test clients in different geographical locations (Sweden and China). The probes are executed as scheduled batch procedures in the *Application A* clients. The probes then run an *Application A* session. They collect data about *Application A* (i.e. server CPU, database calls, operation response times) and its infrastructure (i.e. network latency). The data is stored in the performance tool's own database, aggregated and then exposed by a HTTP server. A PHP-based dashboard provides queries, charts, tables and excel reports of the data stored in the database. Scheduled procedures check the data and send mail alerts if thresholds are exceeded (Siemens AG, 2012).

Figure 4.3 shows *Application A's* domain, however, the thesis mainly investigated the interaction between *Application A* and the test clients running their sessions since this is where the responsiveness of *Application A* to user operations is measured. The server was used to access and extract the historical data that used to validate and evaluate the measurement system.

Since the data was in a raw form and therefore rather complex and hard to understand, there was a need to condense this data into useful KPI's to properly describe *Application A's* responsiveness. The collaborating organization provided the researcher and author of this thesis with relevant information about *Application A*, its domain and the performance measurement tool in form of internal documents and interviews with employees that were responsible for *Application A's* maintenance and the performance tool. Further, the company provided the researcher with relevant data in order to assess current performance measurement actions and develop, test and evaluate the new KPI measurement system.

Diagnosing

This subchapter describes the actions that were performed during the diagnosing phase in each cycle, including defining relevant literature.

Defining literature

For each cycle, there was a need of analysing scientific literature in order to derive information valuable to the process of solving the organization's problem. Thus, in each cycle, relevant literature was identified, reviewed and used as a theoretical basis to plan and take actions together with the stakeholders at the collaborating organization.

For the first cycle, the focus laid on reviewing literature about performance and software measurement, measuring responsiveness and the notion of key performance indicators. Reviewing literature about software measurement provided the researcher with the necessary standards that built the foundation for this thesis. The notion of key performance indicators gave insight of how measures can and should be connected to the organizational context. the literature about measuring responsiveness gave valuable information about the actual information need an led to the discovery of the Apdex.

For the second and third cycle, literature was reviewed that based their work on the previously described software measurement standards. Further, literature about the Apdex was reviewed that helped implementing the KPI into the organization. The literature that is based on the software measurement standards provided the researcher with the development process for the measurement system and the quality model that was used to evaluate the KPI's quality. Also, literature was reviewed and applied to theoretically validate the KPI's and also to reveal threats to the validity of this thesis.

Action Planning

This subchapter describes the steps taken in the *action planning* phase, including reviewing relevant literature, planning interviews and conceptualizing potential KPI's.

Literature Review

The literature review was conducted iteratively for each cycle in order to identify theory that could be applied practically to the problem at hand. It serves as the body of knowledge for this thesis. Most references were retrieved online through accessing databases such as Scopus, Web of Science, ACM Digital Library, Science Direct or Springer Link, but also offline research was conducted.

In cycle 1, the review of literature regarding performance measurement and software performance measurement led to the involvement of the KPI notion into this thesis as well as the software measurement standards which were eventually used as the theoretical foundation. In the second cycle, the literature review led to the discovery of the responsiveness KPI (Apdex) that was used to develop the measurement system. Further, the literature provided a thorough measurement system development process which served as the guideline for the development and validation. In the last cycle, the reviews literature provided a quality model which was ultimately used to determine the KPI's quality and suitability in the organizational context. The findings give answers to RQ1, RQ1.1 and RQ1.2 and are presented in *chapters 2 and 3*.

Planning Interviews

There is a need to collect as much significant data as possible from the selected field experts within the company in order to be able to accomplish a deeper understanding of the problem and expected results. Therefore, it was decided to use semi-structured interviews as a method to collect primary data in cycle 1. The questions of these interviews can be found in *Appendix A*.

Semi-structured interviews are based on a list of topics and questions that the researcher wants to have answered during the session (Bryman and Bell, 2007). Further, Bryman and Bell (2007) describe semi-structured interviews as a technique to motivate the participant to freely express their own opinion on the problem at hand. While structured interviews follow a fixed set of question with expected answers, semi-structured interviews encourage the interviewees own interpretation of things. In other words, semi-structured interviews provide more flexibility than structured interviews or standardised surveys which was needed in the setting at the industrial partner.

The interviews were planned based on the identification of potential stakeholders to the thesis work through several meetings and discussions. The potential interviewees were then approached and, after giving consent, dates for the interviews were scheduled.

Conceptualizing KPI's

With the resulting measure map from assessing the current performance measurement solution and conducting interviews in cycle 1, a gap was identified between what was provided by these solutions and the stakeholders' actual needs, namely to properly monitor *Application A's* responsiveness to user operations. Based on these results from cycle 1 and the findings from the literature review in cycle 2, the Apdex was suggested to the stakeholders as a solution for their information need. After discussing the KPI, its implications and potential impact, it was decided to use the Apdex to develop a measurement system for monitoring Application A's responsiveness to user operations. The work from this step gives answers to RQ1 and RQ1.1.

Action Taking

This subchapter describes the steps taken in the *action taking* phase, including collecting data, assessing current performance measurement solution, conducting interviews, creating the measure map and developing, validating and evaluating the new KPI's.

Data Collection

While the primary data for cycle 1 and 2 was collected qualitatively through interviews, the data set for assessing the current performance measurement solution and developing, validating and evaluating the new KPI's was extracted from the company's current performance measurement tool for *Application A*. It contained two reports filled with history data ranging from the 01.01.2018 to the 30.06.2018. It included 13 technical and 7 non-technical base measures, such as Packet errors, number of packets lost, network latency, client CPU and user operation response times, from 2 geographically different client locations (Sweden and China).

The data set was exported from the performance measurement tool as a table in order to work with it. It was then filtered by office times, ranging from 7 AM to 5 PM depending on geographical location. Moreover, due to the results from cycle 1, the data set was reduced to the base measures relevant to the thesis work, then comprising response times for 6 user operation. Taken together, the reports listed 98.850 measurements, 16.475 for each operation, 8.202 for the Swedish and 8.273 for the Chinese location.

Assessing current performance measurement solutions

During this step in cycle 1, current solutions that provide the collaborating company with the performance measurements for *Application A* were assessed and analysed based on the ISO/IEC/IEEE 15939 measurement information model as described in *chapter 2*. This was done by getting to know the domain and the tools through reading internal guides and testing out the tools and applying the information model to the existing measures. The conducted interviews also helped with establishing a better understanding of the current state and the desired future state. Further, the measured base measures of *Application A* were assessed regarding their suitability to develop KPI's that meet the stakeholders' needs.

Interviews

In cycle 1, the selection of the interviewees was supported by a non-probability approach, which means that the selection was not random but carefully conducted based on the hopes of the researcher to get more valuable information from some employees than from others

(Bryman and Bell, 2007). The group of interviewees comprised of four people, two test engineers responsible for the maintenance of several performance measurement tools, an application manager of the application under investigation, as well as the IT director of the business application systems department at the collaborating company. These people were chosen because, based on the outcome of a number of meetings and discussions, they represented the group of employees with the largest knowledge pool relevant to the thesis.

The interviews were conducted to get a general understanding of the infrastructure that was set up for measuring the performance of *Application A*, identify desired improvements and elicit needs and requirements for a solution. The questions were asked to answer how relevant current measurement actions are and how the results of those relevant measurement actions can be used to develop a useful KPI measurement system. The questions can be found in Appendix A. The outcome of those interviews can be mapped to RQ1.

Measure Map

As described above in the previous subchapter, current performance measurement solutions were assessed. In cycle 1, the identified base measures were then mapped onto a measure map that represents the current state of performance measurement of *Application A*. The map is shown in *figure 5.1* in *chapter 5* and uses the definitions provided by ISO (2017). Moreover, the map helped to identify the gap between the current state and the stakeholders desired future state.

The Assessment of the current performance measurement solutions, the interviews as well as the resulting measure map give answers to RQ1 and its sub question RQ1.1.

KPI Measurement System Development, Validation and Evaluation

As already described in the *Conceptualizing KPI's* subchapter, the Apdex was suggested and chosen to be used to develop the measurement system. In the *action taking* phase however, the actual measurement system was developed, validated and evaluated. After the decision was made to use the Apdex as a KPI in cycle 2, the measurement system development process as described in *chapter 3* was applied to develop the Apdex into a measurement system based on the ISO standards presented in *chapter 2.* After the development, the MS was theoretically and empirically validated. For the theoretical validation, known theory was applied to ensure that the measurement system gains acceptance on scientific ground. The empirical validation was conducted together with the stakeholder by using historical data. The KPI was tested with historical data and presented to the stakeholder in an iterative process until the values of the KPI and the stakeholder's views were align. These steps gave additional answers to RQ1.1.

In cycle 3 however, an evaluation of the new KPI's was conducted by making use of the provided data set and a quality model derived from the literature presented in *chapter 3.* The data set helped with validating the KPI's, showing its consistent ability to monitor *Application A's* responsiveness over a longer period of time. The quality model described by Staron et al. (2016) was used to determine whether the KPI's fit into the organizational context, meet the stakeholders needs and to answer RQ1.2 of whether the KPI's are *"good"*.

Evaluating and Specifying Learning

This subchapter describes the actions taken to evaluate the work and results of each cycle and specify the knowledge that was gained throughout the thesis work.

While the *Evaluating* and *Specifying Learning* phase appear last after the three previous phases, these phases were running simultaneously in order to maximize the learning outcome (Baskerville, 1999). during each *Action Planning* and *Action Taking* phase in each cycle, the findings were evaluated in bi-weekly meetings with the stakeholders at the collaborating company. It was analysed whether the results were going into the right direction of relieving the problem at hand. Based on the feedback, further actions would either continue where left off or pivot into another direction. Furthermore, the knowledge gained by successful or unsuccessful actions give valuable insight for all stakeholders about the company's operations and also about the analysed body of knowledge (Baskerville, 1999).

5 Results

This chapter holds the results from the three cycles of the action research that was conducted during this thesis work. Starting off with displaying the results of cycle 1 following the assessment of the current performance measurement solution and the conducted interviews. Afterwards, the results of cycle 2, the development and validation of the new KPI measurement system, are presented. Lastly, the results following the evaluation of the new KPI are presented.

Cycle 1

The conducted Interviews with practitioners from the host organization, who were responsible for the maintenance of *Application A* and the performance measurement tool, resulted in identifying that the stakeholders were in need of KPI's that are able to track *Application A's* responsiveness. Furthermore, through additional literature review and assessment of *Application A's* performance measurement tools by the researcher, deeper understanding of the domain and the problem could be established.

The literature review about performance and software performance measurement helped the researcher to acquire knowledge about software measurement standards like ISO/IEC 25000 and ISO/IEC/IEEE 15939. These standards helped with the understanding of how software systems are properly measured and how these measures are developed into indicators stakeholders can act upon. Further, these standards provided knowledge on how to evaluate a measure's quality which served as a fundament for the evaluation of the developed MS in cycle 3.

Due to the better understanding in the fields of software measurement, the researcher was able to meet the interviewees on a more equal level and gain deeper insight about their motivation behind measuring *Application A*'s performance. The interviews showed that *Application A* is a crucial part of the collaborating company's daily operations. "*Application A is used by many engineers on different sites for daily design and data management* *operations. It is important for the success of our daily business and R&D. Recently, the application has shown improvement potential when it comes to its responsiveness to user operations, especially when users access from China. It has negative impact on the user satisfaction and productivity.*" (V. Vasekar, personal communication, May 4, 2018). It turned out that users have experienced bad responsiveness results to user operations from that application, which is why different performance measurement tools were set up. The goal of the tools are to measure different entities related to the application (i.e. network, database, responsiveness etc.) in order to establish long term monitoring, including the investigation of causes for bad responsiveness. Manual measurements with a stopwatch were conducted which were eventually decreased to a minimum, due to high resource efforts. Another measurement tool was set up to measure different entities around the application, but it turned out that the results were not accurate enough to use for long-term monitoring. As one interviewees stated "*the measurement of this tool is not accurate enough due to additional server calls of the tool itself that skew the response times*" (A. Jabbar, personal communication, May 5, 2018). In the end, a tool specifically designed for this application was set up at the collaborating company. This tool is vendored by the same company that provided the collaborating company with *Application A*. It comes with a set of features and measures a number of defined values for different entities surrounding the application as well as user operation response times, which is why the values presented in *table 5.1* are collected today.

From the interviews it became clear that although all of these measures are collected today, they are not sure how many of these measures are actually meaningful for them. One of the interviewees stressed that "*many measures are being tracked right now but there is not much action taken from it. Many of them have not the highest priority in the department. Focus is on long-term monitoring of how fast our application reacts to user operations*" (R. Maglica, personal communication, May 5, 2018). The interviewees expressed that as of right now, they do not prioritize technical measures but the response times of the application, which is why it was agreed on pursuing with the development of a responsiveness measurement system. Another issue is that the data in this tool is very unorganized and the visualization is messy because the data is visualized unprocessed in its raw form, making understandability and interpretations very difficult. As one interviewee stated that "*the visualization is too complex.*

We want to find deviations and present these deviations in a simple manner for everyone to understand. Simple visualization is needed" (R. Maglica, personal communication, May 5, 2018). Therefore the need was identified that these measurements, response time in particular, needed to be condensed into meaningful information. With the insight from the the interviewees it was possible to assess the performance tool itself from a perspective relevant to this thesis work.

The assessment of the tool showed that a high number of values are measured, which are presented in in *table 5.1* and give an overview of the status of the application and its related entities. This was done by accessing the tool itself and reading activity logs and experimenting with its measurement features, i.e. visualization feature or exporting data sets to excel. As already mentioned before, the focus lays on the non-technical measures, response times to user operations in particular. The tool is measuring these values several times a day and stores them in an own database. This data is then aggregated and can be visualized in graphs on the tool's web-dashboard using the tool's own visualization feature. Further, the data can be exported to excel. With the issues found from the interviews, the investigation showed quickly where these issues were located and why the needs of the stakeholders were expressed. The tool is measuring a lot of data but it seems to stop at that point and not develop further values from the raw data. With the understanding gained from studying related software measurement standards, the potential was identified to develop meaningful KPI's from the raw data that was perfectly provided by the measurement tool. An advantage was also that the data collection is already automated, which lays a good foundation for further work regarding data preparation and data analysis. In the end, the gap was identified between what the measurement tool provided today and what the stakeholders need was, namely condensed information about Application A's responsiveness to user operations.

To get a better overview of the measures from the tool and how they relate to each other, a measure map was created, which is presented in *figure 5.1.* The knowledge from the literature review helped to apply the measurement information model described by ISO (2017) and map the measures according to the definitions by ISO (2017). This served as a way to describe the relationship between the elements of the measurement tool. The map shows the current state of the performance measurement solution that are performed to track

Application A's performance. The Following table shows a summary the collected base measures and calculated derived measures by the performance measurement tool. A more detailed description of the measures based on the ISO/IEC 25000 format is provided in *Appendix B*.

Table 5.1: Summary of Based and Derived Measures from existing Performance Measurement tool

Base Measures	**Description**
Network Latency	Overall network latency in milliseconds
Network Latency (min)	The minimal network latency in milliseconds
Network Latency (max)	The maximal network latency in milliseconds
Number of Packet Errors	The number of network packet errors
Number of Packets lost	The number of network packets lost
JVM Memory used	The amount of actual memory that is used by the clients Java process
Number of Sessions	The number of *Application A* sessions at the start of the user operation execution
Client CPU Time	The number of seconds of overall time spent by the client CPU elaboration
Server CPU Time	The number of seconds of overall time spent by the Server CPU elaboration
User Operation Response Time	The duration of user operations 1-6 in seconds
SQL Time	The number of seconds of overall time spent executing SQL statement
SQL Calls	The number of SQL calls
SQL Errors	The number of SQL errors
Derived Measures	
Network Latency (avg)	The average network latency in milliseconds
Percentage of Packets lost	The number of network packets lost in relation to the total number of network packets sent
Overall Response Time	The cumulated time spent to execute user operation 1-6

The measure map as described is presented below. It shows what base measures, connected to what entity, are collected. Further, it displays what derived measures and indicators are developed from these base measures. Finally, it presents relevant stakeholders that hold interest in these measures. By definition indicators are characterized by being a product of one or more base or derived measures and the ability to provide an estimate or evaluation of attributes taken from a model considering defined information needs (Staron et al. 2016; ISO 2017). Based on this definition, the performance measurement tool at the collaborating company does not provide any indicators as of today. The company is using these measures and indicators though, since they are trying to interpret the results and act upon them. But this definition paired with the needs expressed in the interviews show that there is improvement potential when it comes to further develop the collected data into meaningful information.

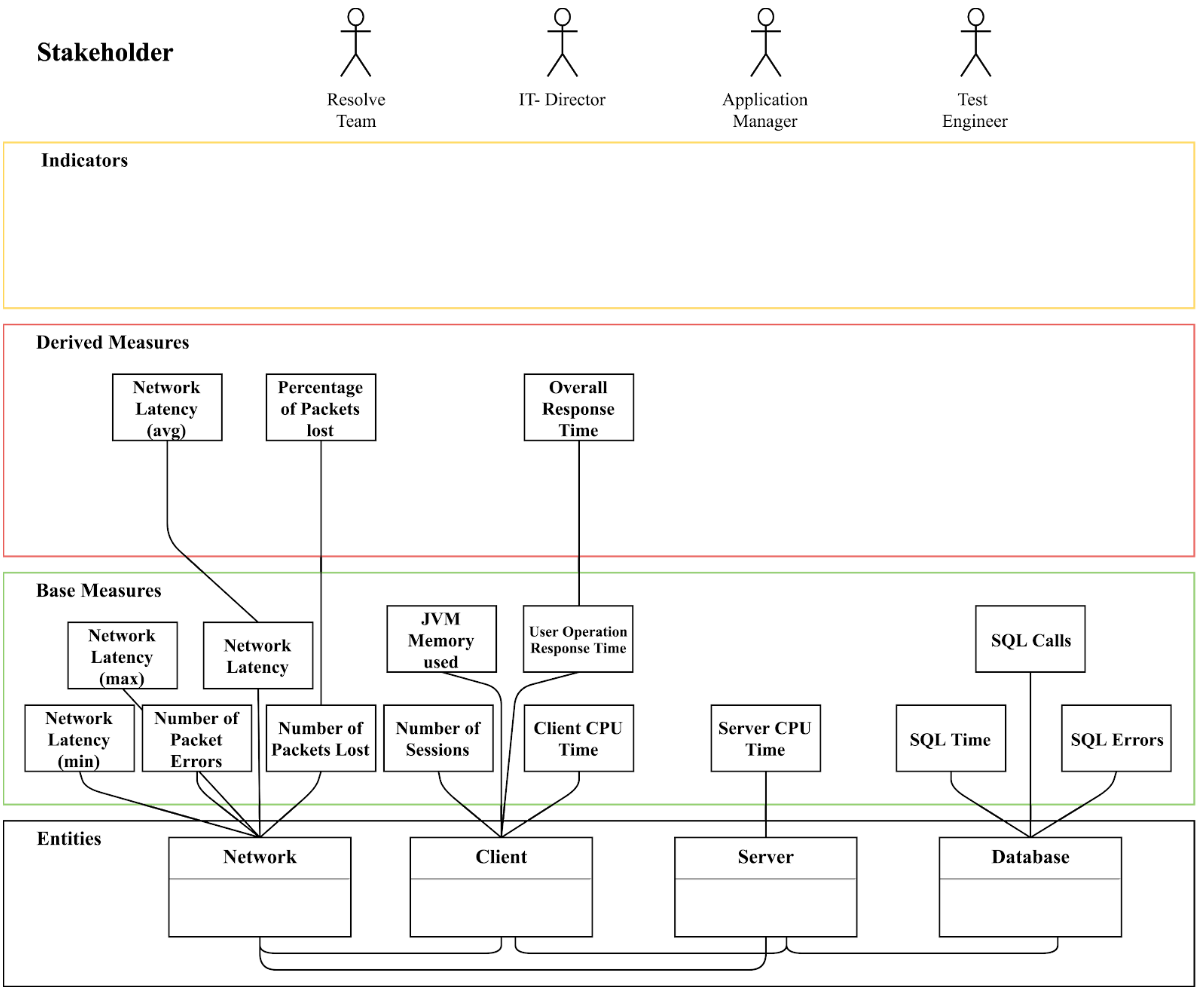

Figure 5.1: Measure Map *Application A*

As mentioned before, the conducted interviews also resulted in identifying a gap between the current state and the desired future state. It turned out that the practitioners do not put any emphasis on technical measures but stressed the need to condense the operation response times into KPI's that are able to describe the applications responsiveness with respect to the application's capabilities.

Further, the stakeholder stated the need of creating reasonable decision criteria in order to properly make interpretations and better support their decision making processes. Moreover, the KPI's should be able to satisfy the information need, which is tracking the responsiveness of *Application A* over time. Lastly, the stakeholders desire KPI which makes it possible to compare the responsiveness of *Application A* in different geographical locations.

Although the stakeholders at the collaborating company call their measures indicators, the application of the measurement information model by ISO (2017) reveals that they are no indicators. Nevertheless, their expression of the need to further condense their data into meaningful information is very align with what software measurement standards provide in terms of definitions of indicators and performance measurement. Further, current measures are not connected to specific stakeholders and their information needs.

In summary, following findings were identified:

- The measure map shows that mostly technical measures are collected and these measures are no actual indicators after ISO (2017), although the stakeholders think of them as indicators
- A gap is identified, because the information needs expressed by the stakeholders put focus on having KPI's for the applications responsiveness by condensing the operation response time data
- The measures are not connected to stakeholders, their information needs or any business goals/strategy
- A need is expressed to define reasonable decision criteria in order to properly make interpretations, evaluations and support the decision making process
- Another need was expressed to create KPI's that would enable proper comparison between the different performances of *Application A* based on different geographical locations

These findings defined the actions that were performed in cycle 2 in order to develop a measurement system for the industrial partner.

Cycle 2

In this cycle, the KPI measurement system was developed. Following subchapters present the results of the five steps of the second action research cycle, namely the development of the Apdex MS using the existing Apdex KPI and putting it into the context of ISO standards. Further, the results of the validation are presented.

Development of the Apdex Measurement System

The first two steps of the second cycle of the action research resulted in identifying the theory that was needed to develop the Apdex measurement system. Following the process of developing a measurement system described by Staron et al. (2010), based on ISO (2017), the steps as described in the process were taken to develop the MS. Further, based on the results from cycle 1 and the findings from the literature review in cycle 2, the Apdex was suggested to the stakeholders as a solution for their information need. After discussing the KPI in several meetings, its implications and potential impact, it was decided to use the Apdex to develop a measurement system for monitoring Application A's responsiveness to user operations. After these steps were successfully conducted, the Apdex measurement system could be developed in the next step.

As described in *chapter 3,* The Apdex is a KPI that reduces collected time data into meaningful information. It is expressed in a numerical number in the range of 0-1 and describes an application's responsiveness to user operations, with 1 being perfectly responsive and 0 being completely non-responsive (Sevcik, 2005). The following *table 5.2* presents a summary of the outcomes of activities 1 to 10 in the measurement system development process described by Staron et al. (2010) in *chapter 3*. The measurement method describes how the measures are being collected that are necessary to create the Apdex KPI. It is documented to make sure that these steps can be reproduced and conducted repeatedly. The base measures represent the values that needed to create derived measures and indicator. The measurement function describes the mathematical transformation from the base measure into a derived measure, based on the theory described in *chapter 3.* The derived measures include the satisfying response time (SRT), tolerating response time (TRT) and frustrating response time (FRT), which are calculated using the defined thresholds F and T. F is a function of T, while T is the only threshold that needs to be defined. How it is defined is further explained in *chapter 3* and the empirical validation section. The analysis model describes the decision criteria that were defined in iteratively occurring meetings together with the stakeholder to make it able to interpret a KPI's status. The indicator presents the KPI itself and the information need expresses the necessary state of fulfillment of their decision criteria. Lastly, the stakeholder is named that is tied to this measurement system.

Table 5.2: Summary of the Specification of the Apdex based on ISO/IEC/IEEE 15939

Stakeholder	Department Manager
Information Need	Are we fulfilling the Green responsiveness promise?
Indicator	Apdex Score: Application Perormance Index for User Operations
Analysis Model	Green: 0.85_T - 1.00_T
	Yellow: 0.70_T - 0.84_T
	Red: 0.00_T - 0.69_T
Derived Measures	Response Time opposed to Target Time T and Limit Time F
Measurement function	Categories:
	Satisfying Response Time (SRT): RT < T
	Tolerating Response Time (TRT): T < RT < F
	Frustrating Response Time (FRT): F < RT
Base Measures	Object: User Operation
	RT: Response Times of user operation x over a given period of 1 month.
	Object: Target Time T
	T: Defined Target Time by Stakeholder
	Object: Limit Time F
	F: Defined Limit Time as a function of T: F = 4T
Measurement Method:	For RT:
	1. Collect all response times of user operation x over the course of 1 month for both geographical locations.
	2. Oppose response and target time T and limit time F according to the measurement function.
	3. Calculate the Apdex according to its formula.
	4. Use analysis model to declare colour flag.
	For T:
	1. Define Target Time based on theory and *Application A's* capabilities
	For F:
	1. Calculate F with the given function: F = 4T

The table shows the definitions of the Apdex measurement system based on the measurement information model described by ISO (2017). It documents all elements that are needed for the creation of the Apdex KPI's.

The specification is further illustrated in *figure 5.2* and contains all measures that are used in the measurement system.

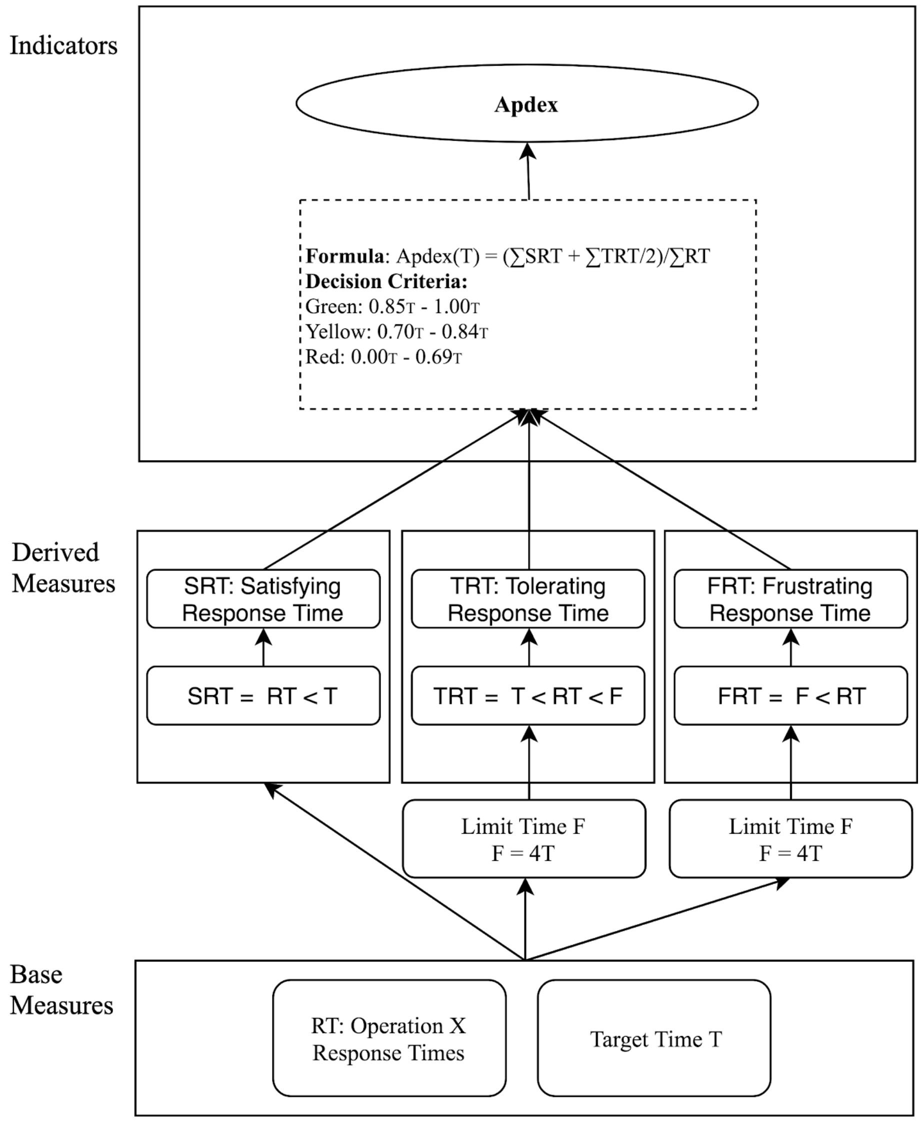

Figure 5.2: Graphical Specification of the Apdex Measurement System

Figure 5.2 serves as a graphical version of the Apdex measurement system to improve the understanding of the KPI. Another crucial factor here is the layout of the specification, which was chosen based on the ISO (2017) measurement information model in order to make it more straightforward to understand. A layout like this encourages stakeholders to spent more time on the ISO standards and improve the learning outcome, which might become useful when engaging in future projects.

The collaborating company is tracking the operation response times for two geographical location, Sweden and China, because the performances for these locations differ a lot. To this day, six user operations related to *Application A* are being measured for both locations. So, for each operation, two instances of derived measures exist. Each of them result in an Apdex KPI and thus provide the desired comparability. *Figure 5.3* shows how the Apdex measurement system is instantiated and illustrates its architecture specification, as inspired by Staron et al. (2010). The figure shows that for each operation there exist two KPI's, one for the location Sweden and one for the location in China. The instantiation is identical for every operation, which is why the figure shows the flow from operation one to six with the arrow at the top of the figure.

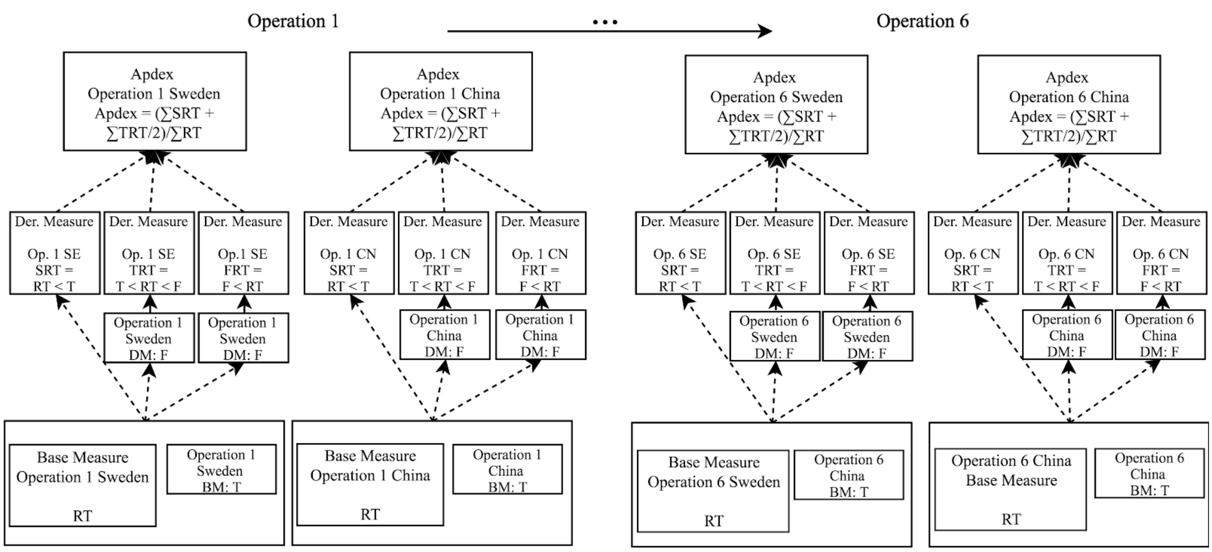

Figure 5.3:Instantiation of Apdex Measurement System (Architecture Specification)

For each operation (6 in total) the response times are collected. Furthermore, the defined thresholds of target time T and limit time F are collected as well. From these measures, 3 different derived measures are calculated, satisfying response time (SRT), tolerating response time (SRT) and frustrating response time (FRT). in the last instantiation, the *Apdex* is calculated using the defined formula.

As one can see from *figure 5.3,* already the presentation of two operations shows that a lot of space is needed for the collection, analysis and management of the data for these KPI's. Even for a rather simple measurement system like this one, the amount of data points can grow rapidly, especially with the fact in mind, that as of now only 6 user operations have been implemented into the performance measurement tool and a lot more are planned to follow. Therefore, the automation of the collection and analysis process is highly emphasized in order to maximize the system's effectiveness (Staron et al., 2016).

Lastly, based on the specification, architecture and instantiation, the measurement system is presented in a dashboard in Microsoft Excel, as shown in *figure 5.4*.

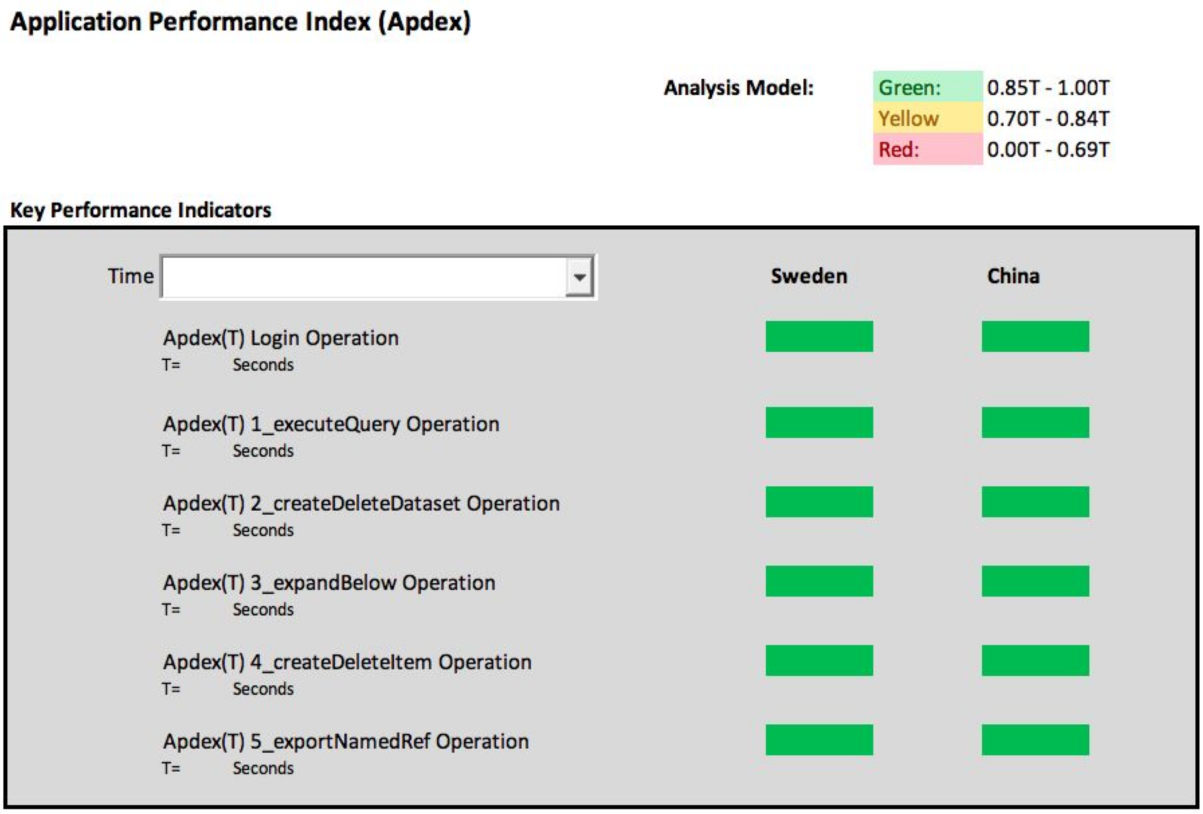

Figure 5.4: Measurement System for the Apdex: Presentation of the KPI's

The analysis model in the figure serve as an orientation for the user to remind him how the KPI is interpreted. In the shaded KPI window, all six operations are listed with coloured cells based on the interpretation result, dedicated for the KPI's for both geographical locations to ensure comparability. Furthermore, under each operation the defined threshold T is presented, so the practitioner can see what target value is defined at times for the user operation.

After the development of the MS, it was analysed together with the stakeholder and the whole process was reflected on. The results conclude that as of now, the measurement system is not automated. The MS was tested, validated and evaluated by the researcher and stakeholders by using the historical data and the measurement system was seen as useful and insightful. Further results from the evaluation and validation of the MS is presented in later subchapters. The automation of the measurement system and thus involved connection to the existing performance measurement tool was seen as not necessary at this stage by the stakeholder. Also, due to the time constraints of this thesis, it was not feasible. However, it was highly emphasized by the researcher that the process should be automated if the company decides to integrate the KPI's into the business operations. Then, for example, a daily update of the KPI's and a connected graphical visualization could be effective next steps. These and other improvement suggestions can be found in the results of cycle 3.

Validation and Example Use of the Apdex

This section presents the results of the theoretical and empirical validation of the Apdex KPI's and its measures.

Theoretical Validation

As described in *chapter 3*, the theoretical validation was conducted based on the theoretical framing by Kitchenham et al. (1995). The theoretical validation of measures allows the researcher to consider his measures as valid, based on a defined set of criteria from theory. Kitchenham et al. (1995) is proposing a framework that requires valid measures to confirm following validities:

- Attribute validity - The attribute under investigation is exhibited by the measured entity
- Unit validity - The unit of the measures fit the purpose of measuring the attribute

- Instrument validity - The measurement instrument and underlying models are accurate and measure the attribute with an appropriate given unit
- Protocol validity - An acceptable measurement protocol is applied

From these validities, Kitchenham et al. (1995) derive a number of properties that need to be fulfilled by the measures and measurement instruments in order to be considered valid. In the following paragraphs, the results of the Apdex and its measures regarding the fulfillment of those properties are presented.

Properties of Measures

The following table presents the validity criteria for measures by Kitchenham et al. (1995) and the results of that theoretical validation.

Table 5.3: Validation Results for Properties of Measure

Property	Validation Result	Valid?
For an attribute to be measurable, it must allow different entities to be distinguished from one another	There must exist two entities for which the measure results in different values. This is the case for the attribute of responsiveness in this thesis. The responsiveness of the application for example can be distinguished from the responsiveness of the connected server	Yes
A valid measure must obey the Representation Condition	The measure must always make sense in terms of the "real world" attribute and entity it is attempting to describe. This is the case because the response time and Apdex will always be able to measure the responsiveness of an entity in the real world (*i.e. Application A*)	Yes
Each unit of an attribute contributing to a valid measure is equivalent	That is standard measurement practice and correct in this case. Since The measures that are involved in the Apdex calculation only inherit the time unit seconds, it is considered valid	Yes
Different entities can have the same attribute value	This is the case. The responsiveness of Application A can theoretically have the same value of the responsiveness of another entity (i.e. another application or a server)	Yes

The table shows that all properties for measures described by Kitchenham et al. (1995) are fulfilled and the measures therefore can be considered valid.

Properties of Indirect Measures

Since indirect measures (in ISO (2017) defined as derived measures) are a combination of several measures, Kitchenham et al. (1995) defines additional properties for them to fulfill. These are presented in *table 5.4*, together with their results.

Table 5.4: Validation Results for Property of Indirect Measures

Property	**Validation Result**	**Valid?**
The indirect measure is based on a model concerning the relationship among attributes/measures defined on specific entities	This is the case. The whole context of the ISO/IEC/IEEE 15939 and ISO/IEC 25000 standards provide definitions based on their models and provide a process that justifies the Apdex' construction from these measures	Yes
The indirect measure based on a dimensionally consistent model	This is the case. The Apdex is measuring what it is supposed to measure, namely responsiveness (time). It is not measuring the size of the application for example	Yes
The indirect measure exhibits no unexpected discontinuities	The original Apdex formula as described in *chapter 3* shows discontinuity when a 0 is inserted. Therefore, for the sake of theoretical validity, the formula is aligned as seen below.	Yes
The indirect measure uses units and scale types correctly	This is the case. The Apdex uses scale types and unit types correctly and consistently	Yes

As it can be seen, the Apdex is fulfilling all properties, but in order to fulfill the property of discontinuity, the formula needed to be edited. The original Apdex formula shows problems when values of 0 are inserted. According to Kitchenham et al. (1995), if an indirect measure fails to fulfill one of these properties, it is considered invalid. Therefore the existing formula was changed to fulfill hundred percent validity. The edited formula is presented below.

$$Apdex = \frac{Satisfying + \frac{(Tolerating + 2)}{2}}{Total\ Amount\ of\ Response\ Times + 1} - 1$$

Like this, the formula ensures mathematical continuity.

Alternative Unit

Kitchenham et al. (1995) describes that the units used by the measurements need to be validated, specifically, if a unit is considered valid, an alternative unit that can result from the original units transformation (i.e. seconds to minutes) needs to be valid as well. The result of this validation is presented in *table 5.5.*

Table 5.5: Validation Results for Alternative Unit

Property	Validation Result	Valid?
If a measurement unit is accepted as a valid means of measuring an attribute, an alternative unit will be valid if the new unit is an admissible transformation from the original unit	This is the case since seconds (s) is a valid measurement unit of time to describe responsiveness as an attribute. And since seconds can be transformed into minutes or hours for this KPI, these units are considered valid as well.	Yes

Based on this property for validating alternative unit, it is considered valid.

Measurement Instrument

Kitchenham et al. (1995) also defines criteria for validating the measurement instrument if one is used. The property for this validation and its result is shown in *table 5.6* below.

Table 5.6: Validation Results for Measurement Instrument

Property	Validation Result	Valid?
The measurement instrument accurately measures attribute values in a given unit	This is the case. The automated performance measurement tool at the collaborating company accurately measures the response times for six user operations with a given unit of seconds.	Yes

Based on this property for validating the measurement instrument, it is considered valid.

Measurement Protocol

Also for the protocol of how to obtain the measures and indirect measures, Kitchenham et al. (1995) formulate a property that has to be fulfilled. It is presented in *table 5.7* with its result.

Table 5.7: Validation Results for Measurement Protocol

Property	Validation Result	Valid?
The measurement protocol is unambiguous, self-consistent and prevents double counting	This is the case. The procedure of how the performance measurement tool at the collaborating company is measuring its entities and attributes is well described and documented in several documents. It also prevents double counting in its setup. Further, the measurement protocol for creating the Apdex KPI is also well documented in this thesis, based on ISO standards.	Yes

Based on this property for validating the measurement protocol, it is considered valid.

All in all, the measures, KPI, measurement instrument and protocol in this thesis work are fully valid, according to the theory provided by Kitchenham et al. (1995).

Empirical Validation

The Apdex KPI's were empirically validated as part of the action taking and evaluating step in the action research. It is following the pragmatic approach described by Staron et al. (2010), as presented in *chapter 3.* After the measurement system was developed it was deployed and presented to the stakeholder. Then, the MS was validated using six months worth of historical data. This data was condensed into Apdex KPI's for six user operations. One critical issue was to define reasonable thresholds for the user operations, so that the resulting KPI's would have proper credibility and would actually satisfy the stakeholders information need. To achieve that, different thresholds were used to calculate different KPI's and presented to the stakeholder. The outcome was discussed and the work revised, if needed. This was an iterative process until the stakeholder's view was align with the values shown in the KPI's. After that, the measurement system was considered empirically valid. This section presents the Apdex KPI's empirical validation for two user operations for two geographical locations.

The first operation the Apdex KPI's were calculated for was a login user operation. The tables with the measurements and the resulting Apdex based on the theory described in this thesis are presented below. For this operation the thresholds were defined together with the stakeholder, also based on the theoretical framing described in *chapter 3* and the system's capabilities, as

- T = 7 seconds and
- F = 28 seconds.

Based on these thresholds, the three performance zones were created and the Apdex KPI's were calculated for both locations using the formula. The results are shown in *table 5.8* for the Swedish location and in *table 5.9* for the chinese location.

Table 5.8: Results Apdex Calculation for Login User Operation Sweden

Sweden					
Month	**Total Amount of Response Times (RT)**	**Satisfying: RT < T**	**Tolerating: T < RT < F**	**Frustrating: F < RT**	**Apdex(T)**
Jan	1293	716	519	58	0,75
Feb	1265	435	829	1	0,67
Mar	1397	358	1026	13	0,62
Apr	1368	558	791	17	0,70
May	1407	679	692	36	0,73
Jun	1342	651	681	10	0,74
Total	**8072**	**3397**	**4538**	**135**	**0,70**

Table 5.9: Results Apdex Calculation for Login User Operation China

China					
Month	**Total Amount of Response Times (RT)**	**Satisfying: RT < T**	**Tolerating: T < RT < F**	**Frustrating: F < RT**	**Apdex(T)**
Jan	1404	0	1381	23	0,49
Feb	1266	0	1264	2	0,50
Mar	1397	0	1387	10	0,50
Apr	1370	0	1352	18	0,49
May	1362	0	1362	0	0,50
Jun	1349	0	1342	7	0,50
Total	**8148**	**0**	**8088**	**60**	**0,50**

For each month, the total amount of measurements is presented together with the number of response times in each performance zone based on the defined thresholds above. The straight comparability between the two locations is visible on first sight. This satisfies one of the stakeholder's main needs expressed in the results of cycle 1. Also, in *table 5.8* one can recognize a trend the Apdex KPI is able to create. Due to the application of an analysis model with colour declared decision criteria, the Apdex KPI's show the trend of the responsiveness of *Application A*. This gives the KPI's the ability to be actionable and allow interpretations and evaluations. A further evaluation of the KPI's is presented in the subchapter of the results of cycle 3.

The second operation shows the usefulness of the KPI's when it comes to providing a proper trend of *Application A's* responsiveness. Further, the KPI's prove to be actionable and allow interpretation and an evaluation of the situation. The operation under investigation is an export operation and the Apdex KPI's were calculated using the same data set, 6 months worth of data for both geographical locations. The thresholds were defined together with the stakeholder, based on the theoretical framing described in *chapter 3* and the system's capabilities, as

- T = 3 seconds and
- F = 12 seconds.

The results from the calculations are shown in *table 5.10* for the Swedish location and in *table 5.11* for the chinese location.

Table 5.10: Results Apdex Calculation for Export User Operation Sweden

Sweden					
Month	**Total Amount of Response Times (RT)**	**Satisfying: RT < T**	**Tolerating: T < RT < F**	**Frustrating: F < RT**	**Apdex(T)**
Jan	1293	1245	31	17	0,97
Feb	1265	1263	1	1	1,00
Mar	1397	1063	334	0	0,88
Apr	1368	59	1308	1	0,52
May	1407	70	1333	4	0,52
Jun	1342	57	1284	1	0,52
Total	**8072**	**3757**	**4291**	**24**	**0,73**

Table 5.11: Results Apdex Calculation for Export User Operation China

China					
Month	**Total Amount of Response Times (RT)**	**Satisfying: RT < T**	**Tolerating: T < RT < F**	**Frustrating: F < RT**	**Apdex(T)**
Jan	1404	281	1123	0	0,60
Feb	1266	725	526	15	0,78
Mar	1397	661	660	76	0,71
Apr	1370	468	876	27	0,66
May	1362	420	899	43	0,64
Jun	1349	486	802	61	0,66
Total	**8148**	**3041**	**4886**	**222**	**0,67**

The tables contain the total amount of measurements for each months, the three performance zones and their number of response times and the calculated Apdex KPI's in the far right column. The KPI's clearly show the trend of the responsiveness for the location in Sweden. This allows the stakeholder to take action and further investigate to get to the root of the problem. Correlations to other components of the system can be made to find out what caused the problem. What is also to note is that through the comparability to another location, the organization can clearly determine that the issue only concerns the location in Sweden because the responsiveness measured in China has not experienced such a drop.

Four more user operations were empirically tested and validated. They can be found in *Appendix C*.

All in all, the empirical validation together with the stakeholder was successful. The Apdex measurement system was validated using historical data and it was perceived as useful and insightful. But also when the KPI's align with the stakeholders views, there is a need for a more detailed evaluation of the KPI's to determine their quality and how well they fit into the whole organizational context. Therefore, the final cycle of the action research focuses on this issue and the results of that evaluation along with potential improvements are presented in the following subchapter.

Cycle 3

The first two steps of the action research for cycle three resulted in the identification of a known quality model that was chosen to evaluate the KPI's quality and fit into the organizational context. Staron et al (2016) define a "good" KPI by its ability to be actionable and supportive towards the company's business goals in an objective. To evaluate KPI's based on this definition, a quality model was developed by the authors, containing a set of attributes which determine how well the measurement procedure and KPI's are defined, fulfill the company's needs and how meaningful and useful they are to the specific stakeholders (Staron et al., 2016). It also allows the organization to understand the difference between a "good" and a "bad" KPI and give suggestions for improvements with regard to their already collected and future indicators.

The evaluation of the *Apdex* KPI in this thesis was conducted in the third and last cycle of the action research, in the action taking and evaluation phases, based on the theoretical framework described in *chapters 2 & 3.* It was conducted by presenting the *Apdex* to the responsible stakeholder and assess it together based on the list of 59 attributes gathered by Staron et al. (2016). These attributes were gathered from the measurement standards ISO/IEC 25000 and ISO/IEC/IEEE 15939 as well as the organizational change adoption theory and the concept of information quality. Afterwards the list was visualized in a graph by counting the number of attributes that the *Apdex* fulfilled and divided the number by the total number of attributes for each category. Subsequently, the outcome was discussed and improvement suggestions for future improvements were outlined. The visualization is presented in following *figure 5.5* and for each category, the results and improvement suggestions are presented further below. The whole evaluation can be found in *Appendix D.*

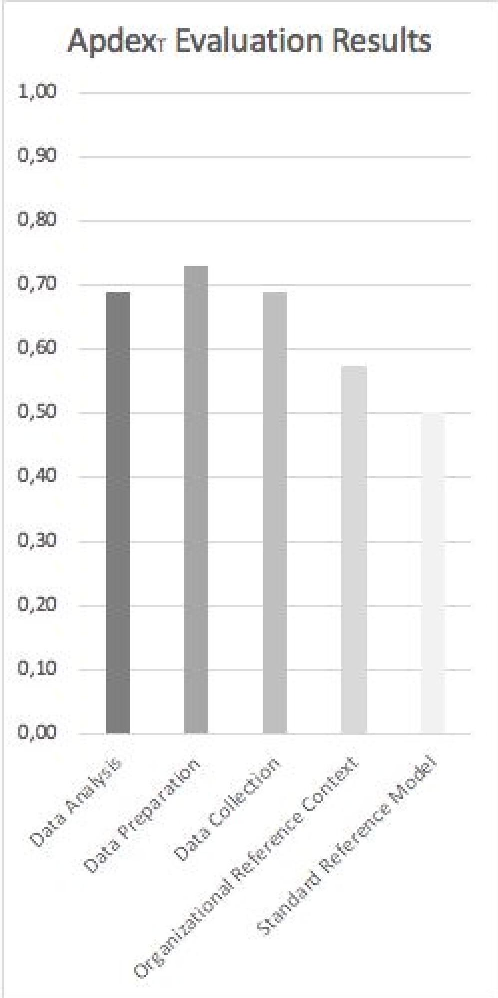

Figure 5.5: Apdex Evaluation Results

The data analysis category results in a ratio of 0.71. While the Apdex shows quality in the measurements procedure and fitting the stakeholder's information need, it lacks quality in being timely and linking towards the company's business goals. For the timeliness, this is due to the fact that the Apdex was validated using historical data and is not measuring up-to date data. Further, the collaborating company has not defined or documented any goals or information needs for tracking *Application A's* responsiveness or performance in general. Improvement suggestions here would be the implementation of the Apdex into the daily operations and tying it to a business goal once it is defined. Also, the Apdex shows quality because it does not include unnecessary information and is consistent with its intention of tracking *Application A's* responsiveness. Improvements could be made within the privacy of the KPI because as of now, it is not protected from unauthorized tampering. The following table summarizes the improvement suggestions for the data analysis category.

Table 5.12: Summary of improvement suggestions for data analysis category

Sub-Category	**Attribute**	**Improvement Suggestion**
Information Product	Up-to-Date (timely)	- Integrate Apdex into daily operation and analyse up-to-date data
Information Product	The KPI is linked to business goals	- Define a higher level business goal for tracking *Application A's* performance and tie the Apdex to it
Analysis Model	Privacy Protected	- Change the files privacy settings (i.e. password protection) and tie them to x persons responsible for them
Analysis Model	Protected from tampering	

Data preparation has the highest ratio of all categories with 0.73. Like in data analysis, attributes regarding the measurement procedure are included in this category, which give the KPI quality. All steps of the procedure are performed and performed correctly. The procedure is documented in files and it is therefore also clear how the procedure is performed and how results are obtained. Moreover, the value of the Apdex reflects the true value of the measurand and does not include irrelevant information. Improvements should be made by providing the the measurement error to the KPI, which is not given as of now. *Table 5.13* sums up the improvement suggestions for this category.

Table 5.13: Summary of improvement suggestions in the data preparation category

Sub-Category	**Attribute**	**Improvement Suggestion**
Derived Measure	Measurement Uncertainty (Measurement Error)	- Determine, provide and document the measurement error for the performance measurement tool and the Apdex KPI

The data collection category also includes the attributes regarding measurement procedure and error and there have analogically the same influence on the ratio as in the previous categories. Additionally, the Apdex shows quality in this category because the measurement procedure is easy to use and very straight forward. The base and derived measures are easily collected and calculated and there are no complex equations to determine the KPI’s. While there is support for the data collection of the Apdex, neither analysis nor management are supported. Also, besides the measurement tool in place itself, there is no tool supporting the measurement procedure. Therefore, improvements can be made here by for example automating the procedure and supporting it with a reporting tool. In total the data collection category reaches a ratio of 0.69. *Table 5.14* provides the list of improvement suggestions for this category.

Table 5.14: Summary of improvement suggestions in the data collection category

Sub-Category	**Attribute**	**Improvement Suggestion**
Measurement Method	Supported by human resources to collect, analyse and manage the KPI's	- Allocate resources for implementing the Apdex and using the dashboard for regular tracking of Application A's Responsiveness
Attributes	Measurement procedure is supported by available tools	- Automate the procedure by connecting the Apdex and the measurement tool for supporting data collection and the visualization - Automate the calculation process with the data from the measurement tool

The organizational reference context category gives a ratio of 0.57. The Apdex has quality because the stakeholder judges the KPI as useful and actionable. Further, it is clear to everyone how the Apdex is applied to *Application A* and it is thought to be believable. It scores a rather low ratio because the KPI has not been integrated into organizational context but much rather on the department or stakeholder level and is therefore not tied to any higher business strategy. Due to that, actions to improve the KPI's quality in this category would include integrating it into the daily operations and let it help making decisions throughout the department and the organization. *Table 5.15* includes the improvement suggestions for this category.

Table 5.15: Summary of improvement suggestions in the organizational reference context category

Sub-Category	**Attribute**	**Improvement Suggestion**
Reference values for context	Reputed (the KPI has a good reputation in the organization)	- Integrate the Apdex into daily operations and use it - communicate the usefulness and value of the KPI within the department and organization
Decision Criteria	The KPI is understood by stakeholder and organization	
Decision Criteria	Linked to specific business strategy	- Define a business strategy for tracking *Application A's* performance and tie the Apdex to it

The standard reference model category scores the lowest ratio with 0.50. While the stakeholders perceive that the Apdex will bring value to the company and is relevant to the information need, the KPI's is not known in the organization yet and is not characterized by being able to be used to make predictions. Further, since no actions have been taken upon the KPI's results, it is not known whether improving the KPI's would actually result in an improvement in the organization. When it comes to making predictions, the stakeholders' stressed that their need was not to be able make predictions but more to have solid KPI's that are able to properly track Application A's responsiveness. Therefore, there is no improvement needed when it comes to predictions. For the remaining attributes, the improvement would also be to integrate the KPI's into the daily business and let become known within the company. *Table 5.16* lists the suggested improvements for this category.

Table 5.16: Summary of improvement suggestions in the standard reference model category

Sub-Category	**Attribute**	**Improvement Suggestion**
Generic model of information needs	Provides consensus that improving the KPI results in an improvement in the company	- Use the Apdex during an improvement process and evaluate its usefulness afterwards
Generic model of information needs	KPI is known in the organization	- Integrate the Apdex into daily operations and use it - Communicate the usefulness and value of the KPI within the department and organization

All in all, the Apdex shows quality when it comes to the measurement procedure and the value it brings to the stakeholders and their information need. It is perceived as useful, believable and easy to use and understand. On the other hand, it lacks quality in categories like privacy and reputation, as well as the integration into the company by tying it to their business goals and strategy. The tables including the improvement suggestions and the evaluation itself also serve as valuable guidelines to use when assessing other existing measures in the company, for example for identifying other relevant or irrelevant indicators. It also helps for the development of future indicators, since the responsible stakeholders have an idea of what characterizes good KPI's.

6 Discussion

In this section the research questions with respect to the results in this thesis are answered. Furthermore, an outlook for future work within the collaborating company and research is presented. Finally, threats to validity of this thesis are discussed.

Answer to Research Questions

In this subchapter, the research question and its sub questions are answered in order to conclude this thesis work.

Research Question 1

The main research question to this thesis is "*How can, in an industrial context, 'good' KPI's be identified and developed?"* In order to answer this question, several steps had to be taken. In general, making use of the research method action research as described by Susman and Evered (1978) allowed the researcher to come as close as possible to an industrial setting in which an effective research environment could be created. By being able to directly apply all findings from theory to the practice at the collaborating company, it was possible to optimize the problem solving process and gain valuable knowledge for both the company and the researcher. Also, the iterative nature of action research enabled a close collaboration which provided a lot of constructive feedback that led to results of better quality. But to answer the question to its full extent, the two sub questions need to be answered. Therefore, the following paragraphs discuss the answers to RQ1's sub questions and sum up the answer to the main question in the end.

Research Question 1.1

To answer RQ1, its sub questions need to be answered first. The first sub question reads *"How can existing performance measurement standards and approaches be applied to develop 'good' KPI's?"* Reviewing literature in the software measurement field led to identifying the ISO/IEC 25000 series and the ISO/IEC/IEEE 15939, which helped describing the measures at the collaborating company. The measurement information model provided by ISO (2017) helped mapping existing measures and identify gaps between their definitions and how the practitioners used their measures. Further, the ISO/IEC 25000 series helped describing and documenting the measures in a more detailed way. Since these standards have a very straight forward approach, there were no difficulties applying them to the work at the industrial partner. Applying these standards helped a lot in terms of understandability and organization of the data.

Further, the MS development process as described by Staron et al. (2010) showed great applicability in this thesis. Firstly, because it is based on the aforementioned ISO standards and secondly, because it provided a perfect guide to develop ‘good’ KPI’s from raw and unorganized data. It was not necessary, though, to conduct all 18 steps provided by Staron et al.’s (2010) process, because some were already fulfilled at the collaborating company (i.e. develop measurement instrument, which was already given by the measurement tool). The pragmatic process to empirically validate the MS was also very supportive and could easily applied to the developed MS.

The Apdex KPI as developed by the Apdex Alliance inc. (2007) has proven to a valuable KPI to apply to this thesis work. In some aspects it needed alignment though. To ensure the theoretical validity of the measures, the original formula needed to be changed. Further, the Apdex makes use of two thresholds from which one needs to be defined. The framework for how to define the target time T is very generic. In order to find reasonable thresholds for T, a lot of effort needs to be put into discussions and experiments together with the stakeholder since its an individual setup for each company.

When it comes to the theoretical validation, this thesis made use of the framework provided by Kitchenham et al. (1995). It showed great applicability but there are issues with theoretical validation in software engineering. In general, there is a lot of different theory on how to conduct theoretical validation of software measures and it mostly is subject to different assumptions and is partly also contradicting each other (Kitchenham et al., 1995; Briand, Morasca, & Basili, 1996; Zuse. 2013; Schneidewind, 1992; Weyuker, 1988; Fenton, 1994; Melton, Gustafson, Bieman, & Baker, 1990). There is a need to come to a common agreement on a valid, consistent and comprehensive theory because otherwise the validation of software measurement is feared to be not accepted on scientific ground (Kitchenham et al., 1995). This thesis used the framework described by Kitchenham et al. (1995), because it has a broad set of criteria that fit the different aspects of software measurement in this thesis (i.e. the measurement instrument, measures and indirect measures). Nevertheless, the results of the theoretical validation might most likely differ if other approaches would have been applied, since there is no one valid framework that is commonly used.

A lot of theory was applied to develop a good KPI measurement system and improve the situation at the collaborating company. But there was still a need to determine how "good" the KPI's are and how well they fit into the organizational context. Which leads us to the second sub questions that is answered in the following paragraphs.

Research Question 1.2

The second sub question is "*How 'good' are the developed KPI's and how well do they fit into the organizational context?".* In this context, to extent the evaluation of the Apdex from the stakeholder in the empirical validation, a known quality model developed by Staron et al. (2016) was applied. The quality model is based on the ISO standards used in this thesis and shows therefore great alignment with the whole thesis work. The quality model as described in *chapter 3* contains a set of 59 attributes that can answer the sub question. Together with the stakeholder the attributes were evaluated as the results in *chapter 5* show. One issue that was encountered during the evaluation was that the measurement system was newly developed and a number of attributes required the KPI's to be already established within the organization (i.e. having a reputation, being known in the organization,etc.). Therefore, the MS shows lower quality than it might show if it was not newly developed.

With the answers to the sub questions the main research question can be answered. Action research is a powerful tool to get close to an industrial setting and directly apply theory to practice. Further, the application of software standards helped significantly with assessing current performance solutions and identifying gaps. They also laid the foundation for further research to be applied, in form of a development process and quality model. An established responsiveness KPI was implemented and put into context with the ISO standards that were used which really suited the stakeholders information need was properly align with the core of this work. The application of theory to validate the measures ensured further acceptance on scientific ground. Finally, the evaluation using a known quality model showed the KPI's strengths and improvement potential for future integration into the company's business operations.

Future Work

This section discusses possible future steps for the collaboration company in an outlook and discusses the contribution to research.

The results in this thesis have proven the usefulness of the Apdex for the collaborating company. Nevertheless, to exploit its full potential there are a number of steps that can be taken to optimize the effectiveness of the KPI. One crucial step is to take a step back and define a business strategy and performance goals. Having defined business goals ensure accountability for the achievement and the KPI itself when it is tied to them. Also, the discussion of defining reasonable thresholds for future user operations will be easier. Another step is the integration into daily operations and the automation of the measurement system. The KPI is condensing large amounts of data and conducting the collection, analysis and management manually costs time and money. The KPI also needs a reputation within the company. Employees need to know what the KPI is measuring and what the values mean. The integration also provides further evaluation for the KPI and ensures that resources are allocated to manage it. Lastly, improvements to the privacy of the measurement system can be made (i.e. password protection for the files to prevent tampering)

When it comes to contribution to research, this thesis does a great deal of proving theory by showing its applicability to an industrial setting. The development process described by Staron et al. (2010) shows its great alignment with ISO standards and an effective way to develop a valid measurement system in an industrial setting. Further, the quality model described by Staron et al. (2016) was well adopted by the stakeholders at the collaboration company and proved to be well applicable. This shows that it can well be used by other researchers and practitioners to evaluate KPI's in their own industrial setting. Moreover, the Apdex has proven to be able to get implemented, framed by ISO standards, in an industrial environment and provide an impactful way of tracking an application's responsiveness. It also proves to show great potential to be actionable and supportive towards stakeholders information needs.

Threats to Validity

Several threats to validity have been identified during this thesis work. These are categorized in four categories described by Wohlin et al. (2000) and are presented in the following sub chapters.

Conclusion Validity

The conclusions drawn from the results based on historical data might differ from the conclusion drawn from current data, which creates a threat to conclusion validity.

Also, since this thesis is not using any sort of statistics, there is a imminent threat to the conclusion validity. The threat was minimized by visualizing the data and discussing it iteratively to avoid misunderstandings.

Further, the measures in this thesis were assessed by applying the theoretical framing provided by ISO standards. The results of this assessment will differ, depending on the company's perspective or other theory.

Internal Validity

One threat to internal validity are the conducted interviews. Although the interviews were carefully chosen and conducted with people relevant to the thesis work, the sample might have been too small to get insight to its full extent.In this context, the researcher also might have missed out on potential interviewees that could have provided more insight that could have led to better results.

Another threat to validity is the usage of historical data. Although the stakeholder has stressed that the data is just as relevant as up-to-date data, results might have turned out different if historical data had not been used.

Construct Validity

The main threat to construct validity is that the thesis is making use different theoretical foundations to develop and evaluate the measurement system. This might result in differences in interpretation or the outcome if someone reproduces the case and wants to develop a measurement system. It is tried to minimize the threat by aligning the thesis work as close as possible with known software measurement standards.

External Validity

Since this thesis was conducted using the method of action research there is a natural threat to external validity due to the risk that the study might only be applicable in this specific setting. This threat was reduced by applying widely adopted standards that help the adoption by potential future researchers.

Lessons Learned

After conducting the work in this thesis, several lessons can be learned when it comes to developing 'good' KPI's in an industrial setting.

Firstly, when developing KPI measurement systems following specific processes in this thesis, it was found that it is often not necessary to follow every step of those processes. This can have various reasons based on the specific industrial setting. It can be that required steps are already fulfilled by the collaborating company or that steps are not needed in the specific context.

Furthermore, when developing KPI's one comes to the point when important thresholds have to be defined and an analysis model has to be created in order to make interpretations. The thesis work shows that it is crucial for the KPI's credibility and usefulness to properly define these thresholds. These thresholds are very dependent on the stakeholders' goals and business strategy. Existing development processes and also existing KPI's and their description lack to stress this issue enough. On the other side, when companies do not have defined well thought through goals for their cases, then defining reasonable thresholds become a real challenge.

Another lesson to learn from this thesis is that based on established quality models and further theory, it is basically impossible to create a ‘good’ KPI from scratch. Several quality attributes require the KPI measurement system to be already in place in order to fulfill them. Hence, when newly introducing KPI’s to the business, the quality of these indicators automatically suffer. It is therefore necessary to integrate the KPI’s into the business and let them establish themselves before their are ready to be properly evaluated and get a chance to fulfill all quality attributes.

7 Conclusion

This thesis aimed to improve the performance measurement situation at a company in the automotive industry by developing a KPI measurement system to improve the monitoring of a business application with respect to its responsiveness to user operations. To achieve that, the current performance measurement solution was assessed to describe existing measures and identify the stakeholders specific information needs.

To satisfy the information need of tracking their application's responsiveness, an existing and established KPI was implemented that suited the company's setting. To ensure the alignment with established software measurement standards, the KPI was put into context with the ISO/IEC/IEEE 15939 and ISO/ITEC 25000 standards. The process of developing the measurement system and its evaluation was based on theory which was also align with the relevant software measurement standards.

To determine the measurement system's significance for the company and scientific research, it was theoretically and empirically validated. The results of the theoretical validation showed that the measurement system shows validity in all properties and therefore ensures acceptance on scientific ground. The empirical validation proved the measurement system's usefulness and reveals some critical aspects to the successful use of it.

The evaluation of the MS determined the quality of the KPI and how well it fits into the organizational context. The results show a positive impact for the collaborating company, while leaving room open for further improvement and integration into the business operations. Further, the thesis does a great deal of proving the applicability of the theory to an industrial case. This study is conducted in the context of software engineering but the development of the measurement system can benefit other industries that work with performance measurement.

8 Bibliography

Susman, G., & Evered, R. (1978). An Assessment of the Scientific Merits of Action Research. *Administrative Science Quarterly,23*(4), 582-603. Retrieved from https://doi.org/10.2307/2392581

Baskerville, R. L. (1999). Investigating information systems with action research. *Communications of the AIS*, 2(3), 4. Retrieved from https://dl.acm.org/citation.cfm?id=374476

Baskerville, R. L. (1997). Distinguishing action research from participative case studies. *Journal of systems and information technology*, 1(1), 24-43. Retrieved from https://doi.org/10.1108/13287269780000733

Baskerville, R. L., & Wood-Harper, A. T. (1996). A critical perspective on action research as a method for information systems research. *Journal of Information Technology*, 11(3), 235-246. Retrieved from https://doi.org/10.1080/026839696345289

Bryman, A., & Bell, E. (2015). Business research methods. Oxford University Press, USA.

Siemens AG (2012). TPM V4. Unpublished internal document.

International Standard Organization. (2017). Software and systems engineering, measurement process (ISO/IEC/IEEE 15939:2017). Retrieved from https://www.iso.org/standard/71197.html

Staron, M., Meding, W., Karlsson, G., & Nilsson, C. (2011). Developing measurement systems: an industrial case study. *Journal of Software Maintenance and Evolution: Research and Practice*, 23(2), 89-107. Retrieved from https://doi.org/10.1002/smr.470

Sevcik, P. (2005). Defining the application performance index. *Business Communications Review*, 20. Retrieved from http://apdex.org/docs/Defining_The_Application_Performance_Index.pdf

Staron, M., Meding, W., Niesel, K., & Abran, A. (2016). A Key performance indicator quality model and its industrial evaluation. *Software Measurement and the International Conference on Software Process and Product Measurement (IWSM-MENSURA), 2016 Joint Conference of the International Workshop*, 170-179. IEEE. Doi: 10.1109/IWSM-Mensura.2016.033

Herbsleb, J. D., & Moitra, D. (2001). Global software development. IEEE software, 18(2), 16-20. Doi: 10.1109/52.914732

Lazar, J., Jones, A., & Shneiderman, B. (2006). Workplace user frustration with computers: An exploratory investigation of the causes and severity. *Behaviour & Information Technology*, 25(03), 239-251. Retrieved from https://doi.org/10.1080/01449290500196963

Lazar, J., Jones, A., Hackley, M., & Shneiderman, B. (2006). Severity and impact of computer user frustration: A comparison of student and workplace users. *Interacting with Computers*, 18(2), 187-207. Doi: 10.1016/j.intcom.2005.06.001

International Standard Organization. (2011). Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- System and software quality models (ISO/IEC 25010:2011). Retrieved from https://www.iso.org/standard/35733.html

International Standard Organization. (2012). Systems and software engineering -- Systems and software Quality Requirements and Evaluation (SQuaRE) -- Quality measure elements (ISO/IEC 20021:2012). Retrieved from https://www.iso.org/standard/55477.html

Upadhaya, B., Munir, R., & Blount, Y. (2014). Association between performance measurement systems and organisational effectiveness. *International Journal of Operations & Production Management*, 34(7), 853-875. Retrieved from https://doi.org/10.1108/IJOPM-02-2013-0091

Schott, G. (1981). Kennzahlen: Instrument der Unternehmensführung. Forkel.

Staron, M., & Meding, W. (2018). Software Development Measurement Programs: Development, Management and Evolution. Springer.

Neely, A., Gregory, M., & Platts, K. (1995) Performance measurement system design: A literature review and research agenda. *International Journal of Operations & Production Management*, 15(4), 80-116. Retrieved from https://doi.org/10.1108/01443579510083622

Staron, M. (2012). Critical role of measures in decision processes: Managerial and technical measures in the context of large software development organizations. *Information and Software Technology*, 54(8), 887-899. Retrieved from https://doi.org/10.1016/j.infsof.2012.02.003

International Bureau of Weights, & Measures. (1984). International vocabulary of basic and general terms in metrology. International organization for standardization.

Gopal, A., Mukhopadhyay, T., & Krishnan, M. S. (2005). The impact of institutional forces on software metrics programs. *IEEE Transactions on Software Engineering*, 31(8), 679-694. Doi: 10.1109/TSE.2005.95

Kilpi, T. (2001). Implementing a software metrics program at Nokia. *IEEE software*, 18(6), 72-77. Doi: 10.1109/52.965808

Kuzniarz, L., Huzar, Z., Reggio, G., Sourrouille, J. L., & Staron, M. (2003). Workshop on Consistency Problems in UML-based Software Development II. Retrieved from http://www.diva-portal.org/smash/get/diva2:838199/FULLTEXT01.pdf

Pandazo, K., Shollo, A., Staron, M., & Meding, W. (2010). Presenting software metrics indicators: a case study. *Proceedings of the 20th International Conference on Software Product and Process Measurement (MENSURA),* 20(1). Retrieved from http://www.cse.chalmers.se/~miroslaw/papers/2010_presenting.pdf

Staron, M., Meding, W., Hansson, J., Höglund, C., Niesel, K., & Bergmann, V. (2015). Dashboards for continuous monitoring of quality for software product under development. *Relating System Quality and Software Architecture*, 209-229. Retrieved from https://doi.org/10.1016/B978-0-12-417009-4.00008-9

St-Louis, D., & Suryn, W. (2012). Enhancing ISO/IEC 25021 quality measure elements for wider application within ISO 25000 series. IEEE. Retrieved from https://www.researchgate.net/profile/Witold_Suryn/publication/261297743_Enhancing_ISOIEC_25021_quality_measure_elements_for_wider_application_within_ISO_25000_series/links/57d1aef508ae6399a38b8434/Enhancing-ISO-IEC-25021-quality-measure-elements-for-wider-application-within-ISO-25000-series.pdf

Abran, A., Al-Qutaish, R. E., Desharnais, J. M., & Habra, N. (2005). An information model for software quality measurement with ISO standards. *Proceedings of the International Conference on Software Development (SWDC-REK), Reykjavik, Iceland,* 104-116. Retrieved from https://s3.amazonaws.com/academia.edu.documents/32335621/Rek%28Square%292005.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1536925823&Signature=DgDj06KwV0qQRNQQEvpbalyDVO0%3D&response-content-disposition=inline%3B%20filename%3DAn_Information_Model_for_Software_Qualit.pdf

Parmenter, D. (2015). Key performance indicators: developing, implementing, and using winning KPIs. John Wiley & Sons.

Kaplan, R. S., Kaplan, R. S., Norton, D. P., & Norton, D. P. (1996). The balanced scorecard: translating strategy into action. Harvard Business Press.

Weik, M. (2000). Computer science and communications dictionary. Springer Science & Business Media.

Nielsen, J. (1994). Usability engineering. Elsevier.

Ramsay, J., Barbesi, A., & Preece, J. (1998). A psychological investigation of long retrieval times on the World Wide Web. *Interacting with computers*, 10(1), 77-86. doi: 10.1016/S0953-5438(97)00019-2

Bhatti, N., Bouch, A., & Kuchinsky, A. (2000). Integrating user-perceived quality into web server design. *Computer Networks*, 33(1-6), 1-16. doi:10.1016/S1389-1286(00)00087-6

Apdex Alliance Inc.. (2007). Application Performance Index - Apdex Technical Specification (V. 1.1). Retrieved from apdex.org

Abran, A. (2010). Software metrics and software metrology. John Wiley & Sons.

Goodman, P. (1993). The Practical Implementation of Software Metrics. McGraw-Hill, Inc..

Goodman, P. S., Bazerman, M., & Conlon, E. (1979). Institutionalization of Planned Organizational Change (No. 1). *Carnegie-Mellon University Pittsburgh PA Graduate School of Industrial Administration*. Retrieved from http://www.dtic.mil/dtic/tr/fulltext/u2/a075221.pdf

Kitchenham, B., Pfleeger, S. L., & Fenton, N. (1995). Towards a framework for software measurement validation. *IEEE Transactions on software Engineering*, 21(12), 929-944. doi: 10.1109/32.489070

Briand, L. C., Morasca, S., & Basili, V. R. (1996). Property-based software engineering measurement. *IEEE transactions on software Engineering,* (1), 68-86. doi: 10.1109/32.481535

Zuse, H. (2013). A framework of software measurement. Walter de Gruyter.

Fenton, N. (1994). Software measurement: A necessary scientific basis. *IEEE Transactions on software engineering*, 20(3), 199-206. doi: 10.1109/32.268921

Melton, A. C., Gustafson, D. A., Bieman, J. M., & Baker, A. L. (1990). A mathematical perspective for software measure research. *Software Engineering Journal*, 5(5), 246-254. doi: 10.1049/sej.1990.0027

Schneidewind, N. F. (1992). Methodology for validating software metrics. *IEEE Transactions on software engineering*, 18(5), 410-422. doi: 10.1109/32.135774

Weyuker, E. J. (1988). Evaluating software complexity measures. *IEEE transactions on Software Engineering*, 14(9), 1357-1365. doi: 10.1109/32.6178

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wess- lén, A. (2000). Experimentation in Software Engineering: An Introduction. Kluwer Academic Publishers. Retrieved from https://doi.org/10.1023/A:1026586415054

Appendix

A - Interview Questions and Consent

A1: Interview Consent - Oral Data Collection Protocol

ORAL DATA COLLECTION PROTOCOL

Informed Consent:

Assessment and Development of Key Performance Indicators to describe the Relative Performance of Software Engineering Support Infrastructure

Jakob Noetzel – 19920916-3535
Lärdomsgatan 11, 417 56 Göteborg
+49 160 90389985
Jakob_noetzel@hotmail.de

Procedure and Risks:

I would like to record the interview, if you are willing, and use the tapes to write our materials. I will record the interview only with your written consent, and will ask that no personal identifiers be used during the interview, to ensure your anonymity. Please feel free to say as much or as little as you want. You can decide not to answer any question, or to stop the interview any time you want. The tapes and transcripts will become the property of the study.

If you choose so, the recordings and recording-transcripts (or copy of notes taken) will be kept anonymous, without any reference to your identity, and your identity will be concealed in any reports written from the interviews.

There are no known risks associated with participation in the study.

Benefits:

It is hoped that the results of this study will benefit the community through providing greater insight into the culture and history of the investigated area.

Confidentiality:

All information collected during the study period will be kept strictly confidential. No publications or reports from this project will include identifying information on any participant without your signed permission, and after your review of the materials. If you agree to join this study, please sign your name on the following page.

INFORMED CONSENT FOR INTERVIEWS

Assessment and Development of Key Performance Indicators to describe the Relative Performance of Software Engineering Support Infrastructure

I, _____________________________________, agree to be interviewed for the project ***Assessment and Development of Key Performance Indicators to describe the Relative Performance of Software Engineering Support Infrastructure***, which is being produced by Jakob Noetzel from Chalmers University of Technology

I certify that I have been told of the confidentiality of information collected for this project and the anonymity of my participation; that I have been given satisfactory answers to my inquiries concerning project procedures and other matters; and that I have been advised that I am free to withdraw my consent and to discontinue participation in the project or activity at any time without prejudice.

I agree to participate in one or more electronically recorded interviews for this project. I understand that such interviews and related materials will be kept completely anonymous, and that the results of this study may be published in an academic paper.

I agree that any information obtained from this research may be used in any way thought best for this study.

____ Date ________________________
Signature of Interviewee

If you cannot obtain satisfactory answers to your questions or have comments or complaints about your treatment in this study, contact:

Jakob Noetzel
+49 160 90389985
Jakob_noetzel@hotmail.de

Final Consent Form

Assessment and Development of Key Performance Indicators to describe the Relative Performance of Software Engineering Support Infrastructure

Dear Participant:

This form gives us final authorization to use material from your interview in **Assessment and Development of Key Performance Indicators to describe the Relative Performance of Software Engineering Support Infrastructure**. A draft of these materials should have been presented to you for your review, correction, or modification. You may grant use rights for this draft "as is," or with the modifications you specify, if any. See "Conditions" at the bottom of the form

I, ____, hereby grant the right to use information from recordings and or notes taken in interviews of me, to Jakob Noetzel from Chalmers University of Technology. I understand that the interview records will be kept by the interviewer and the project, and that the information contained in the interviews may be used in materials to be made available to the general public.

____Date:___________________________
Signature of Interviewee

____ Date:___________________________
Signature of Interviewer

The following conditions limit the release of information, as agreed between the interviewer and the interviewee:

_____ None needed

_____ Material may be released once corrections I specified have been made

_____ Material may be released once it has been edited by a third part (please specify)

A2: Interview Questions

The following questions were asked in the conduction of the semi-structured interviews at the collaborating company in order to elicit information about the current state of the performance measurement and the desired future state.

1. What is your position and responsibility at the company?
2. How are you involved in the performance measurements of *Application A?*
3. What is the focus/purpose of the performance measurements?
    a. What are current performance measurement techniques and how do they work?
    b. Are you satisfied with what values are measured with the performance measurement tools?
    c. Do the methods/techniques have any limitations?
    d. How is the data stored, processed and presented/visualized?
    e. What do you think about the way the data is processed and visualized? Are you satisfied?
    f. Do you think the performance measurement methods can be improved?
4. What happens when a cause of poor performance is identified? How is reported and how do you solve the problem / improve the situation?
    a. Are you satisfied with that process?
5. Do you think the performance measurements bring value to the overall performance of the company? Is the performance measurement beneficial for you?
6. Is there anything else you want to add to the topic?

B - Base measures and derived measures of the current performance measurement tool

The following base and derived measures are measures by the current measurement tool in place at the collaborating company. They are presented in the form based on the ISO 25000 series and summarized in *Table 5.1* in *chapter 5*.

Table B1: Definition of Network Latency using the format of QME from ISO/IEC 25000

QME	**Description**
QME category	Time Duration
QME name	Overall Response Time
QME ID	ORT
Detail	Overall Response Time measures the cumulated response time of all executed user operation implemented in the performance measurement tool
Input	Performance Measurement Tool
Documentation	ORT defines the cumulated responsiveness to all user operations executed in the performance measurement tool of *Application A*
Measurement Scale	Nominal
Measurement Focus	Internal

Table B2: Definition of Network Latency (min) using the format of QME from ISO/IEC 25000

QME	**Description**
QME category	Number of I/O
QME name	Network Latency (min)
QME ID	NL (min)
Detail	Network Latency (min) measures the minimal latency in milliseconds
Input	Performance Measurement Tool
Documentation	NL (min) defines the minimum latency of the application during the execution of a user operation
Measurement Scale	Ordinal
Measurement Focus	Internal

Table B3: Definition of Network Latency (max) using the format of QME from ISO/IEC 25000

QME	Description
QME category	Time Duration
QME name	Overall Response Time
QME ID	ORT
Detail	Overall Response Time measures the cumulated response time of all executed user operation implemented in the performance measurement tool
Input	Performance Measurement Tool
Documentation	ORT defines the cumulated responsiveness to all user operations executed in the performance measurement tool of *Application A*
Measurement Scale	Nominal
Measurement Focus	Internal

Table B4: Definition of Number of Packet Errors using the format of QME from ISO/IEC 25000

QME	Description
QME category	Number of faults
QME name	Number of Packet Errors
QME ID	PE
Detail	Number of Packet Errors measures the number of network packet errors
Input	Performance Measurement Tool
Documentation	PE defines the number of network packet errors counted of the application during the execution of a user operation
Measurement Scale	Nominal
Measurement Focus	Internal

Table B5: Definition of Number of Packet Lost using the format of QME from ISO/IEC 25000

QME	Description
QME category	Number of faults
QME name	Number of Packet lost
QME ID	PL
Detail	Number of Packet Lost measures the number of network packet that has been lost
Input	Performance Measurement Tool
Documentation	PL defines the number of network packets lost of the application during the execution of a user operation
Measurement Scale	Nominal
Measurement Focus	Internal

Table B6: Definition of JVM Memory used using the format of QME from ISO/IEC 25000

QME	Description
QME category	Number of Data Items
QME name	JVM Memory used
QME ID	JVM M
Detail	JVM Memory used measures the amount of actual memory that is used by the clients' Java process
Input	Performance Measurement Tool
Documentation	JVM M defines the amount of memory used by the clients' Java process during the test session
Measurement Scale	Nominal
Measurement Focus	Internal

Table B7: Definition of Number of Sessions using the format of QME from ISO/IEC 25000

QME	Description
QME category	Number of user operations
QME name	Number of Sessions
QME ID	S
Detail	Number of Sessions measures the amount sessions at the start of a user operation execution
Input	Performance Measurement Tool
Documentation	S defines the number of active *Application A* sessions at the start of a user operation execution
Measurement Scale	Nominal
Measurement Focus	Internal

Table B8: Definition of Client CPU Time using the format of QME from ISO/IEC 25000

QME	Description
QME category	Number of I/O
QME name	Client CPU Time
QME ID	C-CPU
Detail	Client CPU Time measures the number of seconds of overall time spent by the client CPU elaboration
Input	Performance Measurement Tool
Documentation	C-CPU defines the clients CPU elaboration during the user operation execution
Measurement Scale	Nominal
Measurement Focus	Internal

Table B9: Definition of Server CPU Time using the format of QME from ISO/IEC 25000

QME	Description
QME category	Number of I/O
QME name	Server CPU Time
QME ID	S-CPU
Detail	Server CPU Time measures the number of seconds of overall time spent by the server CPU elaboration
Input	Performance Measurement Tool
Documentation	S-CPU defines the servers CPU elaboration during the user operation execution
Measurement Scale	Nominal
Measurement Focus	Internal

Table B10: Definition of SQL Time using the format of QME from ISO/IEC 25000

QME	Description
QME category	Time Duration
QME name	SQL Time
QME ID	SQL-T
Detail	SQL Time measures the number of seconds of overall time spent executing SQL statements
Input	Performance Measurement Tool
Documentation	SQL-T defines the responsiveness of the SQL database
Measurement Scale	Nominal
Measurement Focus	Internal

Table B11: Definition of SQL Calls using the format of QME from ISO/IEC 25000

QME	Description
QME category	Number of Tasks
QME name	SQL Calls
QME ID	SQL-C
Detail	SQL Calls measures the number of SQL calls during an execution of an SQL statement
Input	Performance Measurement Tool
Documentation	SQL-C defines the number SQL calls of *Application A* during an executed SQL statement
Measurement Scale	Nominal
Measurement Focus	Internal

Table B12: Definition of SQL Errors using the format of QME from ISO/IEC 25000

QME	Description
QME category	Number of faults
QME name	SQL Errors
QME ID	SQL-E
Detail	SQL Errors measures the number of SQL errors during an execution of an SQL statement
Input	Performance Measurement Tool
Documentation	SQL-E defines the number SQL errors of *Application A* during an executed SQL statement
Measurement Scale	Nominal
Measurement Focus	Internal

Table B13: Definition of Network Latency (avg) using the format of QME from ISO/IEC 25000

QME	Description
QME category	Number of I/O
QME name	Network Latency (avg)
QME ID	NL (avg)
Detail	Network Latency (avg) measures the average network latency during a user operation execution
Input	Performance Measurement Tool
Documentation	NL (avg) defines the average network latency for *Application A* during a user operation execution
Measurement Scale	Ratio
Measurement Focus	Internal

Table B14: Definition of Percentage of Packets Lost using the format of QME from ISO/IEC 25000

QME	Description
QME category	Number of faults
QME name	Percentage of Packets Lost
QME ID	PPL
Detail	Percentage of Packets Lost measures the number of network packets lost in relation to the total number of network packets sent
Input	Performance Measurement Tool
Documentation	PPL defines the ratio of lost network packets during the execution of a user operation of *Application A*
Measurement Scale	Ratio
Measurement Focus	Internal

Table B15: Definition of Overall Response Time using the format of QME from ISO/IEC 25000

QME	**Description**
QME category	Time Duration
QME name	Overall Response Time
QME ID	ORT
Detail	Overall Response Time measures the cumulated response time of all executed user operation implemented in the performance measurement tool
Input	Performance Measurement Tool
Documentation	ORT defines the cumulated responsiveness to all user operations executed in the performance measurement tool of *Application A*
Measurement Scale	Nominal
Measurement Focus	Internal

C - Validation and Example Use of the Apdex (Operation 2,3,4 & 5)

The *tables C1* through *C8* present the calculated Apdex KPI's for the remaining user operations of *Application A*.

Operation 2 - Execute Query

Thresholds: T = 115 seconds; F = 460 seconds

Table C1: Results Apdex Calculation for Execute Query User Operation Sweden

Gothenburg (Sweden)					
Month	**Total Amount of Response Times (RT)**	**Satisfying: RT < T**	**Tolerating: T < RT < F**	**Frustrating: F < RT**	**Apdex(T)**
Jan	1293	642	621	30	0,74
Feb	1265	742	518	5	0,79
Mar	1397	812	584	1	0,79
Apr	1368	927	441	0	0,84
May	1407	1091	308	8	0,88
Jun	1342	965	373	4	0,86
Total	**8072**	**5179**	**2845**	**48**	**0,82**

Table C2: Results Apdex Calculation for Execute Query User Operation China

Jiading (China)					
Month	**Total Amount of Response Times (RT)**	**Satisfying: RT < T**	**Tolerating: T < RT < F**	**Frustrating: F < RT**	**Apdex(T)**
Jan	1404	341	1063	0	0,62
Feb	1266	431	830	5	0,67
Mar	1397	503	894	0	0,68
Apr	1370	638	732	0	0,73
May	1362	739	623	0	0,77
Jun	1349	725	624	0	0,77
Total	**8148**	**3377**	**4766**	**5**	**0,71**

Operation 3 - Create/Delete Dataset

Thresholds: T = 3 seconds; F = 12 seconds

Table C3: Results Apdex Calculation for Create/Delete Dataset User Operation Sweden

Gothenburg (Sweden)					
Month	Total Amount of Response Times (RT)	Satisfying: RT < T	Tolerating: T < RT < F	Frustrating: F < RT	Apdex(T)
Jan	1293	1102	102	89	0,89
Feb	1265	1194	71	0	0,97
Mar	1397	1324	72	1	0,97
Apr	1368	1278	88	2	0,97
May	1407	1226	116	65	0,91
Jun	1342	1210	109	23	0,94
Total	**8072**	**7334**	**558**	**180**	**0,94**

Table C4: Results Apdex Calculation for Create/Delete Dataset User Operation China

Jiading (China)					
Month	Total Amount of Response Times (RT)	Satisfying: RT < T	Tolerating: T < RT < F	Frustrating: F < RT	Apdex(T)
Jan	1404	0	1194	210	0,43
Feb	1266	3	1157	106	0,46
Mar	1397	2	1288	107	0,46
Apr	1370	2	1213	155	0,44
May	1362	3	1225	134	0,45
Jun	1349	4	1215	130	0,45
Total	**8148**	**14**	**7292**	**842**	**0,45**

Operation 4 - Expand Operation

Thresholds: T = 25 seconds; F = 100 seconds

Table C5: Results Apdex Calculation for Expand User Operation Sweden

Gothenburg (Sweden)					
Month	**Total Amount of Response Times (RT)**	**Satisfying: RT < T**	**Tolerating: T < RT < F**	**Frustrating: F < RT**	**Apdex(T)**
Jan	1293	1156	68	69	0,92
Feb	1265	1065	185	15	0,92
Mar	1397	1364	21	12	0,98
Apr	1368	1342	12	14	0,99
May	1407	1336	52	19	0,97
Jun	1342	1302	33	7	0,98
Total	**8072**	**7565**	**371**	**136**	**0,96**

Table C6: Results Apdex Calculation for Expand User Operation China

Jiading (China)					
Month	**Total Amount of Response Times (RT)**	**Satisfying: RT < T**	**Tolerating: T < RT < F**	**Frustrating: F < RT**	**Apdex(T)**
Jan	1404	0	286	1118	0,10
Feb	1266	0	722	544	0,29
Mar	1397	0	923	474	0,33
Apr	1370	0	560	814	0,20
May	1362	2	460	900	0,17
Jun	1349	0	658	691	0,24
Total	**8148**	**2**	**3609**	**4541**	**0,22**

Operation 5 - Create/Delete Item

Thresholds: T = 3 seconds; F = 12 seconds

Table C7: Results Apdex Calculation for Create/Delete Item User Operation Sweden

Gothenburg (Sweden)					
Month	Total Amount of Response Times (RT)	Satisfying: RT < T	Tolerating: T < RT < F	Frustrating: F < RT	Apdex(T)
Jan	1293	863	416	14	0,83
Feb	1265	438	822	5	0,67
Mar	1397	522	874	1	0,69
Apr	1368	601	767	0	0,72
May	1407	845	561	1	0,80
Jun	1342	663	677	2	0,75
Total	**8072**	**3932**	**4117**	**23**	**0,74**

Table C8: Results Apdex Calculation for Create/Delete Item User Operation China

Jiading (China)					
Month	Total Amount of Response Times (RT)	Satisfying: RT < T	Tolerating: T < RT < F	Frustrating: F < RT	Apdex(T)
Jan	1404	669	694	41	0,72
Feb	1266	0	1177	89	0,46
Mar	1397	0	1339	58	0,48
Apr	1370	0	1306	68	0,48
May	1362	215	1067	80	0,55
Jun	1349	35	1192	122	0,47
Total	**8148**	**919**	**6775**	**458**	**0,53**

D - Quality attributes of the quality model and *Apdex* evaluation

The following tables present the attributes from the quality model developed by Staron et al. (2016), divided into their five categories, that were used to evaluate the Apdex measurement system. Furthermore, the evaluation results from *chapter 5* are provided in the far right column.

Table D1*:* Data Analysis Attributes and Evaluation

Sub-Category	**Attribute**	**Description**	**Evaluation** (fulfills - 1 / does not fulfill - 0)
Information Product	Up-to-date	The KPI value is representing the current status of the entity	0
Information Product	Satisfies the as-sumptions of the information product	The KPI has an underlying business goal or strategy	0
Information Product	Supports fitness of purpose	There KPI fits the stakeholders information need and it can be proved	1
Information Product	Appropriate amount	The KPI contains enough (and not too much) information to satisfy the stakeholders information need	1
Interpretation	Objective	The KPI objectively quantifies the measured entity	1
Interpretation	Interpretable	The KPI can be interpreted in the organization based on the gathered data and the organizational context	0
Indicator	Support	The KPI can satisfy (after adjustment) more than one information need of the organization	1
Indicator	Linked to the stakeholder	The KPI is linked to a specific stakeholder who is responsible for the KPI and can act upon it	1
Indicator	Concise	The presentation of the KPI does not provide irrelevant information	1
Indicator	Consistent	The KPIs presentation is consistent with the goals of the KPI.	1
Analysis model	Complete	All steps of the measurement procedure are performed.	1
Analysis model	Correct	All the steps of the measurement procedure are performed correctly.	1
Analysis model	Reproducible/ repeatable	The measurement procedure is either documented or automated so that it is possible to reproduce the results.	1

Analysis model	Transparent	It is clear how the measurement procedure is performed and how the results are obtained.	1
Analysis model	Privacy protected	The procedure is created in such a way that it protects the privacy of the measured entities.	0
Analysis model	Secure	The procedure is created in such a way that it is protected from unauthorized tampering.	0
Analysis model	Accurate	The procedure truly reflects the measured attributes of the measured entities.	1

Table D2: Data Preparation Attributes and Evaluation

Sub-Category	**Attribute**	**Description**	**Evaluation** (fulfills - 1 / does not fulfill - 0)
Derived measure	Measurement accuracy	The measurement error should be as low as possible	1
Derived measure	Measurement trueness	The value should reflect the true value of the measurand	1
Derived measure	Metrological traceability chain	There is a meta-data describing the technical details of the measure (e.g. its data type)	1
Derived measure	Standard measurement uncertainty	The measurement error of the measure is provided	0
Measurement function	Complete	All steps of the measurement procedure are performed.	1
Measurement function	Correct	All the steps of the measurement procedure are performed correctly.	1
Measurement function	Reproducible/ repeatable	The measurement procedure is either documented or automated so that it is possible to reproduce the results.	1
Measurement function	Transparent	It is clear how the measurement procedure is performed and how the results are obtained.	1
Measurement function	Privacy protected	The procedure is created in such a way that it protects the privacy of the measured entities.	0
Measurement function	Secure	The procedure is created in such a way that it is protected from unauthorized tampering.	0
Measurement function	Accurate	The procedure truly reflects the measured attributes of the measured entities.	1

Table D3: Data Collection Attributes and Evaluation

Sub-Category	**Attribute**	**Description**	**Evaluation** (fulfills - 1 / does not fulfill - 0)
Base measure	Measurement accuracy	The measurement error should be as low as possible	1
Base measure	Measurement trueness	The value should reflect the true value of the measurand	1
Base measure	Metrological traceability chain	There is a meta-data describing the technical details of the measure (e.g. its data type)	1
Base measure	Standard measurement uncertainty	The measurement error of the measure is provided	0
Measurement method	Complete	All steps of the measurement procedure are performed.	1
Measurement method	Correct	All the steps of the measurement procedure are performed correctly.	1
Measurement method	Reproducible/ repeatable	The measurement procedure is either documented or automated so that it is possible to reproduce the results.	1
Measurement method	Transparent	It is clear how the measurement procedure is performed and how the results are obtained.	1
Measurement method	Privacy protected	The procedure is created in such a way that it protects the privacy of the measured entities.	0
Measurement method	Secure	The procedure is created in such a way that it is protected from unauthorized tampering.	0
Measurement method	Accurate	The procedure truly reflects the measured attributes of the measured entities.	1
Measurement method	Calibrated	The measurement method is calibrated to the type of the measurand	1
Measurement method	Easy to use	The measurement procedure is easy to use according to its users.	1
Measurement method	Supported by human resources to collect, analyse and manage the KPI	There is a measurement team supporting the data collection, analysis and management for the KPI	0
Attributes	Feasible to collect in the organization	It is possible to objectively collect the data either manually or automatically.	1
Attributes	Supported by available tools	There are tools which support the measurement procedure (e.g. scripts measuring the size of the software)	0

Table D4: Organizational Reference Context Attributes and Evaluation

Sub-Category	Attribute	Description	Evaluation (fulfills - 1 / does not fulfill - 0)
Reference values for context	Preferred	The stakeholders have the preference for the KPI (i.e. judge it as useful).	1
Reference values for context	Actionable	The KPI allows the stakeholders to take concrete actions (i.e. there is an action plan linked to the KPI).	1
Reference values for context	Reputed	The KPI has a good reputation in the organization to lead to the right decisions and actions.	0
Reference values for context	Applicable	It is clear how the KPI is applicable to one or more stages of the product life cycle.	1
Decision criteria	Believable	The KPI is believable.	1
Decision criteria	Understandable	The KPI is understandable by the stakeholders and by the organization.	0
Decision criteria	Linked to the business strategy	The KPI is linked to a specific business strategy of the organization.	0

Table D4: Standard Reference Model Attributes and Evaluation

Sub-Category	Attribute	Description	Evaluation (fulfills - 1 / does not fulfill - 0)
Generic model of information needs	Provides normative consensus	The KPI should provide the consensus that improving the KPI will lead to improvements in the organization.	0
Generic model of inf. needs	Valued by the stakeholders	The KPI is perceived to bring value to the stakeholders and their information needs.	1
Generic model of inf. needs	Known to the organization	The KPI is known (disseminated) in the organization.	0
Generic model of inf. needs	Relevant	The KPI is relevant to the prioritized information needs	1

Generic model of inf. needs	Predictive	The KPI is valid from the empirical perspective and can be used for making predictions.	0
Formal information	Traceable	The results obtainable by using the measurement procedure are able to be traced back to the sources of the data.	1