Investigating the Applicability of Agility Assessment Surveys: A Case Study

Samireh Jalali*, **Claes Wohlin***, **Lefteris Angelis****

*Blekinge Institute of Technology, SE- 371 79 Karlskrona, Sweden

**Aristotle University of Thessaloniki, Department of Informatics, 54124 Thessaloniki, Greece

Corresponding Author: samireh.jalali@bth.se

Abstract

Context: Agile software development has become popular in the past decade despite that it is not a particularly well-defined concept. The general principles in the Agile Manifesto can be instantiated in many different ways, and hence the perception of Agility may differ quite a lot. This has resulted in several conceptual frameworks being presented in the research literature to evaluate the level of Agility. However, the evidence of actual use in practice of these frameworks is limited.
Objective: The objective in this paper is to identify online surveys that can be used to evaluate the level of Agility in practice, and to evaluate the surveys in an industrial setting.
Method: Surveys for evaluating Agility were identified by systematically searching the web. Based on an exploration of the surveys found, two surveys were identified as most promising for our objective. The two surveys selected were evaluated in a case study with three Agile teams in a software consultancy company. The case study included a self-assessment of the Agility level by using the two surveys, interviews with the Scrum master and a team representative, interviews with the customers of the teams and a focus group meeting for each team.
Results: The perception of team Agility was judged by each of the teams and their respective customer, and the outcome was compared with the results from the two surveys. Agility profiles were created based on the surveys.
Conclusions: It is concluded that different surveys may very well judge Agility differently, which support the viewpoint that it is not a well-defined concept. The researchers and practitioners agreed that one of the surveys, at least in this specific case, provided a better and more holistic assessment of the Agility of the teams in the case study.

Keywords Agile, Assessment, Measurement, Tool, Survey.

Acronyms and Abbreviations

GSD: **G**lobal **S**oftware **D**evelopment
CI: **C**ontinuous **I**ntegration
XP: e**X**treme **P**rogramming
RQ: **R**esearch **Q**uestion
S2: **Survey 2**
T2: **T**eam **2**
TDD: **T**est **D**riven **D**evelopment
PP: **P**air **P**rogramming
TR: **T**rouble **R**eport
S1: **S**urvey **1**
T1: **T**eam **1**
T1: **T**eam **3**

1. Introduction

Agile values and principles have emerged as a general approach to developing software. The formulation of Agile principles and practices has been driven from a practitioner's point of view (Beck et al., 2001). Agile software development has received more attention in recent years and different Agile methods have been utilized in software organizations. Each method consists of several practices, however, the extent to which practitioners utilize the practices is not yet fully investigated.

This has motivated researchers to study what practitioners exactly mean when claiming to be Agile through examining their level of adhering to Agile values, principles, and technical practices. Therefore, several frameworks with the purpose of assessing or profiling Agility have been developed e.g. (Sidky 2007; Hossain et al., 2009; Taromirad and Ramsin, 2009; Soundararajan and Arthur, 2011). Although some of these frameworks are developed through empirical studies, to the best of our knowledge, only the participating organizations have used them afterwards. The major reason could be that practitioners view them as context specific so that they cannot apply them without the help of Agile experts.

On the other hand, several questionnaires e.g. (James, 2007; Waters, 2008; Mayberg Consulting, 2008; ThoughtWorks Studio, 2010a; ThoughtWorks Studio, 2010b; Sutherland, 2009) have been developed and used by practitioners that have not been investigated by researchers. This indicates that researchers and practitioners seem to evolve the tenets of Agile software development in parallel to some extent. The objective in this paper is to bring research and practice closer by evaluating some of the questionnaires for assessing Agility in industry.

A number of online tools or surveys (henceforth referred to as surveys) for Agility assessment are identified and evaluated. The surveys identified are first evaluated using a set of criteria after which two surveys are selected for further studies. These two surveys are evaluated in an industrial case with three teams working with two different customers. Thus, the article contributes with an industrial evaluation of surveys for assessing Agility. The results and experiences should be helpful in increasing the understanding of Agility for both researchers and practitioners alike. More generally, Agility assessment forms an important step towards formulating ways of improving software development using Agile methods. It would help organizations identify areas for improvement to become better in practicing Agility and it could also be helpful in identifying specific practices that could be better practiced. Based on the findings in the case study, a comparison of the two selected surveys are provided to enhance the understanding of the differences between them.

The remainder of the paper is structured as follows. Section 2 summarizes related work. Section 3 discusses the research methodology and introduces the different steps taken in the research. The results are discussed from two different perspectives in Sections 4 and 5 respectively. The results from using the surveys are presented in Section 4, and a comparison of the two surveys is provided in Section 5. Threats to validity are discussed in Section 6. Finally, conclusions and future research directions are presented in Section 7.

2. Related Work

The concept of Agility as "flexibility" and "leanness" (Conboy and Fitzgerald, 2004) in software engineering was introduced by practitioners (Beck et al., 2001) to mitigate the limitations of traditional software development approaches such as perceived heavy documentation, long time to market, etc. Agility is also defined to be continuously receiving feedback and making changes in the software rather than rejecting higher rates of change (Williams and Cockburn, 2003).

Different approaches to Agility exist; the most common ones are Extreme Programming (XP), Scrum, Feature Driven Development, Dynamic Systems Development Method, and Lean development (Bose, 2008; Dybå and Dingsøyr, 2008). Most Agile methods, however, encourage frequent and continuous face-to-face communication, customer feedback and requirements gathering, as well as Pair Programming (PP), refactoring, Continuous Integration (CI) and minimal documentation (Bose, 2008).

Despite the popularity of Agile methods, there is a lack of understanding to what extent different practices need to be used to become more efficient in software development, and which practices are actually applied in Agile or Lean software organizations or teams.

To the best of our knowledge, the majority of the available research aims at assisting software organization in the process of Agile adoption by assessing, measuring, or evaluating Agility, rather than building a mutual understanding between academia and industry about Agile methodologies and practices. The relevant research is briefly summarized below.

We have grouped the related work based on their outcome, which are: 1) methods and guidelines and 2) tools and frameworks. At the end of Section 2.2, a brief summary of the existing research is provided to put our contribution into the context of the related work.

2.1. Methods and Guidelines for Evaluating Agility

Agile values and principles depict the fundamental characteristics of Agile methodologies and can be used as evaluation criteria. Conboy and Fitzgerald introduced flexibility and leanness as the essential properties of any Agile software development method thorough reviewing Agility across many disciplines (Conboy and Fitzgerald, 2004). They have concluded that the Agile Manifesto principles do not provide practical understanding of the concept of Agility outside the field of software development.

Furthermore, Boehm and Turner identified project size, project criticality, dynamism of the environment, personnel, and culture as crucial variables in Agile methods (Boehm and Turner, 2004). Their proposed five-step risk-based software development method combines the characteristics of both Agile and plan-driven approaches. Although it is a helpful guide for academics, it may not provide the necessary details for practitioners.

Other researchers evaluated Agile methods in relation to other criteria than Agile values and principles. Abrahamsson et al. introduced a structure in which the process, roles and responsibilities, practices, status of adoption, experiences, scope of use, and current research regarding each method are identified (Abrahamsson et al., 2002). They have used this structure as their analytical foundation to analyze the differences and similarities between different Agile software development methods.

McCaffery et al. proposed an assessment method that integrates the structuredness of the plan-driven software process improvement models i.e. CMMI (Software Engineering Institute, 2011) and Automotive SPICE (International Organization for Standardization, 2012) with the flexibleness of Agile practices (McCaffery et al., 2004). However, the assessment is designed specifically for small-to-medium sized automotive software supplier organizations.

Germain and Robillard provided an empirical comparison between an engineering-based process (Unified Process for Education – UPEDU) and an Agile process built around XP principles, mainly through comparing and analyzing the work and time spent in each of their “cognitive” activities (Germain and Robillard, 2005).

In the latest work, Yauch constructed an objective metric for Agility performance performs work that measures Agility as a performance outcome and captures both organizational success and environmental turbulence (Yauch, 2011). The metric was developed as a theoretical model and then validated through literature review, case studies, and pilot survey data.

In addition to the above-mentioned methods and guidelines, several Agility evaluation tools and frameworks exist, which are summarized in the following section.

2.2. Tools and Frameworks for Evaluating Agility

In 2003, Abrahamsson et al. proposed an analytical framework for the analysis of existing Agile methods (Abrahamsson et al., 2003). They have used software development notions such as lifecycle coverage (including the process), project management, abstract principles vs. concrete guidance, universally predefined vs. situation appropriate, and empirical evidence as “analytical lenses”.

Later, Williams et al. provided a benchmark measurement framework for assessing XP practices adopted in software organizations (Williams et al, 2003) which is composed of three parts: 1) XP Context Factors to record essential contextual information, 2) XP Adherence Metrics to concretely and comparatively express the utilized practices, 3) XP Outcome Measures to assess the team’s outcome when using a full or partial set of XP practices.

In 2006, Hartmann and Dymond collected some of the current thinking on appropriate Agile metrics, and proposed simple tools to be used by teams or organizations (Hartmann and Dymond, 2006) with the purpose of encouraging metrics that are more in alignment with the objectives of Agile teamwork.

Simultaneously, Qumer and Henderson-Sellers proposed 4-DAT, which is a framework-based assessment tool for the analysis and comparison of Agile methods (Qumer and Henderson-Sellers, 2006). It provides evaluation criteria for the detailed assessment of Agile software development methods through defining four dimensions: 1) method scope characterization, 2) Agility characterization, 3) Agile values characterization, and 4) software process characterization. 4-DAT provides a mechanism for quantitatively measuring the degree of Agility of any method at a specific level in a process using specific practices.

In 2007, Sidky proposed Sidky Agile Measurement Index (SAMI), which is based on the concepts and principles of the existing software process improvement standards CMM (Paulk, 1993) and CMMI (Software Engineering Institute, 2011) (Sidky, 2007). SAMI determines five Agile levels by measuring the number of Agile practices adopted by an organization. The objectives for each level are set relatively to Agile values and principles as stated in the Agile Manifesto, and a set of practices is suggested for each level. However, enforcing these predefined practices is not in alignment with the flexibility of Agile methods.

Later in 2009, Hossain et al. introduced a conceptual framework based on the research literature to address the challenges of combining Global Software Development (GSD) and Agile methods (Hossain et al., 2009). It is expected to assist project managers in deciding which Agile strategies are effective for a particular GSD setting considering the contextual information. Furthermore, Taromirad and Ramsin introduced the Comprehensive Evaluation Framework for Agile Methodologies in (Taromirad and Ramsin, 2009), aiming at covering all different aspects of Agile methodologies.

Lastly, Soundararajan and Arthur proposed the Objectives, Principles and Practices framework, in which the "goodness" of an Agile method is assessed by evaluating its adequacy and effectiveness, and the capability of the organization to provide the supporting environment (Soundararajan and Arthur, 2009). It means that for a given set of objectives for an Agile method, it has to be ensured that the appropriate principles and the proper practices are set.

According to the provided summary of the related work, the majority of the existing research does not adequately address all Agility issues in their criteria, and rarely considers the usage context, which is the main concern of project managers. Furthermore, it does not suggest any quantitative metrics that are easily applicable in practice. Therefore, our study aimed at examining the Agility assessment surveys that are proposed by practitioners.

3. Research Methodology

3.1. Research Questions

The major purpose of the research was to investigate the applicability of the existing surveys for assessing the Agility of a team or an organization. The research questions (RQ) were formulated as follows.

RQ1: Which surveys exist to evaluate the Agility of a team or an organization and what are their strengths and weaknesses?

RQ2: Are the existing surveys applicable to assess Agility?

RQ3: Do the existing surveys give the same assessment results?

3.2. Research Steps

The research methodology consists of three steps: preparation, case study and statistical analysis. The core step is the case study, i.e. the second step. The first step provides the basis for conducting the case study and the third step involves the statistical analysis of the data collected in the case study. Each of them is presented in more detail in the following subsections.

1) **Preparation: Survey identification**
The research started by a systematic search of the relevant literature and the web for current frameworks, or surveys with the purpose of evaluating the Agility of a software development team or an organization. Since, we did not find the existing frameworks sufficiently testable (e.g. through a questionnaire that could be used in a case study), we ended up focusing the searches on

the surveys that are developed and used by practitioners. Thus, to answer RQ1, we searched the web and preliminary evaluated the surveys and selected a few of them.

2) **Case study in industry**

The surveys identified in the web search were input to a case study to answer RQ2. The objective was to evaluate the applicability of the surveys. The applicability in this context means to what extent it represents the actual status of the studied teams, and is measured by comparing the results given by the surveys with the participants' perceptions and expectations as well as with their customers (i.e. the product owners).

3) **Statistical analysis**

The data collected from the three teams in the case study were analysed using statistical tests. Furthermore, RQ3 was answered by comparing the results of the surveys examined in the case study. This was also evaluated using statistical tests.

In summary, the research methodology was a mixed approach (Creswell, 2003) applying both qualitative and quantitative data collection and analysis methods in an industrial case study. The mixed approach includes using both interviews and focus groups (qualitative) and a survey (quantitative). The classification of the study as a case study is based on the guidelines for case study research in software engineering (Runeson and Höst, 2009).

3.3. Preparation: Survey Identification

The preparation phase consisted of finding existing surveys and preliminary evaluating them in order to minimize the time required from industrial participants.

3.3.1. Searching the Web

To address RQ1, we have searched the web for existing tools or surveys that are not found in the research literature. Here a tool is viewed as a set of questions (e.g. a survey or questionnaire) plus an analysis on the responses to the questions. The search was conducted using the Google[1] search engine. The intention is to use the following search string: Agile AND (assess OR measure OR framework OR tool OR evaluate OR profile). If this string results in too many items, the objective is to try simpler search strings to be able to handle the amount of information.

3.3.2. Initial Evaluations of the Surveys

The surveys found in the searches are to be evaluated. The objective is to only select a few surveys as input to an industrial case study. The main reason being that it would be too time consuming for the industrial partner to evaluate all surveys potentially available. The following criteria were decided to form the basis for identifying the surveys that should go into the industrial case study:

- *Covering all Agile methods*: We were looking for a survey that assesses Agility in general rather than being focused on one or a combination of Agile methods, e.g. Scrum or XP.
- *Being free of charge*: Since the scope was wide (i.e. all surveys that assess Agility), and the participating teams could potentially be from different companies (as initially planned), we decided to continue with the surveys, which are available for free. In addition, companies normally prefer cost-efficient solutions. It becomes even more expensive if a company would like to repeat the assessment in different points of time and has to pay for the license each time.
  However, we also examined the ones that provide trial version. In addition, they were further examined based on their questions and the covered areas regarding the Agility. Another important factor was the presentation of the results. Therefore, we gave random answers to all questions of available surveys and explored the results in order to figure out how they are extracted.
- *Assessing a team and an organization*: Agility can be measured at different levels i.e. individual, team, department, division, organization, etc. We were interested in the surveys that assess a software development team and its software organization.
- *Having a reasonable number of questions*: Each survey should cover different Agility areas (e.g. development and management practices). On the one hand, too many questions may be hard to have people answering, but on the other hand having too few questions would make it hard to evaluate the actual Agility in an organization.

[1] http://www.google.com

- *Providing a sufficient analysis on the given responses*: The fundamental part of all assessments is the analysis and presentation of the given data. Therefore, we omitted the surveys that did not provide sufficient analysis or the presentation was not clear.
- *Simplicity*: An assessment survey must be easy to use, both in entering the data and in interpreting the results. Most small to medium size organizations prefer surveys that can be utilized with the internal expertise and as often as they need to.

3.4. Case Study in Industry

Using the selection of surveys, we planned a case study to evaluate their applicability through comparing their results with the perceptions/expectations of the participating teams and their customers.

3.4.1. Participants

In the following subsections, we describe the participating organization and the teams taking part in the case study as well as their customers. The contextual information is reported according to the guidelines recommended by Petersen and Wohlin (Petersen and Wohlin, 2009).

3.4.1.1. Softhouse Consulting

Softhouse Consulting was founded in 1996 as an independent IT consultancy company in Sweden, and currently is one of the leading Scandinavian suppliers of Lean Software Development with over 100 employees[2]. It has sites located in Malmö, Karlskrona, Växjö, Stockholm and Gothenburg.

The company encourages simplicity, reliability, and professionalism in its work process and employs cross-functional and self-organizing teams to deliver fully functional software with the desired quality.

This study was conducted in cooperation with the development sites located in Karlskrona and Malmö. Three teams participated in the surveys. Two of the teams studied were located in Karlskrona and the third team in Malmö. The teams in Karlskrona work with the same customer while the team in Malmö has a different customer. More detailed information about the teams is given in the following subsections.

3.4.1.2. Telecom Development Team (Team 1)

The team is responsible to add new features to an already existing system.

Product: The work is contract-based and the business relationship is onshore outsourcing (Prikladnicki et al., 2007). The final product is commercial and is customized for specific telecom operator(s). At the time of data collection, the collaboration was built over 2.5 years through different projects, and each project is about 500 person-hours. The programming languages are C++ and Java.

Processes: If the end customers are not satisfied with the standard features, they can order new features. The product owner prioritizes the orders applying first-in-first-out strategy. A project to add a new feature is normally about 200-400 hours.

Requirements (i.e. features) are small and therefore a number of small projects are run simultaneously within the team. One project is normally accomplished within four weeks (one sprint). The team receives a fixed set of requirements and a fixed deadline, and then independently performs Sprint planning and the design. However, the team must handshake the design with the customer's architect team to ensure its consistency with the whole product.

Although the team applies Scrum practices, the work process is adapted to the customer's processes. For example, the practices within the team are Scrum, but the reports are made according to the customer's preferences. The team follows coding conventions (which are accessible in a wiki). In addition, it applies automated testing and Test Driven Development (TDD). Daily meetings are held within the team to report the progress and discuss the issues, and weekly meetings with all personnel.

[2] http://www.softhouse.se/en/index.php/about-us/about-softhouse

Practices, Tools and Techniques: Since the customer does not provide any Agile specific tools, Excel sheets are used for Sprint planning (e.g. to visualize the task loads, reprioritize the tasks, and provide burn down graphs to be sent to the customer as weekly reports).

People: One functional team shall consist of four to five people according to the customer's demands. However, eight people are in Team 1 (T1) so that they can form two teams at the same time working on different tasks if needed. One person out of eight takes the role of system analyst, one is Scrum master, and six are developers/testers. The average prior experience of Agile (before this project) is about one year excluding the Scrum master and 1.2 years including the Scrum master.

Organization: The team is co-located at the customer site in the same city, but works independently. The team uses informal and face-to-face communication with no bookings.

Market: The project is developed for only one customer (although the customer would sell it to more than one other customer), and the studied organization's strategy is to reduce the delivery time with maintained good quality.

Customer: The customer organization is globally distributed and its software development teams have adopted customized Agile.

3.4.1.3. Telecom Maintenance Team (Team 2)

Team 2 (T2) was formed 0.5 year after T1 to work with reported defects and maintenance issues (the same product as for T1). Most of the contextual information is the same as for T1 and therefore we describe only the differences.

Processes: Any types of problems with the product found by the end users, market, or test departments, are documented as "trouble report" (TR). T2 receives TRs and runs trouble shooting i.e. recreating the failure, finding where to fix it in the code, writing the test case, preparing the follow up forms, and sending the follow up to the test team.

The team has the freedom to plan only the tasks of a day, because the customer prioritizes the tasks, which TRs should be done first. So, the operation is somehow sprint-less. Regular meetings (three times per week) are held with the customer to collect the TRs and discuss the progress.

Practices, Tools and Techniques: A customized Kanban board is used which shows only backlog, team, and done columns to represent the plan and the workflow. Since the work is not really planned in Sprints, retrospective meetings have not been held regularly.

People: T2 consists of 10 people in total, two system analysts, one Scrum master, and seven developers/testers. The average prior experience of Agile is about 0.5 year excluding the Scrum master and 0.75 years including the Scrum master.

Customer: Although the customer organization is the same as for T1, the customer unit is different.

3.4.1.4. Building Management Development Team (Team 3)

Product: The business relationship is onshore outsourcing, and the product is a building management system, which is supervising and controlling technical systems in buildings. For example, it is used for monitoring heating/cooling, ventilation, fire prevention alarm systems, light control, etc. At the time of data collection, the collaboration was built over four years.

Processes: The product owner prioritizes the tasks for the development team. Team 3 (T3) performs Sprint planning and the design themselves. The team applies Scrum practices with some adaptations to the customer's processes. Daily meetings are held within the team to report the progress and discuss the issues, together with all personnel.

Practices, Tools and Techniques: The team applies only practices that are needed, but TDD is not used since testing is done externally by a test department at the customer site.

People: The team consists of six developers, one Scrum master, one product owner, one project leader (who works as an interface towards the customer organization), and one technical publication person (who writes how the system works). It accomplishes four to five user stories in a two-week Sprint. Tasks are done in parallel, and PP seems to increase the quality of the output (the number of defects seems to be lower). The average prior experience of Agile (before this project) is about 4 months including the Scrum master and the team leader, and is 6 months when excluding them.

Organization: The team is co-located at the customer site in the same city, but works independently. The team uses informal and face-to-face communication.

Market: The project is developed for only one customer (although the customer would sell it to more than one other customer).

Customer: The customer organization is a multicultural company and is globally distributed.

3.4.2. Design and Conduct of Case Study

In this section, we explain the details of the research design and conduct in relation to the case study, including the questionnaire, interviews, and open discussions.

3.4.2.1. Questionnaire

All questions from the selected surveys were to be merged into one Excel file (in separate sheets). In addition, some questions were added to capture the demographic information of the participants (e.g. role and years of experience) as well as project information (e.g. type and customer).

For each question in the surveys, we added the option to skip answering if it was not applicable or the respondent did not want to reply to it for any reasons. Furthermore, participants were supposed to declare how sure they were of the given answer for each question. The reason for this was to weigh the answers with more confidence level higher than the answers with low confidence.

In addition, classification of the questions as it was in the original surveys was removed in order to reduce the bias when answering the questions although hiding the classification might have caused confusions when answering the questions. A sample of the survey is presented in Appendix 1.

The detailed planning of the case study was done when it was known that two surveys were selected from the survey identification step. The two chosen surveys are denoted S1 and S2. At the time we designed the questionnaires for each survey, S2 consisted of 127 questions excluding the demographic questions and S1 had 20 questions. Hence, in total, the study includes 147 questions. Therefore, the time to finish the questionnaire was estimated to be maximum 150 minutes. All participants were informed about the purpose of the study as well as its potential benefits in advance.

3.4.2.2. Interview with Scrum Master and a Team Representative

The major reasons for conducting interviews with the Scrum master and a team representative (separate interviews) were to discuss the any inconsistencies observed in the collected answers, and to gather additional contextual information of the organization, team, project, and the customer.

For this purpose, shortly after data collection, the objective was that two of the researchers should scan the answers separately and find inconsistencies and prioritize them before going into the interviews.

The candidate questions to bring up at the interviews were identified as follows: 1) with no majority (if at least half of the participants did not give the same response), 2) with a big difference in the answers (e.g. "strong yes" vs. "strong no" for the same question), 3) questions that could dramatically change the score in specific areas, 4) with a big agreement, and 5) with the majority of "Not Applicable" response.

They were prioritized in the same order too (i.e. the ones "with no majority response" were more important to be discussed and so on). The main motivation for prioritization was the time limit for the interview, which was agreed to be maximum 90 minutes. Since we had to collect additional contextual information, it was impossible to discuss important identified inconsistencies during the interview if they were not prioritized.

The team representative was to be nominated by the Scrum master and the company's director. The same items were to be discussed with the Scrum master and the team representative intending to openly discuss the reasons causing the inconsistent responses. If both interviewees had given the same or similar response, that was considered as the team's response. Otherwise, we considered the question as "Not Applicable".

3.4.2.3. Interview with Customer (i.e. Product Owner)

In order to collect the customers' perception on the team Agility, we designed a 60-minute interview with their customer representatives (separate customer interviews for T1, T2, and T3). The purpose of the interviews was explained beforehand to the interviewees in an email.

Since the customers were not expected to be aware of all details about the teams in particular the technical aspects, we designed the interviews in a way to generally discuss the following items, which are similar to the categorizations made by the studied surveys:

Teamwork: The perception of the Agility of the team's composition, management and communication.

Requirements: This area includes the level of details of requirements, accommodating changes, technical design of requirements, and collaborating with the product owner.

Planning: It concerns the suitability of planning activities e.g. planning time, levels of planning and progress tracking.

Technical Practices: Technical practices are TDD, PP, refactoring, CI, coding standards, and collective code ownership.

Quality: This relates to automated unit testing, customer acceptance tests, and timing.

Culture: To what extent the customer views the team's management style, response to stress, and the customer involvement as Agile.

Knowledge Creating: If the team learning is evident to the customer and is useful.

In addition, each interviewee was asked to assign a score from 0 to 5 (not Agile to highly Agile) to each before mentioned area. A sample of interview questions is presented in Appendix 2.

3.4.2.4. Focus Group

The results were from the survey and interviews were to be presented to the whole teams in a retrospective meeting, where team members openly could discuss if it matched their own perceptions. The objective was to involve the teams actively in discussing the reasons for not being Agile in some areas, and how it can be improved or why it is an external factor that cannot be changed.

After presenting the results to the teams and discussing how it matched their own perceptions of practicing Agile as well as the reasons for not being Agile in relation to specific aspects, we conducted additional interviews (in addition to the interviews described in Section 3.4.2.2) with two Scrum masters (one for T1 and T2, and one for T3). The main intention of the interview was to plan for an in-depth discussion around the applicability of the studied surveys as well as the potential combination of them. The applicability in this context means to what extent it represents the actual status of the studied teams.

3.4.3. Statistical Analysis

Statistical analysis was to be performed to objectively examine the surveys' applicability. Finally, the plan is to compare the results given by the surveys with the results of the statistical analysis, and to propose an applicable survey based on this comparison. Statistical analysis was conducted to provide answers to the posed research questions.

RQ1: Which surveys exist to evaluate the Agility of a team or an organization and what are their strengths and weaknesses?

Descriptive analysis was to be utilized to provide answer to RQ1 by summarizing and classifying the surveys that were found in the searches.

RQ2: Are the existing surveys applicable to assess Agility?

Applicability was defined as to what extent a survey represents the actual status of a studied team, and it was measured by performing comparative analysis (Marascuilo and Serlin, 1998) i.e. comparing the results given by the surveys with the participants' perceptions/expectations as well as with their customers' view.

RQ3: Do the existing surveys give the same assessment results?
This question was to be answered through conducting both a comparative analysis and statistical tests (Marascuilo and Serlin, 1998).

Firstly, we apply fuzzy approval voting (Lapresta and Panero, 2002; Kangas et al, 2006) to calculate team representative answers for all questions as input for the studied surveys to get the teams' Agility scores, and compare the scores provided by the two studied surveys. Secondly, we enter the data of each participating individual in the studied surveys to calculate their scores, and apply statistical tests on the individual scores to examine differences and similarities between the teams as well as between the surveys. The findings of statistical tests are then to be compared with the results given by the surveys. Both approaches are elaborated in the following subsections.

3.4.3.1. First Approach: Comparing Teams' Agility Ranks Provided by the Studied Surveys

This part of the data analysis consist of: calculating the mode for each question of the survey to represent each team's answer, calculating the level of Agility based on the customer interview, and discussing each team's Agility according to the studied surveys.

Participants could determine how confident they are about the given answer for each question by selecting an option among: "sure", "more sure than unsure", "neither sure not unsure", "more unsure than sure", and "unsure" (i.e. to weight the answer between 1 to 0). In addition, it was possible to write comments.

Not all surveys consider the certainty of respondents about their given answers. Therefore, the same person can answer differently in a replication of the study if he/she is not certain about his/her given answer. We considered this as a threat to the reliability of all surveys and addressed it by adding an option of self-confidence as described above.

Immediately after data collection, we check whether the participants had provided all required information. Otherwise, we plan to contact them for clarification or to complete the questions not answered.

We use a special case of fuzzy approval voting (Lapresta and Panero, 2002; Kangas et al, 2006) in which a person was allowed to choose only one alternative answer among a given set of answers for each question. In the following, it is elaborated how we calculate each team's representative answer.

We sum the confidence levels for each specific option/answer given for a question. The answer with the highest sum of confidence (mode value) is then considered as each team's answer. It is explained in the following example.

Suppose six team members have answered question 1 as it is shown in Table 1. Option 1 has 0.75+0.5+0 confidence level (persons: 1, 3 and 6), option 2 with 1.0 (person: 2), and option 3 with 0.5+0.25 (persons: 4 and 5) as it is shown in Figure 1. After weighting the answer with the level of confidence, answer number 1 becomes the team's answer, since the total confidence level is 1.25 compared to 1.0 and 0.75 respectively for the other options.

Table 1. Sample of Given Answers

Person 1		Person 2		Person 3		Person 4		Person 5		Person 6	
Sure	Answer	Sure	Answer	Sure	Answer	Sure	Answer	Sure	Answer	Sure	Answer
0.75	1	1.0	2	0.5	1	0.5	3	0.25	3	0	1

For some questions, it may be possible that a mode value is not identified because either more than one answer/option had the highest weight or less than the majority of the team members (i.e. (team size/2)+1) had given an identical answer regardless of its weight. Those questions were to be brought to the discussions with the Scrum master and each team's representative separately. If both gave similar explanation and selected the same answer, we considered that as the team's answer, otherwise it was set to "Not Applicable". Then, all answers were entered into the studied online surveys and the results were presented and discussed with the team. The actual analysis is elaborated in more details in Section 4.

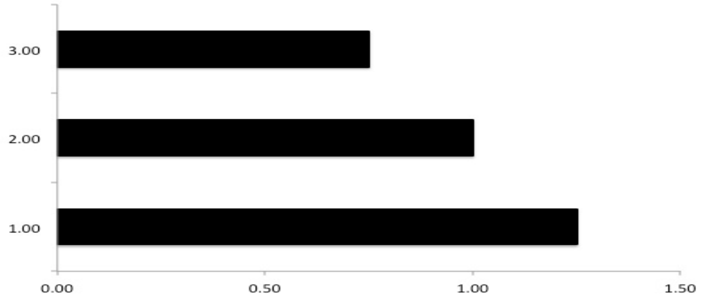

Figure 1. Example of Calculating Mode Value

In the meeting with the product owners, they are asked to give a score from 0 to 5 (not Agile to highly Agile) to each Agile area as discussed in Section 3.4.2.3. Each area is weighed the same (i.e. all are equally important) and the mean value is calculated to represent each team's Agility level from the customer's perspective. If the customer did not have enough information of a specific area, it is removed from the calculations. The results are elaborated in Section 4.

3.4.3.2. Second Approach: Comparing Teams and Comparing Surveys using Statistical Tests

The results given by the studied surveys are complemented by additional statistical tests. To be able to perform the calculations and relevant tests, we need to calculate the individual Agility ranks first. Therefore, the respondents' answers were to be put into the surveys and the rank provided by the surveys was considered as her/his Agility rank.

It is anticipated that data points will be missing, so we plan to apply the nearest neighbor approach for data imputation (Chen and Shao, 2000) before entering data into the surveys, which is elaborated in the following.

Suppose person X has a missing data point (i.e. an answer) for question A. We calculate the similarity between person X and other team members regarding the given answers (no missing, no imputed) for all other questions. The similarity value for all participants in the team was set to zero at the beginning, and we will go through all questions (without missing data and without imputed data) and follow the procedure below:

- For each other question:
    - If person Y had selected the same option as person X, then the similarity value was increased by 0.5 point
    - If the certainty value (see Table 1) was also the same, the similarity value was increased by another 0.5 point
- The person with the greatest similarity value was considered as the nearest neighbor for person X
    - If there were more than one nearest neighbors, we chose one person randomly, using the random function in MS Excel[3]
- Both the selected answer and its certainty value (see Table 1) for question A from the nearest neighbor were imputed for person X

3.5. Summary

The major objective of the research was to investigate the applicability of the existing surveys for assessing/profiling Agility. Therefore, we studied several surveys to select a few of them as input to a case study. The case study was the participating organization and the three units of analysis were the three participating teams (Runeson and Höst, 2009).

[3] RAND()*(b-a)+a: to generate a random real number between a and b

The participants of the case study were: 1) three software development teams in a consultancy company in Sweden that answered the questions of the selected surveys, 2) the Scrum master of the teams to clarify the answers of questions that required discussions, 3) representatives of each team with the same purpose as meeting the Scrum masters, and 4) customers of all teams to collect their perception of each team's Agility.

The extracted results of surveys on the level of practicing Agility by teams were compared against the teams' own perceptions of Agility, and their customer's perception. The results of the investigations are presented in the following section.

4. Results

The results from using the two surveys are presented based on the three research steps presented in Section 3.2. Section 4.1 presents the selection of two surveys starting from the search for surveys on the web and the evaluation of them based on a set of criteria. In Section 4.2, first imputation is conducted and then the case study is presented where the results from using the two selected surveys are discussed and compared. Furthermore, the results from the surveys are compared with the customer's perception of the Agility of the teams, and the teams also reflect on the results. Section 4.3 presents the statistical analysis, where the results from the point of view of the teams are compared. A detailed comparison of the two surveys is provided in Section 5.

4.1. Survey Identification

4.1.1. Surveys from Searching the Web

The use of the following search string: Agile AND (assess OR measure OR framework OR tool OR evaluate OR profile) resulted (as feared) in an enormous number of items (i.e. about 323000). Therefore, simpler search strings were tried such as "Agile assessment tool" and "Agile measurement survey". However, the search string that gave the most reasonable results was "Agility assessment tool". Thus, the surveys found and evaluated were found using these search strings.

Some commonly used checklists by Agile practitioners seems to be (1) the Nokia Test (Sutherland, 2009) for Scrum, (2) How Agile Are You: 42-Point Test (Waters, 2008), (3) the Scrum Master Checklist (James, 2007), and (4) the Do It Yourself (DIY) Project Process Evaluation Kit (Dinwiddie, 2009). It should be noted that most of these checklists are tailored to one or a few specific Agile methods.

Another concern for software organizations is to figure out how efficient the Agile adoption has been, which implies identifying problem areas and taking proper actions to address them. For this purpose, retrospective meetings at the end of each iteration are held to "fine-tune" the Agile practices. Furthermore, external consultants or surveys can also help the assessment. Most assessment surveys are, however, not publicly available free of charge, and therefore, information about them is also hard to find. A summary of Agile assessment surveys found in our searches is presented in Table 2 and Table 3.

Table 2 - Summary of Available Surveys 1

Test	Assessment unit				Score		Free	Agile method		Focus		
	Team	Department	Division	Organization	Comparative	Absolute		Generic Agile	Only	Agile principles	Generic Agile practices	Specific Agile practices
*Agile Karlskrona Test (*Mayberg Consulting, 2008*)*	√					√	√	√		√		
CMI Lean Agility (CMI Lean Agility, 2009)				√		√		√				
Comparative Agility Assessment (Cohn and Rubin, 2007)	√	√	√	√	√[4]		√	√			√	
Dr Agile (Santeon Group, 2012)				√		√		√				
Nokia Test (Sutherland, 2009)	√					√	√		Scrum			√
Scrum Checklist (Kniberg, 2010)	√					√	√		Scrum			√
Signet Consulting (Signet Research and Consulting, 2005)	√	√	√	√		√	√	√			√	
Thoughtworks Agile Self Evaluation (Thoughtworks Studio, 2010a)	√			√		√	√	√		√	√	
Thoughtworks Build And Release Management (Thoughtworks Studio, 2010b)	√					√	√	√				√
42-Point Test (Waters, 2008)	√					√	√	√		√		

4.1.2. Initial Evaluation of Surveys

Among the presented surveys above (see Table 2 and Table 3), two Surveys i.e. (Thoughtworks Studio, 2010a) and (Cohn & Rubin, 2007) were selected for further evaluation after studying and discussing them with the practitioners. The basis for the selection was the criteria presented in Section 3.3.2. The chosen surveys were judged having a good coverage of different perspectives of Agility (e.g. team and organization) as well as different areas (e.g. requirements, testing and communication). Thus, the two surveys remaining for further investigation in the case study were:

1) **Thoughtworks' Agile Self Evaluation (S1)**: It is developed by Thoughtworks, which is a leading Agile development/consulting organization (Thoughtworks Studio, 2010a). The survey consists of 20 multiple-choice questions addressing Agile development and management practices. Management practices are related to requirements analysis, business responsiveness, collaboration and communication, project management, and governance. Development practices address simplicity, build management, configuration management, and testing and quality assurance. Practitioners can complete the surveys online and get a report on the level of Agility within their organization/team, as well as improvement opportunities (Thoughtworks Studio, 2010a).

   Although the questions are related to different Agile areas and perspectives, the number of questions per area is not sufficient for all categories. However, Agility and the highest level of practices are clearly defined and are transparent to the users. In addition, the results are presented visually along with relevant improvement recommendations for each Agile area.

[4] Comparing to the answers in the database which are provided by other respondents

Table 3 - Summary of Available Surveys 2

Test	**Usage**				**Questions**		**Criteria**	**Analysis**	
	Generic Agile Adoption	Partial Agile Adoption	Generic Agile Assessment	Partial Agile Assessment	Number	Type		Sufficient	Partial
*Agile Karlskrona Test (*Mayberg Consulting, 2008*)*	√				11	Multiple			√
CMI Lean Agility (CMI Lean Agility, 2009)	√								
Comparative Agility Assessment (Cohn and Rubin, 2007)			√		±120	Five-point scale		√	
Dr Agile (Santeon Group, 2012)	√				22		Organization's readiness to adopt Agile practices	√	
Nokia Test (Sutherland, 2009)		√		√	9				√
Scrum Checklist (Kniberg, 2010)		√		√			Existence of required practices		√
Signet Consulting (Signet Research and Consulting, 2005)			√		22		Level of applying practices		√
Thoughtworks Agile Self Evaluation (Thoughtworks Studio, 2010a)			√		20	Multiple choice		√	
Thoughtworks Build And Release Management (Thoughtworks Studio, 2010b)				√	20	Multiple choice		√	
42-Point Test (Waters, 2008)			√		42	True/false statements	Team's consistency rather than Agility score	√	

2) **Comparative Agility Assessment (S2)**: It is developed based on the assumption that most organizations prefer to be more Agile than their competitors rather than striving to be "perfectly Agile" (Williams et al, 2010). Therefore, it assesses the Agility of teams/organizations relative to other teams/organizations that responded to the questionnaire.
It is a survey-based tool, in which answers are recorded on a five-point scale (true, more true than false, neither true nor false, more false then true and false). Questions can be answered from four different views, which are team, department, division, and organization. Eight dimensions are considered as the basis for the assessment: 1) teamwork, 2) requirements, 3) planning, 4) technical practices, 5) quality, 6) culture, 7) knowledge creating, and 8) outcomes (Cohn and Rubin, 2007).
Each dimension consists of three to six characteristics, and each characteristic of a number of statements that have to be verified by the respondents. Each statement indicates an Agile practice, and the respondent should determine the truth of the statement based on their actual practices in the team/organization.
The questions sufficiently address different areas of Agility and the number of questions per area is reasonable. Participants get a free report indicating their level of practicing Agility in each dimension and each characteristic separately in comparison to other practitioners.

4.1.3. Reflections on Survey Identification

It may be perceived that Agility ought to be defined before being able to assess it. However, we intentionally did not pick any definition from any sources in order to be able to objectively conduct searches. The intention was to compare the definitions of Agility found in several surveys with the opinions and expectations of the practitioners who participated in this research study. Therefore, we did not begin our research with a strict definition for Agility.

To enable process improvement systematically and efficiently in an agile development environment, software companies/teams need to know how Agile they are and what Agile practices they should be implementing or improving. For process improvement purpose, it is crucial to understand the current situation (baseline) before planning any improvement initiatives. However, we do not claim that the objective should be to become fully Agile and implementing all Agile practices, but we encourage examining the status of practicing Agility within a team/organization in alignment with the goals and objectives of the team and/or organization.

Thus, the major purpose of conducting this research study was to understand how Agility can/should be assessed. Hence, we searched for proper surveys that evaluate Agility of a team or an organization. The existing frameworks found in the research literature did not seem to be applicable in practice because it was not known whether they could be used without help of a consultant. On the other hand, Agile researchers have not evaluated the available checklists and surveys developed by practitioners.

In this research, we searched and evaluated the available surveys and utilized two of them as input to an industrial case study with the purpose of evaluating the applicability of the surveys. Therefore, instead of introducing a new framework, we scientifically examined the Agile assessment surveys. This helps both researchers and practitioners to gain awareness of the existing work in the area and to benefit from an analysis on the strengths and weaknesses of the studied surveys.

4.2. Case Study

The results of each survey for the three teams T1, T2, and T3 are presented separately as well as for T1 and T2 together (since they have the same customer) and in comparison with T3, followed by the customers' perception. It should be noted that one person is common in both T1 and T2, i.e. both teams have the same Scrum master.

4.2.1. Data Imputation

As expected, data imputation was needed since not all participants were able to provide responses to all questions or they forgot to answer some questions given the number of questions. Table 4 summarizes the results of data imputation. All teams have lower imputation rates in S2 compared to S1. T2 has the greatest imputation rate of 12.6% in S1.

After data imputation, we calculated each person's Agility score/rank in each survey (i.e. two distinctive scores from two surveys). Then, these scores were used to calculate descriptive statistics (i.e. minimum value, maximum, mean, standard deviation, and median) for each team. In addition, a variety of statistical methods were used in order to compare the surveys and the teams.

These techniques include graphical methods, statistics like the Pearson correlation coefficient and agreement indices and furthermore tests of hypotheses like ANOVA and Mann-Whitney. The statistical analysis is presented in depth in Section 4.3.

Table 4. Summary of Data Imputation

S1		**Total number of options = 64**	
Team	**Participants**	**Total options**	**Imputed**
1	6	64*6=384	10 out of 384, i.e. 2.6%
2	6	384	12.6%
3	5	320	4%
S2		**Total number of options = 348**	
Team	**Participants**	**Total options**	**Imputed**
1	6	348*6=2088	1.4%
2	6	2088	1.4%
3	5	1740	1.4%

4.2.2. S1 – Thoughtworks' Agile Self Evaluation Survey

It considers Agility as the ability to adapt to ever-changing environments, and the questions are based on the Agile principles including both management and development practices. Based on the answers, it ranks the team/project on a scale ranging from "regressive" to "Agile".

Being in the "regressive" state represents behaviors that hinder the ability to adapt and the "Agile" state represents behaviors that focus almost exclusively on adapting to changes.

When presenting the results to the team, we made minor modifications, which were changing “ad-hoc” to “neutral”, and “regressive” to “non-Agile”, which sounds more positive and encourages the team to participate in the discussions around the results instead of risking a more defensive mode.

Figure 2 summarizes the results of the survey for all teams separately as well as for T1 and T2 together (represented as C1). T1 and T2 have the same customer and practically are seen as one team by the customer organization, while T3 has a different customer. Therefore, T1 and T2 can be compared against each other, and both together (C1) against T3 with the purpose of learning from each other. In addition, the comparisons would also help to estimate the Agility of the studied organization.

During the presentation session, the teams also categorized the improvement areas as external and internal. The external refers to the factors that are introduced outside the team (e.g. due to the customer side). The internal areas were listed down in the action plan of the team.

Figure 2 indicates that only T2 is Agile. T3 is more Agile than C1 and T1 is less Agile than C1, but all of them are in the neutral area.

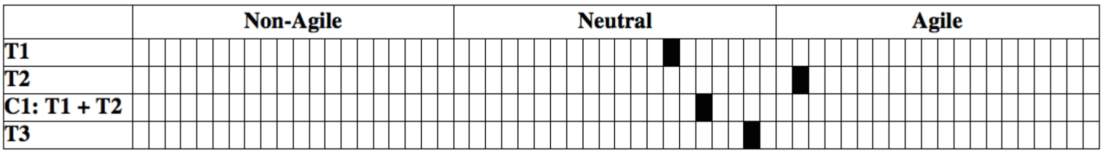

Figure 2. Results of Survey 1

Surprisingly, the score for C1 (T1+T2) is closer to T1’s score. Considering the weighted mode response for each question as the majority might lead to domination of the answers with higher level of confidence, and hence this type of outcome. By looking at the collected data, T1 team members are more confident about the given answers to some questions comparing to T2, which can be one explanation to this unexpected result.

4.2.3. S2 – Comparative Agility Assessment Survey

This survey produces two different graphs based on the answers. The first graph shows the level of Agility in different areas/dimensions (Cohn and Rubin, 2007) in comparison to the other answers in the survey’s database. The second graph represents Agility level for all characteristics of each area. The analysis for each question is given in terms of the number of standard deviations from the mean value in the survey’s database. So, having a positive score means the answer is “better than” the average answers in the database. More details regarding the analysis can be found in (Cohn and Rubin, 2007).

Figure 3 summarizes the results (only the first graph) for T1, T2, T3, and T1 together with T2. Surprisingly, T2 is ranked less Agile than T1 in this survey, which is different from S1. Also, T3 is significantly more Agile than T1 and T2 separately and together unlike S1. We will discuss the possible reasons for this contradiction in Section 5.2.

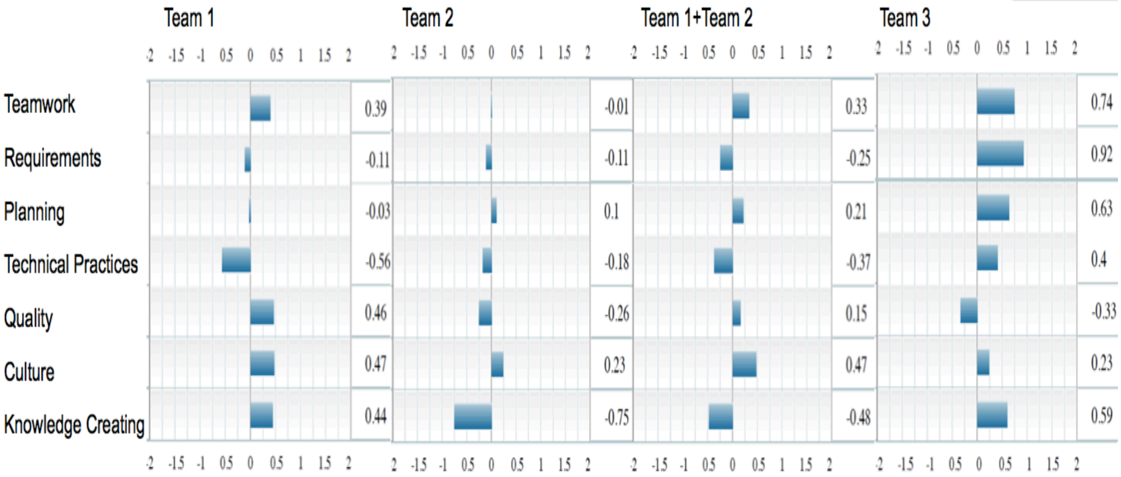

Figure 3. Results of Survey 2

4.2.4. Collective Agile Areas – All Surveys

One critical topic of discussion was around the contradictions in the results of different surveys. For example, T2 is more Agile than T1 in S1, but less in S2. Furthermore, T1 is Agile in most areas of S2 while T2 is not, but all together becomes sufficiently Agile (Figure 3). On the other hand, T3 is not Agile according to S1, but according to S2 it is significantly more Agile than T1 and T2 and both together.

To examine the combination of surveys, we put all Agile areas from all surveys in a scale from "not Agile" to "very Agile" and discussed whether it depicts the teams' status better than each individual survey. Table 5 maps the studied Agile areas to the Agile scales. If one specific area was examined in more than one survey, we considered the one with higher Agility rank.

It is surprising that T1 and T2 perceive project management (PM) differently although the same person (i.e. Scrum master) is responsible. It however might be due to the differences in their tasks and the way the work is formulated e.g. T1 uses sprints while T2 tasks are oriented differently to handle TRs. The results for T3, however, seem to be closer to the results of S2 than the results of S1.

Table 5. Collective Agile Areas from Both Surveys

Team 1				
Not Agile	**Neutral**	**Acceptable Agile**	**Agile**	**Very Agile**
1) Project Management 2) Business Responsiveness 3) Requirements 4) Technical practices	1) Build Management 2) Simplicity 3) Testing	1) Planning	1) Collaboration & Communication 2) Teamwork 3) Quality 4) Culture 5) Knowledge Creation	1) Governance 2) Configuration Management
Team 2				
Not Agile	**Neutral**	**Acceptable Agile**	**Agile**	**Very Agile**
1) Requirements 2) Technical practices 3) Knowledge creation 4) Quality	1) Simplicity 2) Testing	1) Teamwork	1) Collaboration & Communication 2) Project Management 3) Culture 4) Planning	1) Business Responsiveness 2) Governance 3) Build Management 4) Configuration Management
Team 1 + Team 2				
Not Agile	**Neutral**	**Acceptable Agile**	**Agile**	**Very Agile**
1) Business Responsiveness 2) Requirements 3) Technical practices 4) Knowledge Creation	1) Simplicity 2) Testing	1) Project Management 2) Quality 3) Planning	1) Collaboration & Communication 2) Culture 3) Teamwork	1) Build Management 2) Configuration Management 3) Governance
Team 3				
Not Agile	**Neutral**	**Acceptable Agile**	**Agile**	**Very Agile**
1) Testing 2) Quality	1) Simplicity		1) Project Management 2) Configuration Management 3) Technical Practices 4) Culture	1) Business Responsiveness 2) Governance 3) Collaboration & Communication 4) Build Management 5) Requirements 6) Teamwork 7) Planning 8) Knowledge Creation

4.2.5. Customer's Perception and Team's Self-Assessment

The customer representatives were interviewed in order to complete the contextual information of the project as well as to gather their view on the teams' Agility. Figure 4 summarizes all teams' scores (0 to 5) in different areas. It should be noted that the customer representative for T2 did not assign any score to the "knowledge creating" area due to unawareness of the details regarding how the team performs knowledge creation. Similarly, T3 did not receive any score in the "technical practices" area. Therefore, these areas were not considered in calculating the mean Agile value for T2 and T3.

The mean values were calculated to be 4.4 for T1, 3.5 for T2, and 3.9 for T3. This is in alignment with the Scrum masters' perception. Accepting these perceptions, T1 is more Agile in practice than T2 (which matches the results of S2), and T3 (which contradicts the results of S2). More in-depth discussions are provided in Section 5.2.

In the team self-assessment, the teams had many discussions and reflections on their own practice of Agility. This helped create a better joint understanding of their strong areas in terms of Agility. The teams agreed that S2 better captured their perception of their own Agility as well as a comparatively to the other teams. The main advantage was that it had more questions, and hence the results became for fine-grained and less dependent on single questions. A more in-depth comparison of the surveys is provided in Section 5.

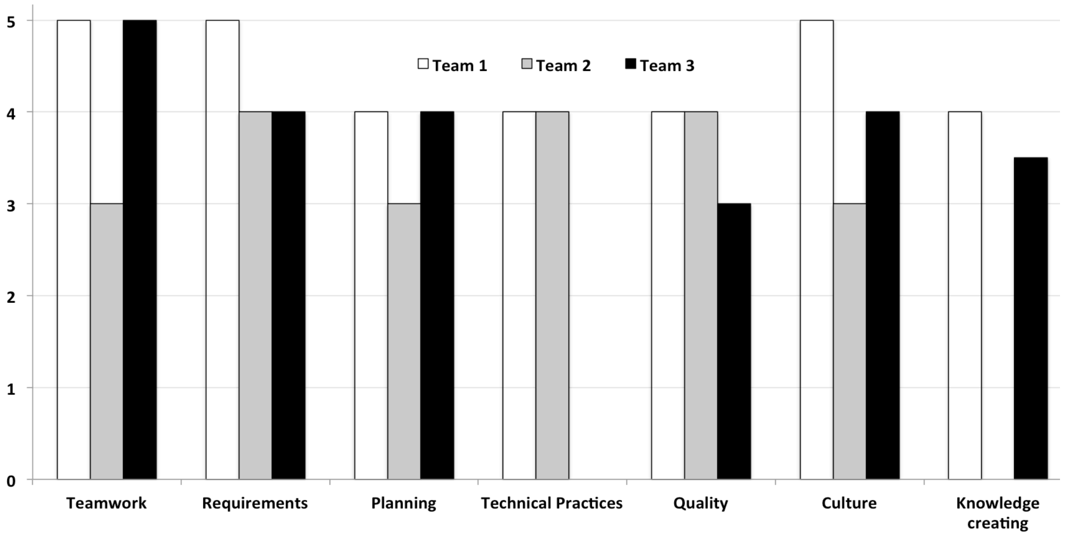

Figure 4. Customers' Perception

4.3. Using Statistical Tests for Comparing Teams and Surveys

The statistical analysis described below includes methods applied to the individual Agility scores obtained from the studied Surveys.

Table 6. Summary of Descriptive Statistics for the Agility Scores

Team		Score of S1	Score of S2
T1	Participants	6	6
	Minimum	0.40	0.37
	Maximum	0.57	0.57
	Mean	0.51	0.46
	Std. Deviation	0.06	0.065
	Median	0.51	0.44
T2	Participants	6	6
	Minimum	0.30	0.30
	Maximum	0.65	0.59
	Mean	0.53	0.45
	Std. Deviation	0.12	0.11
	Median	0.54	0.49
T3	Participants	5	5
	Minimum	0.50	0.53
	Maximum	0.67	0.67
	Mean	0.61	0.61
	Std. Deviation	0.069	0.07
	Median	0.65	0.63
Total	Participants	17	17
	Minimum	0.30	0.30
	Maximum	0.67	0.67
	Mean	0.55	0.50
	Std. Deviation	0.10	0.11
	Median	0.55	0.50

4.3.1. Descriptive Statistics for Agility Scores

The descriptive statistics of the Agility scores, separately for each team and for all individuals, as outcome of S1 and S2 are presented in Table 6.

All samples (six in total which is three teams participating in two surveys) are very small, and the Kolmogorov–Smirnov test of normality for each team and each survey separately did not show statistically significant difference from the normal distribution for any of the six samples (p>0.05).

The dot plots in Figure 5 show the distribution of the scores of the three teams from S1 and S2. One individual score in each team seems to be an outlier in S1 while the scores are more divergent in S2. However, we have not removed any individuals from analyses because our intention was to reflect the participating teams' Agility as it is and the dot plots only show that either individuals are practicing different level of Agility or their awareness or perception of actual practices are different.

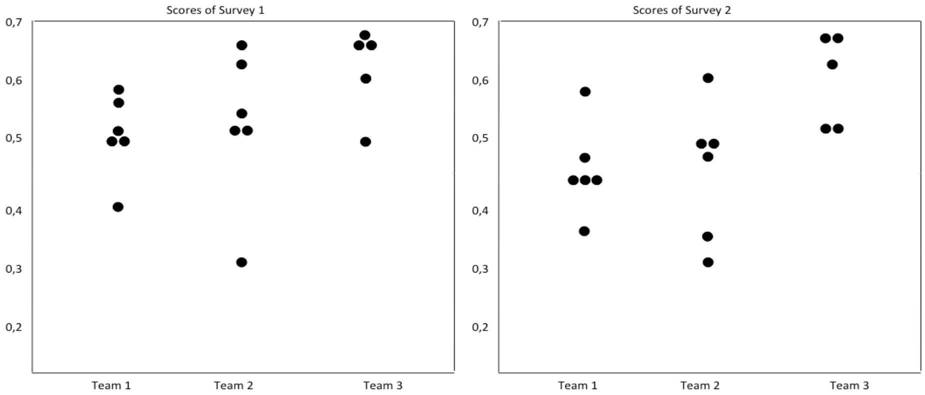

Figure 5. Distributions of Agility Scores from S1 And S2

4.3.2. Correlation Analysis of Agility Scores

We used the Pearson's r correlation coefficient (Marascuilo and Serlin, 1998) for each team separately to test whether the correlation between the scores of different surveys is significant. Correlated scores would indicate alignment of different Surveys' results.

There is a significant correlation (r=0.897, p=0.021) between the scores of T1 in S1 and S2. On the contrary, the scores of T2 are not significantly correlated (r=0.682, p=0.136) and the same holds for T3 (r=0.376, p=0.533).

In other words, this test examines whether a significant correlation exists between the Agility scores given by S1 and S2. We found a significant correlation only for one team (out of three), and hence the surveys do not show strong correlation.

4.3.3. Comparison Between Average of Individual Scores and Team Score from Fuzzy Approval Voting

We applied the "one-sample t-test" (Marascuilo and Serlin, 1998) and the results are summarized in Table 7. The t-test was used to compare the mean value of the individual scores (represented as "Mean value of individual scores") of each team with the unique score that each team received after entering the fuzzy approval voting results, separately for each question (represented as "Team score"). The p-value is also given for each test.

The goal of this test is to compare the mean value of individual Agility scores with the team's score. This test is useful to examine which team's score has probably been affected by uncertainty and in which survey, because the uncertainty of individual responses is considered when computing the team's Agility score while it is not considered in computation of individual Agility scores.

In general, there are no significant differences. The only exception is T2 in S1. The mean score of individuals (0.53) is significantly different (p=0.018) from the Team score (0.70). In an effort to look closer at this result, we calculated the mean confidence of each individual based on their answers to the fuzzy approval voting, and from them the mean confidence of each team.

We can see in Table 7 that the lowest mean value for team confidence appears in T2 and in S1. We can also see that the mean confidence for all teams in S2 is very high, close to 1. This may be an indication that the Agility score derived from automated surveys is vulnerable to uncertainty, which cannot actually be measured by such implementations.

Table 7. Summary of t-tests – Comparing Mean Values of Individual Scores Versus Team Score

	S1				S2			
	Team score	**Mean value of individual scores**	**p-value**	**Mean team confidence**	**Team score**	**Mean value of individual scores**	**p-value**	**Mean team confidence**
T1	0.52	0.51	0.605	0.84	0.49	0.46	0.247	0.92
T2	0.70	0.53	0.018	0.81	0.48	0.45	0.546	0.90
T3	0.67	0.61	0.143	0.85	0.62	0.61	0.663	0.96

4.3.4. Differences Among Mean Scores of Teams for Each Survey

One-way ANOVA (Analysis of Variance) (Marascuilo and Serlin, 1998) was used in order to test if there are significant differences between teams. The ANOVA test was applied separately for the two different surveys. The test did not show significant difference between the teams (p=0.153) for S1, but S2 gave p=0.013, which indicates a significant difference between teams. The Bonferroni post hoc test (Marascuilo and Serlin, 1998) for S2 revealed a significant difference between T2 and T3 (see also Figure 5).

4.3.5. Degree of Agreement on Team Level

Each question in both surveys (1 and 2) has multiple choices. In order to examine the degree of agreement between teams, we firstly calculated the Pearson correlation coefficient from the results of fuzzy approval voting for each question separately. So if the four options of a question receive "votes" (0.75, 0.75, 3.75, 0) from the members of T1 and (0, 2.75, 0, 0) from the members of T3 the Pearson coefficient is r=-0.225. Since the Pearson coefficient ranges from -1 to 1, a value of -1 means that two teams have voted in complete disagreement for a specific question while +1 means that they have voted in complete agreement. Figure 6 shows the distribution of these indices of agreement and Table 8 summarizes their descriptive statistics.

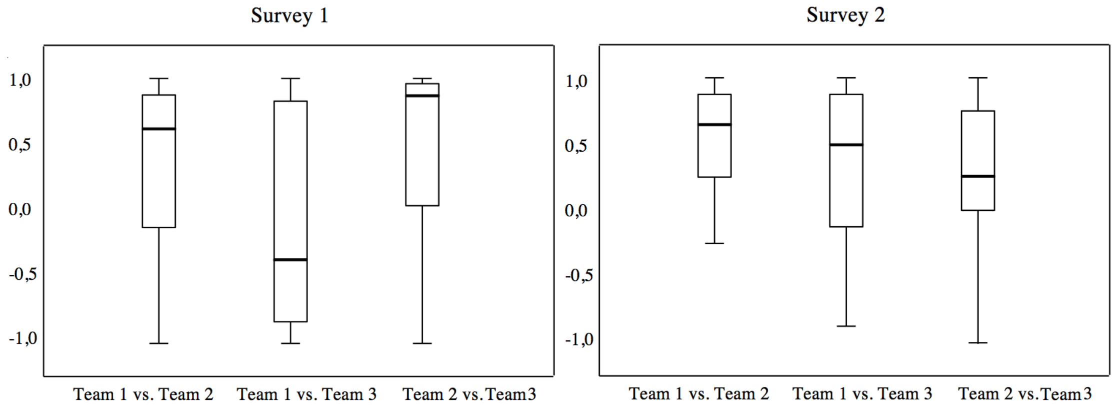

Figure 6. Degree of Agreement between Teams in S1 and S2

As Figure 6 and Table 8 show, there are questions where we had absolute or large disagreement and questions with perfect agreement. However, the differences between teams were not found to be statistically significant according to the results of the Friedman's test for related samples (they are not normally distributed) separately in S1 and S2.

The differences between the two surveys were also not significant at the 0.05-level according to the Mann-Whitney test (although there are indications at the 0.10-level).

In general, the picture in S1 seems to be different from S2. In S1, we have a high degree of disagreement between T1 and T3 and a high degree of agreement between T2 and T3. In S2, the disagreement seems to be smoothened since the means and medians of all coefficients computed are positive.

Table 8. Descriptive Statistics – Pearson Coefficients Between Pairs of Teams – S1 and S2

S1

	T1 vs. T2	T1 vs. T3	T2 vs. T3
N	20	20	20
Minimum	-1.00	-1.00	-1.00
Maximum	1.00	1.00	1.00
Mean	0.377	-0.052	0.437
Std. Deviation	0.696	0.820	0.656
Median	0.688	-0.033	0.840

S2

	T1 vs. T2	T1 vs. T3	T2 vs. T3
N	58	58	58
Minimum	-0.378	-0.794	-0.908
Maximum	1.00	0.980	0.980
Mean	0.515	0.357	0.276
Std. Deviation	0.389	0.499	0.501
Median	0.624	0.445	0.312

These results are indications of divergence between the two surveys. It is also interesting that the standard deviations of agreements are large in S1, which can be due to the fact that too few questions are asked in S1 to assess a team's Agility, and hence, the score may change dramatically depending on the given answers. S2 is less sensitive in this regard since many more questions are asked. It should also be noted that the same team has shown different behavior in different surveys, which can be related to the number of questions as well.

4.3.6. Degree of Agreement within Teams

Here, we tested the level of agreement within teams. In this regard, we focused on the maximum number of participants within each team who had provided the same answer in response to each question in S1 and S2 (regardless of their confidence).

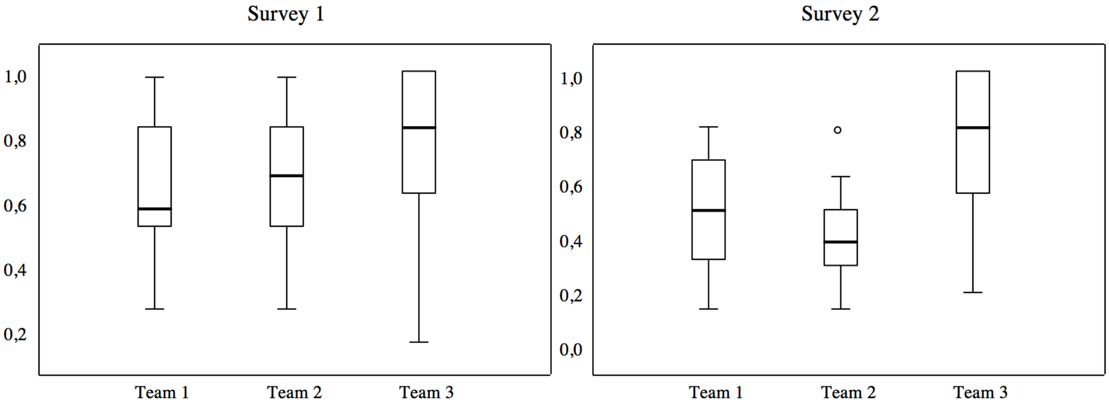

Figure 7. Distribution of Agreement Index for Each Question within Teams in S1 and S2

Hence, for each question in S1 and S2, we computed the agreement index for the teams separately. The agreement index is the ratio of the maximum number of team members that chose the same answer divided by the total number of team members. So, if in a question, regardless of its options, the maximum number of participants that agreed in the same option is 4, while the total number of team members is 6, the agreement index is 4/6=0.67. An agreement index equal to 1 means that all team members chose the same answer while the lowest agreement level can be 1 divided by the number of team members in the case where each team member had different opinion. The results are presented in Figure 7 and Table 9.

Table 9. Descriptive Statistics – Agreement Index Within Teams – All Questions in S1 and S2

S1

	T1	T2	T3
N	1.1.1 20	1.1.2 20	1.1.3 20
Minimum	0.33	0.33	0.20
Maximum	1.00	1.00	1.00
Mean	0.62	0.66	0.73
Std. Deviation	0.22	0.18	0.25
Median	0.58	0.67	0.80

S2

	T1	T2	T3
N	58	58	58
Minimum	0.17	0.17	0.20
Maximum	0.83	0.83	1.00
Mean	0.53	0.44	0.72
Std. Deviation	0.17	0.14	0.24
Median	0.50	0.42	0.80

The difference of each team between the two surveys was tested for significance with the Mann-Whitney (M-W) test. The T1 agreement does not have a significant difference in both surveys (M-W p=0.135) although in S1 the agreement seems to be generally higher. T2 has significantly lower agreement in S2 (M-W p<0.001). The consistency of T3 in both surveys is impressive. Indeed, T3 shows the same highest mean and median agreement index in both surveys (M-W p=0.809).

The agreement indices of the three teams have no significant difference in S1 (related samples Friedman's test, p=0.170), while the difference is significant (p<0.001) in S2. More specifically, the agreement within T3 is significantly higher than within the other two teams, which can be due to the fact that team members have been working together over a longer period of time compared to the other two studied teams and hence are more familiar with the actual practices and processes.

4.3.7. Interpretation and Conclusion of The Scores

Based on the analyses of the final scores, it seems that different surveys give different results regarding both the individual and the team scores, which is not completely surprising given the characteristics of each team and posing different questions in different surveys.

However, we cannot claim that the influence of imputation can be neglected since the score of T2 after data imputation in S1 is significantly different from the mean of individual scores. It should be noted that this combination had the highest imputation rate as well (see Table 4).

5. Comparison of the Two Surveys

This section provides a reflection and discussion of the use of the two surveys. Section 5.1 presents a comparison of the two surveys chosen based on the evaluation criteria presented in Section 3.3.2. The outcome with respect to the different teams and the use of the two surveys is discussed in Section 5.2. Here, the outcomes are discussed from the surveys' point of view, while in Section 4 the results were discussed as outcomes for the organization and teams. In Section 5.3, a candidate survey is suggested based on the experiences from the case study. Finally, in Section 5.4 the definition of Agility based on the selected survey is presented.

5.1. Strengths and Weaknesses of Surveys

S1: Although the examined Agile areas seem to be sufficient, the number of questions per area is rather low. The potential risk is that some areas are assessed only by one question (e.g. "governance"). On the other hand, the total number of questions is 20, which can be considered as an advantage since it does not take long time to answer all of them. The participants had some difficulties in interpreting a few questions, but in general we cannot claim that they are ambiguous. For each question, different number of options are given that can be seen as a disadvantage if one plans to analyze the data besides the provided analysis conducted automatically by the survey.

The analysis is transparent and the presentation of the results is visual along with valuable recommendations. The organization providing the survey is well known and the survey has been recently mentioned in the research literature e.g. in (Taromirad and Ramsin, 2009).

S2: A sufficient number of Agile areas are examined with a rather high number of questions. The total number of questions is high in comparison to the previous survey, but answering them is quite simple since one should determine to what extent given statements are correct (using a five grade Likert scale as described in Section 4.1). Like S1, it is hard to judge its clarity (there is always some possibilities to misinterpret a specific question, but this is valid for most surveys). This survey is mentioned in literature e.g. in (Taromirad and Ramsin, 2009) and has been updated several times (e.g. between March 2013 and December 2013) and changes such as reducing the number of questions, reformulating the questions, etc. were made. We believe that the main disadvantage with this survey is that the analysis is not transparent although the results are sufficiently well presented. In S2, it is possible to ask for a separate link so that all team members could participate and give their own answers independently whilst this feature is not provided in other studied surveys. A summary of this evaluation is presented in Table 10.

Table 10 - Strengths and Weaknesses of the Studied Surveys

Criterion	**Survey 1**	**Survey 2**
Covered Agile areas	Sufficient	Sufficient
Number of questions per area	Low	Sufficient
Total number of questions	20	127
Formulation/readability of questions	Acceptable	Acceptable
Transparency of the analyses	Acceptable	Low
Presentation of the results	Acceptable	Acceptable
Updates/reformulation of the questions/analyses	No	Yes
Being acknowledged by academia	Yes	Yes

5.2. Comparisons of the Survey Results

The Agility of three different teams was assessed through two surveys as well as interviews with the Scrum master, a team representative and the respective customer (i.e. product owner) for each team. Below, we compare the results of the two surveys with each other and with the perception/expectation of the teams, the customers, and the Scrum master. It should be noted that the results were not discussed with the customers and interviews with them were held separately with a different set of questions. In brief, the objective here is to compare the outcomes of the surveys from the surveys' point of view, while in Section 4 the outcome is discussed from the organization's and teams' point of view. The comparison here is conducted to form a basis for identifying a candidate survey based on the results from the case study.

A summary of different teams' Agility ranks given by the studied surveys is provided in Table 11 to help better understanding the discussions presented in the following sections.

Table 11 – Teams' Agility Scores Given by the Studied Surveys

	Survey 1	**Survey 2**
Team 1	Neutral	*Agile*
Team 2	*Agile*	Non-Agile
Score of Team 1 > Score of Team 2		
Team 1+2	Neutral	Agile
Score of Team 1+2 = Score of Team 1 *Score of Team 1+2 > Score of Team 2*		
Team 3	Neutral	*Agile*
Score of Team 3 = Score of Team 1 *Score of Team 3 = Score of Team 1+2* *Score of Team 3 > Score of Team 2*		

5.2.1. Team 1

T1 is "neutral" in S1, and "Agile" in S2. One reason for this contradiction could be that S2 is comparative and T1 could be more Agile than the teams in the database whilst not highly Agile by itself. In addition, the number of questions per area in S1 is smaller than in S2, which means if the participants had difficulty interpreting a question in S1 and answered wrongly, the rank could become "non-Agile" only by that question in that specific area whilst more questions were asked in S2 to judge the Agility.

Another type of contradiction in the results is having different scores in the same areas that are examined in the surveys. Such an area is "requirements analysis". S1 ranks the team as "non-Agile" whilst it is "neutral" in the "requirements" area according to S2.

The difference could be due to the number of questions that is three in S1 and 14 in S2. The greater number of questions could lead to the higher score since misinterpreting one or two questions out of three could lead to being "non-Agile" in S1.

The team and the Scrum master believe that not being Agile in requirements analysis is not unexpected since the team's influence on this process is minimal because it is completely handled by the customer.

In addition, they think that they can be interpreted as "non-Agile" in "technical practices" too since they do not follow them exactly as in literature. However, they perceive themselves as Agile regarding the quality, testing, and communication. The customer has scored the team highly Agile in all areas and is satisfied with its working style and the performance.

For T1, it is hard to evaluate whether the outcome from S1 or S2 is most correct. There was a perception that more questions provided a more fine-grained judgment, which was perceived as good, and hence gave an advantage to S2.

5.2.2. Team 2

Surprisingly, T2 is "Agile" in S1, and "non-Agile" in S2, which is different from T1. This could be due to the large number of "not applicable" answers given by the team members in S2, which could simply be caused by the order of surveys in the questionnaire or the questions were not perceived relevant given their focus on TRs[5]. S2 questions were the last in the list of questions and participants could become tired or less motivated to respond to them carefully in comparison to S1 (supposing that one responds to the questions from the beginning to the end). However, we contacted the team members to confirm their responses, but did not receive any new responses.

The observed contradiction in "requirements analysis" is, however, exactly the same as for T1. In addition, S1 scores T2 "Agile" in "collaboration and communication" and S2 "neutral" in "teamwork", which is contradictory.

Both teams and the Scrum master were not surprised by not being scored Agile in "requirements analysis" "technical practices", and "knowledge creating". They agree with being Agile in "quality", "testing", and "communication". Their customer perceives the team Agile in all areas as for T1.

The customer and the team perceived the team as more agile than the outcome of S2. Thus, the team favored S1 over S2.

5.2.3. Team 3

Similar to T1, T3 is "neutral" in S1, and "Agile" in S2. The reasons for this contradiction are also similar to the motivations provided in Section 5.2.1.

The observed contradiction is also similar to the previous teams, which is in "requirements analysis" area. S1 scores T3 in "requirements analysis" as "acceptable Agile" while S2 gives "highly Agile" score for "requirements".

The results were shared and discussed with the team. Some activities such as "testing" and "requirements analysis" that are externally performed shall be further discussed before making any decisions. In other areas, the team believes that respective practices are customized over time according to the needs of the project and in alignment with the team's culture and capabilities and hence shall not be changed.

All surveys have given a "non-Agile" score to T3 in areas related to "testing" or "quality", which is defined as externally performed by the team. The customer has also given the lowest score (3 out of 5)

[5] This speculation has been discussed with the Scrum master and it is confirmed.

to the “quality” area compared to other areas. In order to be consistent with the testing process in the customer organization, a few sprints is taken together as input for testing. This may result in late detection of defects.

In summary, the team perceived S2 to provide the best picture of their status with regards to agility.

5.2.4. Comparing T1 with T2

T2 is scored as more Agile than T1 in S1 and less Agile in S2. After discussions with the teams and the Scrum master, we believe that the results of S2 are closer to reality (as perceived by the participants). However, we have observed some inconsistencies that are explained as follows.

T1 and T2 had different scores in the areas that are expected to be exactly the same mainly because either the same people have been involved in performing them or the execution approach has been the same. These areas are “project management” and “build management”.

In the results, T1 is “non-Agile” whilst T2 is “Agile” in “project management” which pinpoints a big difference between the teams’ perceptions. Furthermore, T2 is “Agile” in “build management” but T1 is “neutral”. Although the difference between “Agile” and “neutral” is smaller than between “Agile” and “non-Agile”, it shows misalignment in the teams’ perceptions/expectations.

T1 and T2 work in the same organization, the same field, for the same customer organization (but different units) and are managed by the same Scrum master. Having these similar conditions, we expect the teams to be equally Agile and aligned in the Scrum way of working. The differences however could be due to utilizing slightly different practices by the teams. For example, retrospective meetings are skipped by T2 while T1 regularly holds them. In addition, if the communication among team members is not effective, misalignment could be seen as we observed in this research study.

On the other hand, both teams were very open to the discussions and actively participated in interpreting the results of the surveys, which is according to the Agile principles.

Based on the discussions with the teams and the Scrum master, it was concluded that S2 provided a truer picture of the agility.

5.2.5. Comparing All Teams against Each Other

S1: the results of S1 show that T2 is more Agile than T3, and T3 is more Agile than T1. However, T1 and T2 as a team (represented as C1 in Figure 2) is less Agile than T3. Thus, the order of Agility according to S1 is: T2 > T3 > C1 >T1

S2: T3 is more Agile than T1, and T1 is more Agile than T2, but C1 is less Agile than T3. Thus, the order of Agility according to S2 is: T3 > T1 > C1 > T2

This ranking was also discussed with the Scrum masters and teams, and the conclusion was that the results of S2 are again closer to reality.

In summary, the findings illustrate that different surveys may give different results, and hence it becomes important to use the survey that best reflect the perception of the people working in the unit being assessed.

The comparison illustrates that different surveys come up with quite different results, and hence it becomes a matter for an organization to judge which survey provides the truest picture with respect to their own perception of their agility. In this case study, S2 is judged to provide the truest picture of agility (as defined by the context of the case study).

5.2.6. Comparing Results of Surveys with the Statistical Tests

The results of the statistical tests indicated a significant correlation between the scores of T1 for S1 and S2. This means if T1’s score is high in S1, it most probably is high in S2 as well. This correlation is, however, not observed for the other teams.

In addition, we found a significant difference between T2 and T3 for S2, which is similar to the findings presented in Figure 2. Finding a significant difference between T3 and T2 or T1 is not surprising since the context is not similar e.g. product and contract features are different. We would have been suspicious if the difference would have been between T1 and T2 (see Section 5.2.4 for detailed discussion).

5.3. Candidate Survey

As discussed in the previous sections, the studied surveys did not assign the same scores of Agility to the studied teams. Therefore, we discussed the results with the participants to decide which survey or a combination of surveys is the best tool for assessing a team/project's Agility (at least in our context).

We compared the results with the perceptions of the teams, their customers, and the Scrum masters. Generally, the perception is that the S2 results are more aligned with the collected perceptions for T1 and T3. However, its results for T2 are slightly different from the expectations. Furthermore, we discussed the differences between the teams with the Scrum masters and figured out that S2 shows the differences better than S1.

The main advantage of S2 is that it provides more questions for each area in comparison to S1. In addition, it partially examines the goodness of utilizing Agile rather than only presence/absence of certain practices, which is an advantage over most available surveys. Therefore, S2 seems to be sufficiently good for the purpose of assessing a team/project's Agility. Furthermore, it compares teams/projects/organizations to similar ones in the world and gives a comparative rank.

5.4. Definition of Agility According to the Selected Survey

According to the structure of questions for Survey 2 (Cohn and Rubin, 2007), the survey defines Agility as constant changes in a team or organization as it evolves in its lifecycle. The changes are classified as related to: (1) teamwork, (2) requirements, (3) planning, (4) technical practices, (5) quality, (6) culture, and (7) knowledge creating.

Each category of change has its own characteristics that must be fulfilled through applying certain practices (i.e. Agile practices).

6. Threats to Validity

The validity threats regarding reliability and generalizability of this research as well as what we did to overcome them are discussed as follows.

Internal Validity: In order to draw valid conclusions, we applied a triangulation technique, which is a method that compares three or more types of independent perspectives on a given aspect of the research process e.g. methodology, data, etc. (Guion, 2002). The triangulations used in this study were data, investigator, and methodology triangulations.

Data Triangulation: The data was collected from three sources (team, Scrum master and customer) to capture different perspectives/expectations on Agility.

Investigator Triangulation: In data collection and data analyses, more than one researcher was involved in performing and validating the work.

Methodology Triangulation: We collected data both qualitatively (interviews and open discussions) and quantitatively (survey). Hence, both quantitative and qualitative data analyses were performed.

Furthermore, we studied three units of analysis (T1, T2, and T3) in the case study rather than relying on single data points for a single team.

It should be noted that we did not transcribe the interviews immediately and did not directly confirm the content with the interviewees. However, the results of analyses and the conclusions were openly presented and discussed with all participants, which reduces the risk of misinterpretations and misunderstandings.

A concern with the results is that the analyses performed by the surveys are not completely transparent. For example, the number of participants is small in comparison to the number of questions, and we are not aware how it is managed in S2 to ensure the reliability of the statistical calculations. In addition, the S2 questions are updated over time, which makes it hard for us to know how the new questions are compared against the existing database with a different set of questions. This threat may affect our statistical calculations as well since we used the surveys to calculate the Agility rank for each individual.

S2 provides a separate link that all team members can answer the questions independently, and it provides the collective results for the team rather than for each individual participant. We did not use this feature mainly because we intended to independently calculate the team's representative answer. Furthermore, before inserting the responses to the surveys, we discussed ambiguities around the responses in separate meetings with the purpose of clarifying the answers. The team's representative answer for each question was finalized after discussions with the Scrum master and the person representing the team.

We used the survey questions and options with no modifications and therefore any ambiguity with them remained. However, we added the option to skip a question as well as to provide the confidence level for the given answer in a way that the participants could determine how sure they are about the given answer rather than skipping it completely. Furthermore, the questions are grouped in the original surveys whilst we hid the classifications to the participants mainly to avoid bias when answering the questions. On the other hand, this could cause confusions and misinterpreting the intention of the question.

Optimally, from a research methodology point of view, the order of questions in each survey and order of the surveys should have been random. We did not manipulate the list of questions in the surveys, but placed each studied survey in a separate Excel sheet so one could answer them in any desirable order, although most likely the respondents answered them in the same order.

Another threat is that one participant (the Scrum master) is in both T1 and T2. We performed separate analyses for the teams excluding the Scrum master, and concluded that the scores for T1 and T2 become less alike in this case. In this paper, we have not reported the results of separate analyses excluding the Scrum master.

Finally, due to other obligations of the participating teams, some of the team members could not take part in the case study. Nevertheless, 6 members out of 8 from T1, and 6 out of 10 from T2, 5 out of 9 from T3 indicate 75%, 60%, and 55% participation respectively, which is more than average for all three teams. However, the responses of other members of the team (e.g. 4 others in T2) might have influenced the team's score in different Agile areas and hence the results of this study.

External Validity: We discuss the external validity regarding generalizability, which is to what extent findings can be generalized to and across populations of persons, settings, and time (Creswell, 2003). The results of assessing Agility are specific to the participating teams. Hence, the conclusions might be context specific. Therefore, we cannot claim that a similar study on different teams could result in nominating the same survey for assessing/profiling Agility.

There is not much reason to believe that the results can be generalized over time because the studied surveys may evolve and new surveys will be introduced to support Agility assessment.

Finally, the differences in the results achieved by S1 and S2 might indicate that conducting the same research in different context might result in finding other surveys than S2 as the best representative of the Agile status.

7. Conclusions

In this research, surveys for assessing Agility have been studied. Although the initial searches indicated that several surveys exist for the purpose of evaluating Agile software development, they have different focuses and perform different analyses. We studied a selection of surveys based on a search on the web, and chose two of them based on the criteria described for further examination in an industrial case study, including three software development teams. The objective was to evaluate the practical usefulness of the surveys. The different strengths and weaknesses have been discussed for those found and a more in-depth comparison has been done for two of the surveys, which answers RQ1.

Based on the case study, results for the organization and the three teams were reported, and then the results were used to compare the two surveys. It is concluded that one of the surveys (S2) was perceived as portraying the Agility in the three teams and the organization better than the other survey used. This responds to RQ2.

The answer to the third research question (RQ3) is that different surveys do not assess Agility similarly and hence they are not expected to give the same scores to a specific team/organization. This is confirmed by the results from the case study and when comparing the two surveys used and studied in more detail. In summary, it means that Agile is not necessarily assessed in the same way and hence Agility becomes context-dependent and it could be practiced in different way. Each user of Agile methods or practices must decide what Agile means to them, and hence select the best assessment survey in relation to their chosen definition of Agile.

Rapid changes in the market, customer demands, and technology push software organizations towards more Agility, which means if Agile is not applied properly, it could introduce difficulties in the organization instead of flexibility. Therefore, some way of assessing and measuring Agility would be helpful. However, besides the need for assessing/profiling Agile, there is not one clear answer how it should be assessed.

We would like to recommend open discussions of the evaluation results with all team members and managers to prioritize the practices that are critical for the organization (e.g. are in alignment with the organizational goals). This implies a selective approach in adopting/improving Agility rather than encouraging being perfectly Agile.

Validating S2 in a separate case study on a different team is an item for future research directions. In addition, we would like to examine possible approaches of combining product goals and an Agile assessment survey.

Acknowledgements

This work was partly funded by the Industrial Excellence Center EASE - Embedded Applications Software Engineering, (http://ease.cs.lth.se).

References

P. Abrahamsson, O. Salo, J. Ronkainen, J. Warsta (2002): "Agile Software Development Methods: Review and Analysis", VTT Publication, Finland.

P. Abrahamsson, J. Warsta, M. T. Siponen, J. Ronkainen (2003): "New directions on agile methods: a comparative analysis", *Proceedings of the 25th International Conference on Software Engineering*, ACM Press, pp. 244-254.

K. Beck et al. (2001): "Manifesto for Agile Software Development", Available: http://www.agilemanifesto.org, Accessed 2014-06-29.

B. Boehm, R. Turner (2004): "Balancing agility and discipline: a guide for the perplexed", Addison-Wesley.

I. Bose (2008): "Lessons learned from distributed agile software projects: a case-based analysis", Communications of the Association for Information Systems 23(34), pp. 619-632.

J. Chen, J. Shao (2000): "Nearest neighbor imputation for survey data", Journal of Official Statistics-Stockholm 16(2), pp. 113-132.

CMI Lean Agility (2009): "CMI Lean Agility Assessment", Available: http://cimes.promes.be, Accessed 2014-06-29.

M. Cohn, K. Rubin (2007): "Comparative Agility", Available: http://comparativeagility.com, Accessed 2014-06-29.

K. Conboy, B. Fitzgerald (2004): "Toward a conceptual framework of agile methods: a study of agility in different disciplines", *Proceedings of XP/Agile Universe*, Springer Verlag, pp. 37-44.

J. W. Creswell (2003): "Research design: qualitative, quantitative, and mixed method approaches", Second Edition, SAGE, ISBN: 0761924426, 9780761924425.

G. Dinwiddie (2009): "DIY Project/Process Evaluation Kit", Available: http://blog.gdinwiddie.com/2009/08/18/diy-projectprocess-evaluation-kit, Accessed 2014-06-29.

T. Dybå, T. Dingsøyr (2008): "Empirical studies of agile software development: a systematic review", Journal of Information and Software Technology 50, pp. 833-859.

E. Germain, P. Robillard, "Engineering-based Processes and Agile Methodologies for Software Development: a Comparative Case Study", *The Journal of Systems and Software*, Elsevier, February 2005, pp. 17-27.

L. Guion (2002): "Triangulation: establishing the validity of qualitative studies", University of Florida Extension, Institute of Food and Agricultural Sciences.

D. Hartmann, R. Dymond (2006): "Appropriate agile measurements: using metrics and diagnostics to deliver business value", *Agile 2006 Conference*, pp. 126-131.

E. Hossain, M. Ali Babar, J. Verner (2009): "Towards a framework for using agile approaches in global software development", Product-Focused Software Process Improvement, pp. 126-140.

International Organization for Standardization (2012): ISO/IEC 15504-5:2012, Information technology - Process assessment - Part 5: An exemplar software life cycle process assessment model, Available:

http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=60555, Accessed 2014-06-29.

M. James (2007): “A Scrum Master’s Checklist”, Available: http://blogs.collab.net/agile/2007/08/13/a-scrummasters-checklist, Accessed 2014-06-29.

A. Kangas, S. Laukkanen, J. Kangas (2006): “Social choice theory and its applications in sustainable forest management-a review”, Forest Policy and Economics 9(1), pp. 77-92.

H. Kniberg (2010): “Scrum Checklist”, Available: http://www.crisp.se/scrum/checklist, Accessed 2014-06-29.

J. L. G. Lapresta, M. M. Panero (2002): “Borda Count Versus Approval Voting: A Fuzzy Approach”, Public Choice 112(1-2), pp. 167-184.

L. A. Marascuilo, R. C. Serlin (1998): “Statistical Methods for the Social and Behavioral Sciences”, W. H. Freeman and Company.

Mayberg Consulting (2008): “Agile Karlskrona Test”, Available: http://mayberg.se/learning/karlskrona-test-online, Accessed 2014-06-29.

F. McCaffery, M. Pikkarainen, I. Richardson (2008): “Ahaa -agile, hybrid assessment method for automotive, safety critical SMEs”, *Proceedings of International Conference on Software Engineering (ICSE 2008)*, Leipzig, Germany.

M. C. Paulk, B. Curtis, M. B. Chrissis, C. V. Weber (1993): “Capability Maturity Model for Software (Version 1.1)”, Technical Report, Pittsburgh, PA: Software Engineering Institute, Carnegie Mellon University, Available: https://resources.sei.cmu.edu/asset_files/TechnicalReport/1993_005_001_16211.pdf, Accessed 2014-06-29.

K. Petersen, C. Wohlin (2009): “Context in industrial software engineering research”, *3rd International Symposium on Empirical Software Engineering and Measurement*, pp. 401-404.

R. Prikladnicki, J. L. N. Audy, D. Damian, T. C. de Oliveira (2007): “Distributed software development: practices and challenges in different business strategies of offshoring and onshoring”, *Proceedings of the IEEE International Conference on Global Software Engineering (ICGSE)*, pp. 262-274.

A. Qumer, B. Henderson-Sellers (2006): “Comparative evaluation of XP and scrum using the 4D analytical tool (4-DAT)”, *Proceedings of the European and Mediterranean Conference on Information Systems (EMCIS)*, Spain.

P. Runeson, M. Höst (2009): “Guidelines for conducting and reporting case study research in software engineering”, Empirical Software Engineering 14, pp. 131-164.

Santeon Group (2012): “Dr Agile Agile Assessment”, Available: http://www.dragile.com, Accessed 2014-06-29.

A. Sidky (2007): “A structured approach to adopting agile practices: the agile adoption framework”, Ph.D. Dissertation, Computer Science, Virginia Tech, Blacksburg.

Signet Research and Consulting (2005): “A quick self-assessment of your organization’s agility”, Available: http://signetconsulting.com, Accessed 2014-06-29.

Software Engineering Institute (2011): “CMMI Version 1.3 Information Center”, Software Engineering Institute, Available: http://www.sei.cmu.edu/cmmi/index.cfm, Accessed 2014-06-29.

S. Soundararajan, J. D. Arthur (2011): “A structured framework for assessing the goodness of agile methods”, *18th IEEE International Conference and Workshops on Engineering of Computer Based Systems (ECBS),* Las Vegas NV, pp. 14-23.

J. Sutherland (2009): The Nokia Test, Available: http://antoine.vernois.net/scrumbut/?page=intro&lang=en, Accessed 2014-06-29.

M. Taromirad, R. Ramsin (2009): “CEFAM: comprehensive evaluation framework for agile methodologies”, *32nd Annual IEEE Software Engineering Workshop (SEW '08)*, Kassandra, Greece, pp. 195-204.

Thoughtworks Studio (2010a): “Agile Self Evaluation”, Available: http://www.agileassessments.com/online-assessments/agile-self-evaluation, Accessed 2014-06-29.

Thoughtworks Studio (2010b): “Build and Release Management Assessment”, Available: http://www.agileassessments.com/online-assessments/brm-self-evaluation, Accessed 2014-06-29.

K. Waters (2008): “How Agile Are You? (Take This 42 Point Test)”, Available: http://www.allaboutagile.com/how-agile-are-you-take-this-42-point-test, Accessed 2014-06-29.

L. Williams, A. Cockburn (2003): “Agile software development: it’s about feedback and change”, IEEE Computer 36 (6), pp. 39-43.

L. Williams, W. Kerbs, L. Layman, A. I. Anton, P. Abrahamsson (2004): ‘Toward a framework for evaluating extreme programming”, *8th International Conference on Empirical Assessment in Software Engineering*, pp. 11-20.

L. Williams, K. Rubin, M. Cohn (2010): "Driving process improvement via comparative agility assessment", *AGILE Conference*, pp. 3-10.

A. Yauch (2011): "Measuring agility as a performance outcome", Journal of Manufacturing Technology Management 22(3), pp. 384-404.

Appendix 1

A Sample of Survey Questions

S1 - Question	Answer	Comments
How is work assigned?	Sure	
a) People are given specific tasks to perform (coding, analysis, etc.) by leads / managers.	☐	
b) People choose what they are going to work on from a backlog.	☒	
Which of the following most closely describes the ratio of business analysts to developers within your organization?	More sure than unsure	
a) 1 business analyst to 4 or fewer developers.	☐	
b) 1 business analyst to 7 or fewer developers.	☒	
c) 1 business analyst to 8 or more developers.	☐	
…		

S2 - Question	Answer	Sure	Comments
Team members are kept together as long as possible.	True	Sure	
Testers and programmers are on the same team.	True	Sure	
Teams have 5-9 people on them.	True	Sure	
Teams can determine who is on or off the team.	False	Unsure	
…			

Appendix 2
A Sample of Interview Questions

Teamwork

How is the team built? Composed? Located? Who does it?
How do the team work together?
Who decides on priorities or changes them?
Standup meetings? Duration?
…

Requirements

Who is product owner and how collaborates with the team during an iteration?
How is the requirement handling and agreement process?
…

Planning

When and who does the technical design?
How often the team updates the iteration burn down charts?
…

Technical Practices

Is TDD applied? PP? Refactoring? Etc.
…

Quality

Is unit testing done before checking in the code?
What type of testing is performed for iteration?
…

Culture

Is productivity on the focus or the overwork?
…

Knowledge Creating

How good is the team's knowledge to Agile?
Who is present in retrospective meetings?