CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

# Sampling Affects of Software Developers to Understand Individual & Team Performance

Master's thesis in Software Engineering and Technology

KEVIN HEDBERG GRIFFITH & ERIK NGUYEN

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

Master's thesis 2018

# Sampling Affects of Software Developers to Understand Individual & Team Performance

KEVIN HEDBERG GRIFFITH
ERIK NGUYEN



Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2018

Sampling Affects of Software Developers to Understand Individual & Team Performance
KEVIN HEDBERG GRIFFITH
ERIK NGUYEN



Supervisor: Robert Feldt, Department of Computer Science and Engineering
Advisor: Johan Lundberg, Volvo Car Retail Solutions
Examiner: Jan-Philipp Steghöfer, Department of Computer Science and Engineering

Master's Thesis 2018
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2018

Sampling Affects of Software Developers to Understand Individual & Team Performance
KEVIN HEDBERG GRIFFITH & ERIK NGUYEN
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

# Abstract

**Background** Software development is human-centred and consists of intellectual activities and teamwork which requires people skills. However, work towards improving individual and team performance has put much effort into improving technology and processes, and less on human factors such as moods or feelings. Affect can be used as an umbrella term for moods and emotions, and this thesis adheres to the dimensional approach of affects, in which valence, arousal, and dominance describes people's state of feeling.

**Objective** The purpose of this thesis is to understand the impacts of valence, arousal, and dominance on individual and team performance of software developers.

**Method** Experience sampling method (ESM) was chosen for collecting data on affects as it was created to study what people do, feel and think in their natural settings. Data on affects and individual performance was collected with ESM and analysed using manual interpretation and linear mixed-effects model (LMM). Team performance data was gathered using self-assessment surveys and compared with the affect results from the ESM study. Data were analysed using manual interpretation and Kendall's tau-b correlation. The study was conducted in an industrial setting consisting of 28 developers in 4 teams from Volvo Car Retail Solutions.

**Results** Results showed that valence have a significant impact on individual performance. For team performance, manual interpretation indicated a close relationship between valence and team performance.

**Conclusions** We demonstrated how performance and human factors could be measured, and showed how data could be analysed using a LMM, manual interpretation, and Kendall's correlation. Results showed a significant correlation between affects and performance. The results introduces a new perspective by including both individual and team performance in an industrial setting.

Keywords: Software engineering, affects, experience sampling method, team performance, individual performance.

# Acknowledgements

Completing this thesis could not have been without outside help, hence, we have a lot of people to thank. We begin with showing our appreciation to our university supervisor Robert Feldt, who has, with his expertise in the thesis subject, provided valuable feedback and guidance. We would also like to thank our supervisor at VCRS, Johan Lundberg, for always being available for endless discussions and providing us with a place to be.

Further, we would like to give many thanks to Per Lenberg for helping us during the early stages of the thesis, as well as Daniel Graziotin for providing feedback and guiding us in the right direction.

Special thanks to Andreas Bäckevik and Erik Tholén for providing valuable feedback to improve the quality of our thesis.

Thanks to Louis at Expimetrics for making an exception and providing us with a great offer for their service.

Finally, a special thanks to everyone that participated in the study.

Kevin Hedberg Griffith & Erik Nguyen, Gothenburg, June 2018

# Contents

# List of Figures

# List of Tables

# 1
# Introduction

In the context of software development, you rarely come across an organisation that is not composed of project teams. Some organisations consist of great teams, and then there are organisations with teams that are struggling. High-performance teams are cooperative processes of people that achieve extraordinary results (Scarnati, 2001), which are what any organisation strive for. The work towards improving team performance in companies has put much effort in improving technologies and processes, and less on human factors (John, Maurer, & Tessem, 2005). Regardless of latest technology or most efficient processes, software development is human-centred (Hazzan & Hadar, 2008; Amrit, Daneva, & Damian, 2014), and consists of intellectual activities and teamwork which requires people skills. One may have heard the expression 'people trump process', and one cornerstone of the Agile Manifesto is 'Individuals and interactions over processes and tools.'("Agile Manifesto", 2001).

Emotions and moods are commonly used terms of human factors. The term affect (or affective state) is commonly associated with moods and emotions. Further, the term affect can be used as an umbrella term for emotions and moods (Graziotin, Wang, & Abrahamsson, 2014, 2015a). Hence, emotions and moods will be treated as interchangeable terms, and affects will be measured and used as a collective term for emotions and moods. Graziotin, Wang, and Abrahamsson (2015b) presents two major frameworks for affect theories; discrete and dimensional.

The dimensional approach will be applied in this thesis. It is composed of the three dimensions; valence, arousal, and dominance, where valence is the attractiveness of an event, arousal is the feeling of being excited or calm, and dominance is concerned with feelings of control. Hence, these three dimensions will be measured and used to understand the impact of affects on individual and team performance. Team performance will be measured by using self-assessment on team members. The self-assessment data will further be supported by an additional team self-assessment, and by having managers assess the teams. Measuring individual performance will be done by solely using self-assessment.

This thesis will be conducted together with Volvo Car Retail Solutions (VCRS), which is a wholly owned subsidiary of Volvo Car Sverige AB and works with Volvo retailers across Sweden, Norway and Japan. Their primary goal is to make automotive retail easy by providing software solutions to both customers and retailers. They are located in Gothenburg, Sweden with approximate 200 co-workers. The organisation consist of teams of 4-10 members, with each applying the approaches

and principles of agile development. The proposed target group to study will be members of software development teams working at VCRS.

## 1.1 Background

Human factors are drivers of team performance and companies acknowledge the impact it has on teams. It is well known and concluded that human factors indeed impacts team performance (Graziotin, Wang, & Abrahamsson, 2013; Khan, Brinkman, & Hierons, 2011; Lenberg & Feldt, 2018). DeMarco and Lister (2013) have made significant work regarding human aspects in software engineering and acknowledge the importance of human factors over technical factors in successful projects. Feldt, Torkar, Angelis, and Samuelsson (2008) argues that researchers in the field of software engineering should focus more on aspects regarding humans as software development is a human-centred activity. The researchers proposed a method to measure personality, attitudes, motivations and emotions. Furthermore, Lenberg, Feldt, and Wallgren (2014) have identified the need for human-focused software engineering research and proposed the research area behavioural software engineering (BSE) 'as an umbrella concept for research that focuses on behavioural and social aspects in the work activities of software engineers.'. BSE defines the study of cognitive, behavioural and social aspects of software engineering performed by individuals, groups or organisations (Lenberg, Feldt, & Wallgren, 2015). It further defines the research area of human factors in software engineering and will shine more light on that area.

High-performing individuals and teams are often contributing factors to company success. Understanding performance in software engineering is essential, but measuring it is a very complicated task. There exists no clear framework for how to measure team performance (Goodman, Ravlin, & Schminke, 1987) or individual performance and studies have used various kind of measurement methods. For instance, measuring team performance has been done by using objective measurements (Downey & Sutherland, 2013), or having team members self-assess their abilities as a team (Lenberg & Feldt, 2018; Google, 2016). Measuring individual performance can be done by using objective measurements and self-assessments (Graziotin et al., 2015a), or by having team members assess the person in question.

## 1.2 Purpose

The purpose of this thesis is to understand the affects of team-oriented software developers and see how these correlate to individual and team performance. To the authors' knowledge, there is a lack of research in software engineering on this. For instance, Dutra, Prikladnicki, and França (2015) stated that the quality of research of software teams needs to improve. Of the few research studies on human factors in software engineering, Graziotin et al.; Khan et al.; Lenberg and Feldt (2013; 2011; 2018) provide results that emphasise the impact of human factors. This thesis is set

out to both contribute to previous research on the topic as well as deliver valuable results for IT-organisations.

With regards to IT-organisations, the findings of this thesis will help bring increased awareness and knowledge of the importance of human factors. Since organisations highly depend on the success of their teams, it is essential to understand and identify the motive that impedes a team from performing well and act accordingly.

## 1.3 Problem statement

Regardless of the field or task that you are working on, insurance of your team is performing at their best is desired. To achieve well performing teams, and to improve, one must have a holistic understanding of what performance is. In the field of software engineering, there is yet to exist a definite answer on how performance is defined and measured. To the authors' knowledge, many articles and research papers on performance improvements tend to focus on tools, methods and processes. While this may be the more popular and well-practised approach to take, the other side of the coin is being omitted, i.e. human factors. Hence, there is a need for studying human factors, and Lenberg et al. (2014) have presented the research field BSE to fill this gap. It is in fact, the developers that are at the core of a team, and they are also the depending factor that determines the outcome. Graziotin et al. (2014) considered their study to be the first on examining the correlation of affects and performance of software developers working in natural settings on real-world software problems. They further encouraged future studies to aim at finding suitable measurement intervals for more extended sessions, such as the duration of an iteration.

In recent years, researchers have conducted more studies on performance with human factors at the centrepiece. This thesis will also focus on human factors and is set out to contribute to previous research by providing further findings and results.

## 1.4 Research questions

As aforementioned, this thesis is set out to look at how affects impacts team performance. With that being said, the main research question that we aim to answer is:

What impacts do affects have on performance in software teams and on individual developers?

To answer this question, the following hypotheses have been developed:

$H_{01}$: The valence affective of a software developer has no impact on self-assessed team performance

$H_{11}$: The valence affective state of a software developer impacts self-assessed team performance

$H_{02}$: The valence affective state of a software developer has no impact on self-assessed individual performance
$H_{12}$: The valence affective state of a software developer impacts self-assessed individual performance

$H_{03}$: The arousal affective state of a software developer has no impact on self-assessed team performance
$H_{13}$: The arousal affective state of a software developer impacts self-assessed team performance

$H_{04}$: The arousal affective state of a software developer has no impact on self-assessed individual performance
$H_{14}$: The arousal affective state of a software developer impacts self-assessed individual performance

$H_{05}$: The dominance affective state of a software developer has no impact on self-assessed team performance
$H_{15}$: The dominance affective state of a software developer impacts self-assessed team performance

$H_{06}$: The dominance affective state of a software developer has no impact on self-assessed individual performance
$H_{16}$: The dominance affective state of a software developer impacts self-assessed individual performance

## 1.5 Scope & limitations

The scope of this thesis is broad and covers many aspects regarding affects, ESM, individual and team performance. However, limitations were set to achieve the goal of this thesis. Underlying causes that impact participants affects will not be taken into consideration, for instance, personal issues, events, and paydays. Tuckman (1977) describes four group stages of maturity that impact team performance depending on how mature the group is, and hence, group maturity may have a significant impact. Gren, Torkar, and Feldt (2017) acknowledges the impact of group maturity as many aspects of it correlates to agile development teams. Although group maturity is essential, it is disregarded in this thesis. Participant information presented exists solely to give the reader a sense of the participant characteristics and will not be taken into account when analysing individuals and teams.

## 1.6 Significance of thesis

When studying the performance of individuals and teams of software engineering, it is very seldom that you consider human factors. Studies on human factors are progressing over the years, which is an indication of the importance of further studies. Graziotin, Abrahamsson and Wang (Graziotin et al., 2014, 2015a, 2015b), Berkel, Ferreira and Kostakos (Berkel, Ferreira, & Kostakos, 2017), and Feldt, Lenberg and

Wallgren (Lenberg & Feldt, 2018; Lenberg et al., 2015, 2014) are among the likes of researchers that, in recent years, have provided studies on human factors.

We have yet to come across research that studies the impact of human factors on both individual and team performance level. Instead, studies have considered sub-parts of performance, human factors or both, e.g. feelings on productivity (Graziotin et al., 2014), happiness on debug performance (Khan et al., 2011) and motivation on developers (Hall, Sharp, Beecham, Baddoo, & Robinson, 2008). This thesis provides a unique perspective as it will look at how performance is measured as a whole (for both individuals and teams), and then studying the impacts of affects.

Furthermore, this thesis is significant because it is conducted with participants from a software development company in an industrial setting which increases the external validity, whereas many studies analyse students in an academic environment.

## 1.7 Structure of the article

Chapter 2, Related work, presents research that are related to the goal of this thesis, i.e. human factors and how it correlates with performance.

Chapter 3, Theory, provides the reader with relevant background information that is necessary to understand various techniques, methods and concepts that the thesis has applied to.

Chapter 4, Method, is divided into different sections to describes the steps taken to be able to find an answer to the research question.

Chapter 5, Result, presents key results derived from the surveys and the ESM study.

Chapter 6, Discussion, provides the authors' interpretation of the results based on an analysis of the results.

Chapter 7, Conclusion, complete the thesis and summarises key findings.

Chapter 8, Future work, provides suggestions for further research.

# 2
# Related work

This chapter presents research related to human factors and its impact on individuals and teams.

Psychology of Programming (PoP) is a research area that concerns developers cognition, tools and methods for programming related activities and programming education (Sajaniemi, 2008). PoP emerged when researchers understood that evaluating tools and technologies should not entirely be based on computational power as a human point of view is essential to consider. PoP date back to 1960s when most of human-centred software engineering research was regarded as 'exception rather than the rule', and a majority of the studies conducted were focused on technical aspects of programming (Hoc, Green, Samurçay, & Gilmore, 1990). Weinberg (1971) did pioneering work regarding human aspects of software engineering and argued that any matters regarding programming should include psychological viewpoints. The importance of human aspects was also acknowledged by DeMarco and Lister (1987) as the authors realised that human factors matter much more than technical factors. Pew, Rollins, and Williams (1976) and Newman (1977) outlined guidelines regarding human factors and system design in the 1970s. However, Rudy Ramsey and Atwood (1979) argues that research in the area was insufficient and existing literature was 'badly fragmented' due to its foundation in other disciplines that were not regarding computer systems. Therefore, such guidelines are limited. However, since then, researchers have acknowledged and focused their studies more on human aspects in software engineering, e.g. Lenberg et al. (2014) defines and propose the research area BSE which focuses on behavioural and social aspects of software engineers.

Google's study on high-performance teams concluded that it is more important how the team work together rather than who is on the team. Psychological safety, dependability, structure and clarity, meaning, and impact are factors that envision team effectiveness, with psychological safety being far most important (Google, 2016). Participants in the study performed double-blinded interviews and were asked about factors that might impact team effectiveness, such as group dynamics, skill sets, personality traits and emotional intelligence. Being able to speak one's mind and not be ridiculed in teams (high psychological safety) results in that individuals are less likely to leave the company and more likely to bring in more revenue and be more effective than teams with low psychological safety. Additionally, Safdar, Badir, and Afsar (2017) conducted a study on psychological safety among new development team members and concluded that individuals with high psychological safety were more likely to seek advice from their team members.

Lenberg and Feldt (2018) conducted a study on the psychological safety and team norm clarity, and its impact on team performance and job satisfaction. Their research was conducted by collecting data from 217 participants in 38 teams from 5 different organisations using surveys. They showed that psychological safety indeed has an impact on team performance and individual job satisfaction, but team norm clarity is a stronger predictor of both. How to act in different situations and understanding the expected team member behaviour leads to an environment that reduces uncertainty. The importance of team norm is further noticed by Acuña, Gómez, and Juristo (2008) which conducted a study about the correlation between team climate and software quality by examining 35 random selected three-member-teams of students who worked with software development. The researchers wanted to find out how comfort level of a team member, depending on how the team climate matches his/her preferences, relates to software quality. They discovered that high participant safety within the team and team vision are linked to better software quality; however, the two effects could not be separated. The researchers stretched the importance to track team climate or well-being as it has an impact on performance and software quality.

Collective positive emotions impact team resilience, according to Meneghel, Salanova, and Martínez (2016). The authors evaluated the emotions enthusiasm, optimism, satisfaction, comfort, and relaxation and additionally team resilience. Assessments from supervisors measured team performance, and the researchers had a sample size of 1076 employees in 216 teams from 40 companies. Meneghel et al. (2016) concluded that the evaluated emotions were positively related to team resilience, and team resilience impacts team performance. It was further stated that it is essential for teams to have positive collective emotions regarding enthusiasm, optimism, satisfaction, comfort, and relaxation to increase team resilience and performance.

Happiness is proven to have a positive impact on problem-solving abilities, and in a study by Graziotin et al. (2015a) it was stated that the happiest software developers are significantly better analytical problem solvers. The study was conducted by measuring participants affects before completing a creativity task and an analytic problem-solving task. During creativity task, the participants were asked to write the best caption they could come up with for two photographs. In the problem-solving task, participants were asked to complete the Tower of London (or Shallice test). Each task lasted for 30 minutes, and the interviews for measuring affects took less than 10 minutes to complete. The study had 42 participants who were computer science students with diverse nationality. The results from the data showed that happiness has a positive impact on problem-solving abilities. However, it could not prove that affects impacts creativity. The researchers stated an increased need for further research on affects of software developers and perceived their research as a step towards validating that people are more important than processes and tools.

Khan et al. (2011) conducted a study on how mood impacts the performance of programming debugging tasks. The researchers concluded, after two separate studies, that moods indeed do influence debugging tasks. In the first study, arousal showed to have a significant impact on debugging performance while valence did not. The

second study showed that an increase in arousal and valence 'coincided with an improvement in programmers' task performance', however, the effects could not be separated. This study paves the path for further studies on impacts of mood and the researchers stated that their study "could be regarded as a first step in developing a deeper understanding".

Graziotin et al. (2014) conducted a study to analyse the affect dimensions valence, arousal and dominance of software developers. Participants were asked to complete a pre-task interview, a software development task, and a post-task interview. The software development task was observed during 90 min, and the participant had to fill in a short survey each 10 min. Results showed that both valence and dominance impacts self-assessed productivity positively, with 35% of the deviance explained. However, the results could not prove that arousal had any effect on self-assessed productivity. The researchers stated that the participants were misunderstanding the arousal dimension because there were many questions about it in the survey explanations.

# 3
# Theory

The following sections will provide information about the three central pillars of the thesis; measuring performance, affects and the experience sampling method. This chapter provides a clear overview of the three topics and will help you comprehend the remaining part of the thesis.

## 3.1 Measuring performance

Today, many teams adhere to agile practices where performance indicators often are based on objective measurements such as time, cost and quality. In companies with agile practices, it is common to use Scrum in the development process. The usage of Scrum comes with a variety of available objective measurements that can be used to measure individual and team performance. Downey and Sutherland (2013) argues that a combination of objective measurements in Scrum such as velocity, work capacity, focus factor, the accuracy of estimation and accuracy of forecasts can be accurately used to measure team performance and compare between teams.

Regardless of the use of objective measurements, Hariharan and Arpasuteerat (2017) concluded that it could not provide a holistic view of team performance unless considering human factors, which is also stated by Hackman (1987). Perceived team effectiveness is something that emphasis on both internal and external criteria. Internal criteria are human factors (e.g. member satisfaction and team viability), whereas external criteria are objective measurements (e.g. productivity and performance) (Hackman, 1987). When Google (2016) conducted their study on team performance, a combination of both quantitative and qualitative assessment was used to understand team effectiveness. “Google’s leaders, who had initially pushed for objective effectiveness measures, realised that every suggested measure could be inherently flawed – more lines of code aren’t necessarily a good thing, and more bugs fixed means that more bugs were initially created.”. Lenberg and Feldt (2018) suggests using a combination of objective measurements and self-assessments to raise the validity of the data. As for individual performance, Graziotin et al. (2015a) conducted a study on the impacts of affects on self-assessed productivity. In their research, individual performance was measured by having participants self-assess their productivity with ESM. In this thesis, individual and team performance will be measured by solely using self-assessment as it is the most feasible method to use in VCRS’s company setting. Further, team performance self-assessment will adhere to Hoegl and Gemuenden (2001) statement about the importance of including multiple viewpoints.

## 3.2 Affects

Emotions and moods are commonly used terms of human factors. In previous literature in the field of psychology, one will find many different meanings of emotions and moods, as researchers have yet to come to a consensus on the definition of the terms. According to Ekman (2003) and Frijda (1993) moods are affective states that, in comparison to emotions, last for an extended period. Frijda (1993), further explained emotions as being intense, with a definite object or cause, whereas moods are of weaker states of uncertain origin, and Kleinginna and Kleinginna (1981) identified the existence of nearly a hundred of definitions for the term emotion. However, Weiss and Cropanzano (1996) stated that the need for a clear distinction between the terms are not always necessary, and further literature has considered using the terms interchangeably (Baas, De Dreu, & Nijstad, 2008; Schwarz & Clore, 1983; Schwarz, 1990; Wegge, van Dick, Fisher, West, & Dawson, 2006). As mentioned in section 1, emotions and moods will be treated as interchangeable terms, and affects will be measured and used as a collective term emotions and moods.

### 3.2.1 Measuring affects

To the authors' knowledge, researchers have yet to agree on a common metric for assessing affects. Hence, the following presents approaches that have been used throughout the years to measure affects.

A lot of frameworks exists for measuring affects (mood and emotions). However, Huang (2001) concluded the existence of the following four major theories:

*Differential Emotions Theory*
Based on ten emotions (7 negative, 2 positive, and 1 neutral) to constitute the human motivational system (Izard, 1977).

*Circular Model of Emotion*
consisting of eight primary emotions (4 negative, 2 positive, and 2 neutral), where all other emotions were considered as mixtures of the primary ones (Plutchik, 1980).

*PAD Model of Affect*
Built on three dimensions pleasure, arousal, and dominance, which included moods, feelings, and any other feeling-related concepts (Mehrabian & Russell, 1974).

*PANAS*
Designed to present a mood scale with positive and negative affects as the primary dimensions (Watson & Clark, 1992; Watson & Tellegen, 1985).

Despite the claims made by Huang (2001) on the existence of four major theories, a more recent study conducted by Graziotin et al. (2015b) presented only two major frameworks; *discrete* and *dimensional*. The *discrete* approach refers to a collection of basic affective states that can be distinguished uniquely (Plutchik & Kellerman, 1980), e.g. happiness, sadness, surprise, fear, anger, and disgust (Ekman, 1971). The other approach, *dimensional*, refers to three independent emotional dimension;

valence, arousal, and dominance, which describe people's state of feelings (Mehrabian & Russell, 1974). Sánchez, Kirschning, Palacio, and Ostróvskaya (2005) describes valence (or pleasure) as a subjective measure ranging from unpleasantness to pleasantness, including adjectives such as happy - unhappy, pleased - annoyed, and satisfied - unsatisfied (Mehrabian & Russell, 1974). *Valence* can further be explained as the attractiveness (or adverseness) of an event, object, or situation (Lewin, 1935; Lang, Greenwald, Bradley, & Hamm, 1993). *Arousal* can be described as a subjective state of feeling activated or deactivated (Sánchez et al., 2005). It represents the intensity of emotional activation (Lane, Chua, & Dolan, 1999). *Dominance* is related to feelings of control and the extent to which an individual feels restricted in his behaviour, including adjectives such as controlling, influential and autonomous (Mehrabian & Russell, 1974). It is the sensation by which the individual's skills are higher than the challenge level for a task (Csikszentmihalyi, 1997).

A comparison between discrete and dimensional models were conducted by Eerola and Vuoskoski (2011). The result concluded that the overall consistencies between emotion ratings in the dimensional and discrete models did not display any significant differences. However, in comparison to the dimensional model, the discrete model was shown to be less reliable in rating excerpts that were ambiguous examples of an emotion category.

Because no differences were found between the two approaches, dimensional and discrete, the selection of which to chose was based on what tool that was selected (section 3.2.2).

### 3.2.2 Self-assessment manakin (SAM)

In year 1994, Bradley and Lang (1994) developed the Self-Assessment Manikin (SAM) and described it as a non-verbal pictorial assessment technique that directly measures the pleasure, arousal, and dominance associated with a person's affective reaction to a wide variety of stimuli. It has proven to be popular in recent years as it has been successfully used in many studies (Backs, da Silva, & Han, 2007; Betella & Verschure, 2016; Imbir, 2016; Graziotin et al., 2013, 2014). SAM follows the PAD theory as it is composed of three separate sets of figures (figure 3.1), where a range from frowning, unhappy to smiling, happy represents valence, arousal ranges from relaxed, sleepy to excited, wide-eyed, and the size of the figure represents dominance. As the saying goes, a picture is worth a thousand words, which can be likened to SAM in its way to use images, as opposed to written text, to develop a survey that is very easy to understand and use.

Betella and Verschure (2016) identified a problem with SAM in that participants frequently asked for further clarification on the meaning of the figures despite having received official rating instructions. Hence, when conducting SAM, clear instructions and descriptions must be provided to the participants to mitigate possible misinterpretations. Betella and Verschure (2016) further states that the problem might be in that the paper-and-pencil approach upon which SAM was based on were obsolete and did not match up to present tools and technologies. Hence, they designed a new

digital scale called the 'Affective Slider'. Composed of two slider controls, the new design presented two advantages compared to SAM in which no written instructions were needed and the simplicity to reproduce it in digital devices such as smartphones and tablets.

Despite the 'Affective Slider' showing characteristics proven to be an improved version of SAM, it was not chosen for measuring affects. No other study had been conducted using this approach, and hence, the risks were considered too high.

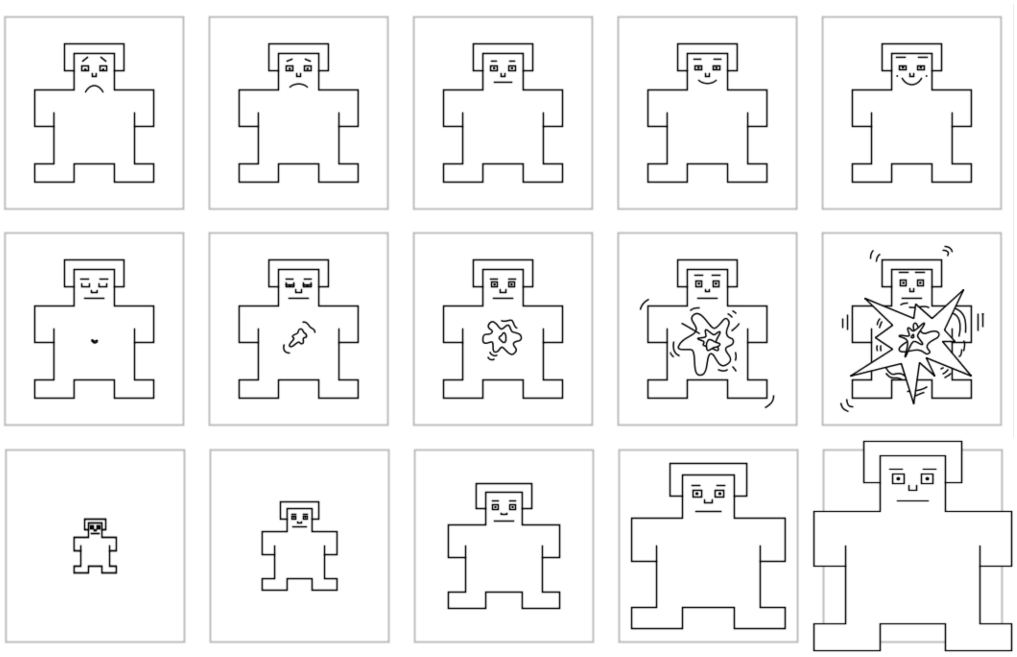

**Figure 3.1:** The Self-Assessment Manikin based on Bradley and Lang (1994) with images derived from PXLab.

## 3.3 Experience sampling method (ESM)

Larson and Csikszentmihalyi (2014) explained the Experience Sampling Method (ESM) as a research procedure to study what people do, feel, and think during their daily lives by having them provide self-reports throughout their daily lives multiple times a day. Further, Fisher and To (2012) explains ESM as a method for collecting data about people's current or very recent affect, behaviour and thoughts. As aforementioned, Bradburn, Rips, and Shevell (1987) identified a problem with recall bias and recall loss when making assessment sometime after an event, but by conducting an ESM study, the impact of recall problems will be mitigated because of the short time between the signal to provide self-report and the response (Scollon, Kim-Prieto, & Diener, 2003). Furthermore, Scollon et al. (2003) stated other promises that come with the use of ESM, such as the possibility to analyse behaviour contingencies and increase the ecological validity.

The interest in studying people's day-to-day activities can be found in traditional

daily diary (DD) literature studies dating back to the early 1900s by Bevans (1913) and Altshuller (1923). During the 1990s, both ESM and DD drew a lot of attention in health, clinical, and social psychology (Fisher & To, 2012). ESM and DD have their similarities in their way of gathering data through self-reports provided by participants. Over time, data from these self-reports are merged into one accumulated report making it is possible to distinguish the old behaviours and feelings of a single participant. With DD, researchers usually ask participants to report only once per day, whereas with ESM, participants are asked to provide multiple self-reports during per day. Further, ESM excels in its ability to gather data directly from a participant's natural setting at the very moment of answering a self-report. The use of observational methods introduces a problem in that participant's are taken away from their natural setting, which in turn, can skew the gathered data Bolger and Laurenceau (2013). This problem does not exist with ESM.

As a consequence of increased awareness and popularity around psychological factors in recent years, there is a notable increase in research that has been conducted using ESM. As ESM was created to study what people do, feel and think, this thesis will use ESM and adhere to its best practices for conducting a successful data gathering process.

### 3.3.1 ESM on mobile devices

Raento, Oulasvirta, and Eagle (2009) stated that the availability of personal devices enables widespread deployment of mobile phones as a research tool. Hence, in this thesis, a more modern approach will be applied by conducting a smartphone-based ESM study. The original approach for carrying out ESM self-reports has been through electronic pagers that would signal participants accordingly to a random schedule. This signal was a cue to complete a self-report survey that asked about their experience at that moment of time (Larson & Csikszentmihalyi, 2014). The surveys were in paper forms which implied that the participants had to bring these forms with them at all time. However, as we are shifting towards the use of mobile devices over traditional pen-and-paper, we are seeing more ESM studies being carried out on smartphones (Berkel et al., 2017). Further, Weber, Voit, Kratzer, and Henze (2012) conducted a study on notifications in multi-device environments and concluded that smartphones (followed by smartwatch, PC and tablet), in fact, is the preferred device on which to receive notifications. Considering that the targeted group for this thesis are software teams, you would assume everyone to own a smartphone, or at least be familiar with such a device. The following list briefly presents some of the advantages of smartphone-based ESM studies:

- Real-time study status, researchers can receive and analyse study data in real-time (Berkel et al., 2017).
- Advanced question logic, questions can depend on previous input from the participant or the participant's current context (Berkel et al., 2017).

These technological advances have given rise to new possibilities for the ESM (Barrett & Barrett, 2001). In a recent study on ESM on mobile devices, Berkel et al.

(2017) conducted an extensive and systematic analysis based on a total of 110 papers in which they, among other things, identified common study parameters of ESM. The findings provide useful guidelines and recommendations as to how to conduct ESM in this thesis.

*ESM study duration*
A majority of the analysed studies (70.9%) lasted less than a month, with an average duration of 32 days, and a median of 14 days (because of high standard deviations). Key takeaways from here are that this thesis will be limited to a maximum of 1 month, which is a reasonable duration to maintain motivated participants and reduce participant burden. It is necessary to find an even balance between study duration and the number of surveys in the self-reports. If the length of a study is very long and the self-reports are time-consuming, the participants will quickly start to feel annoyed and unmotivated which in turn will most likely lead to a drastic drop in the number of completed reports as well as the quality of the answers.

*Number of participants*
Deciding on the number of participants to include in a study plays a significant role. Researchers have to find a balance that allows for enough data to be gathered while still being able to manage all of it. Statistics from the study by Berkel et al. (2017) revealed a median of 19 number of participants, with an average of 53. Therefore, we believe that the optimal setup for our thesis would be to have between 3 to 6 teams, each consisting of 4 to 9 members.

*Response rate*
In ESM, the response rate is a measure of the number of completed self-reports divided by the total number of submitted self-reports, i.e. the ratio of completed self-reports. Berkel et al. (2017) stated that a high response ratio provides a complete picture of the study, as well as indicating that collecting the data is more likely to be contextually diverse. A low response ratio can be the cause of multiple factors such as notification expiry time and low participant motivation. They further discovered that 59.1% of the 110 included papers did not report any response rate at all. For the articles that did report it, the average response rate was 69.6%. In an approach to achieve such a ratio, Adams (1963) states that participants are more willing to provide input if the costs to participation (e.g., time, energy, resources) are lower than the value of the expected outcome. Considering the costs to participants will be at centre when forming the surveys for the ESM study in this thesis.

*ESM trigger*
The ways to notify participants in an ESM is divided into three different types, signal contingent, interval contingent and event contingent. With signal contingent implying sending self-reports at randomised times throughout the day, interval contingent implying presenting self-reports following a schedule or predefined time intervals, and event contingent implying submitting self-reports when specific predefined events occur.

Berkel et al. (2017) discovered that interval contingent trigger was the most common

type of trigger, with signal contingent and event contingent coming in at a close second and third place respectively.

*Device ownership*
Whether the researchers provide the participants with study-specific mobile devices or the participants uses a personal device can have an impact on the study, with the latter becoming the more popular and favourable alternative. Asking the participants to use a mobile device provided by the researchers might take them away from their natural setting, and hence, introducing a feeling of discomfort and being in an experimental environment. "In principle, the less aware the subject is of the presence of the observing device, the less its presence should affect the study." (Raento et al., 2009).

# 4
# Methods

This chapter explains the steps that towards achieving the goal of this thesis. As a pictorial helper, figure 4.1 presents an infographic roadmap on the major steps taken. The data gathering process consisted of an ESM study along with a set of surveys, but before any work could begin on this, preparatory work had to be done. This work was divided into two parts; performance and affects, with the goal of acquiring knowledge and understanding for how to define and measure each.

Data about the participants is presented in section 4.1. Sections 4.2 and 4.3 demonstrates the process of defining and measuring performance and affects respectively. An introduction to the main tool that was used for both the ESM study and the surveys are presented in section 4.4. Further, section 4.5 presents the design of the ESM study, followed by the data analysis process in section 4.6.

Throughout the early stages of the thesis, there were a lot of uncertainties on the definitions and measurements for performance and affects. As a consequence, several approaches and methods were tested that, unfortunately, did not result in the final solution. Hence, to maintain clear guidance on the methodology of the thesis, these deviations and trade-offs are presented in the final section, i.e. section 4.7.

## 4.1 Participants

28 employees in four teams from VCRS participated in this study. 24 participants choose to answer the Participant Information (PI) survey, whereas one participant provided information that could not be interpreted, hence N = 23.

As for participant characteristics, 20 of the participants were male, two were female, and one did not want to state the gender. The mean age was 35.04 years old with standard deviation (SD) = 9.580 and an age gap between 22-60 years old. As for years of working experience, the mean was 11.3 years, SD = 9.223 and the experience ranged from 1-36 years. The participants consisted mostly of developers, there were 19 developers, two were designers, and two were managers. The PI survey can be found in appendix B.

To participate in the study two criteria had to be matched. First, a participant was required to be in a software development team, and second, he or she was expected to work with the team's development project actively. Also, at least two teams were required to work on separate products. The participants were asked to submit

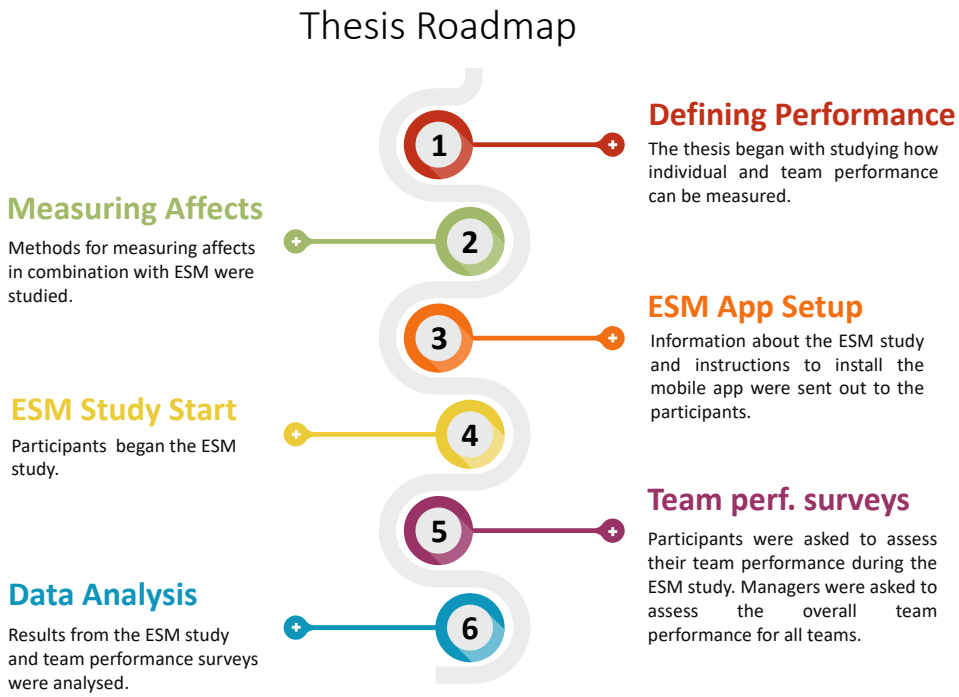


**Figure 4.1:** Overview of the steps taken in the thesis.

answers to the ESM study during working hours, and full anonymity was ensured. The participants were not rewarded.

## 4.2 Performance

A widely used approach for measuring performance are self-assessments. Lenberg and Feldt; Google (2018; 2016) used self-assessments to measure performance in their study. Team members are the ones that best know how their team is performing, and hence, self-assessment was the chosen method to measure performance. As the goal of this thesis is to understand the impact of affects on an individual- and team level, performance was divided into two parts, individual performance and team performance. Sections 4.2.1 and 4.2.2 presents the process of constructing surveys for self-assessing individual performance and team performance respectively.

### 4.2.1 Individual performance

The self-assessment surveys for individual performance were a part of the ESM study and included measurements on efficiency, productivity and quality of work. Individual performance surveys were based on a 5-point ordinal scale ranging from *'strongly disagree'* to *'strongly agree'*. Graziotin et al. (2014) had a similar approach in which they measured individual productivity by including a measurement on productivity in the ESM survey using a 5-point ordinal scale.

Table 4.1 presents the performance indicators used for assessing individual performance. The first two indicators were developed based on studies by Lenberg and Feldt; Henderson and Lee (2018; 1992). The last one was used in the study by Graziotin et al. (2014).

The performance indicators mentioned above were chosen because they were considered to be dynamic, meaning that they could change drastically throughout a day or in different stages of a sprint. Participants assessed these items twice a day and hence, it would have been useless if they were not changing frequently.

**Table 4.1:** Indicators used for assessing individual performance

| |
|---|
| I am working efficiently |
| I am producing quality work |
| I am working productively |

### 4.2.2 Team performance

A problem with self-assessments is the threat to validity due to overestimation bias. That is, developers might overestimate their or the team's actual level of performance. Lenberg and Feldt; Meneghel et al.; Hoegl and Gemuenden (2018; 2016; 2001) addressed this problem and suggested including secondary data sources to triangulate the data. Hence, in this thesis, a total of three surveys were developed for measuring team performance, two for team members and one for managers. The first survey for team members was a self-assessment on the overall team performance (OTP survey), and the second one was a self-assessment on team performance during the ESM study (sprint) period (STP survey). The third survey for managers was an assessment of the teams that were participating in the ESM study (MTP survey).

*Overall team performance survey*
When developing the OTP survey, insight was taken from two studies made by Lenberg and Feldt; Henderson and Lee (2018; 1992). Lenberg and Feldt (2018) was studying impacts of human factors on team performance similar to this thesis, and Henderson and Lee (1992) had done significant work on team performance. Table 4.2 presents the team performance items derived from the two studies, and as shown, many similarities were found.

Further literature reviews were made to confirm the current findings, and to find additional measurements. More specific, discoveries were made in studies by Google; Acuña et al.; Hariharan and Arpasuteerat (2016; 2008; 2017) that helped to complete the list. Table 4.3 presents the final set of performance items, along with related studies that were used.

**Table 4.2:** Team performance items from the studies by Lenberg and Feldt (2018) and Henderson and Lee (1992). Items are shown on the same row if they correspond.

| *Lenberg and Feldt (2018)* | *Henderson and Lee (1992)* |
|---|---|
| This team produce quality work. | The quality of work the team produces. |
| This team delivers according to schedule. | The team's adherence to schedules. |
| My team communicates efficiently with others, e.g. product owner, line manager, and other teams. | Effectiveness of the team's interactions with people outside of the team. |
| The team could become more efficient. | The efficiency of team operations. |
| My team is productive. | - |
| - | The amount of work the team produces. |
| - | The team's adherence to budgets. |
| - | The team's ability to meet the goals of the project. |
| - | The team could have done its work faster with the same level of quality. |
| - | The team met the goals as quickly as possible. |

**Table 4.3:** Presentation of the final set of team performance questionnaire items along with their related studies

| Statement | Related study |
|---|---|
| My team produces quality work | *Lenberg and Feldt; Henderson and Lee (2018; 1992)* |
| My team is able to deliver the expected results | *Henderson and Lee (1992)* |
| My team delivers on time | *Lenberg and Feldt; Henderson and Lee; Google (2018; 1992; 2016)* |
| My team operates efficiently | *Lenberg and Feldt; Henderson and Lee (2018; 1992)* |
| My team feels safe to take risks & to be vulnerable in front of each other | *Google; Acuña, Gómez, and Juristo (2016; 2008)* |
| My team understands what is, and what is not, acceptable member behaviour | *Lenberg and Feldt; Acuña, Gómez, and Juristo (2018; 2008)* |
| The team communication is good within the team | *Hariharan and Arpasuteerat (2017)* |
| My team communicates well with other teams | *Lenberg and Feldt; Hariharan and Arpasuteerat (2018; 2017)* |

The OTP survey was sent out a few days before the ESM study was initiated, and the participants had seven days to submit it through a mobile application used in the ESM study (more on this in section 4.4).

*Sprint team performance survey*
The STP survey and the OTP survey consisted of the same performance items with the exception that the latter survey asked participants to assess their team's performance during the ESM study period. By including both overall and sprint assessments, it was possible to make comparisons and, further, explain potential deviations in the ESM study. The STP survey was sent out at the end of the ESM study using Google Forms.

*Manager team performance survey*
This survey was the same as the OTP survey, except that with this survey, managers with team insight assessed the overall performance of each team. Meneghel et al. (2016) stated that managers could evaluate team performance to better control bias, and Google (2016) made a similar approach where they had managers assess their teams. Hoegl and Gemuenden (2001) emphasises the importance of including multiple viewpoints when evaluating team performance because the success rate of a project depends, in some sense, by the one inspecting. The team performance survey for managers was sent out to four managers one week after the ESM study had started using Google Forms.

## 4.3 Affects

The process for defining affects consisted mostly of reviewing literature. No systematic approach was applied, rather, the authors mainly used Chalmers Online Library to search for relevant research papers and articles. Section 3.2 presents two main approaches for measuring affects; discrete and dimensional.

Hoping to receive confirmation on that the thesis was heading in the right direction, researchers were contacted. Mail conversation was obtained with researcher Daniel Graziotin from the University of Stuttgart, and a meeting was arranged with Per Lenberg, a PhD student from Chalmers University of Technology. These discussion provided feedback on findings made so far, and advice and guidance were given on how to proceed with the thesis were given.

Regarding the two approaches for measuring affects, the dimensional method was applied, with the grounds that, compared to the discrete approach, fewer factors were required for assessing affect. Hence, the dimensional way is most likely the preferable option for tasks that need simplicity and speed, or that are limited in time. Given that affects was only a sub-part of this thesis, other approaches were deemed risky since it could require more work, and hence, possibly divert focus away from the intended purpose of this thesis.

Assessing and measuring affects based on the dimensional approach was done with the self-assessed manikin (SAM). As stated in section 3.2.2, SAM is a pictorial

questionnaire that is composed of 3 sets of 5 figures to assess valence, arousal and dominance. The uniqueness of SAM, compared to other methods and techniques, is that it uses only pictures. As opposed to written text, images are easier to understand, as it eliminates the problems that come with mistranslations, and possible misinterpretation because of cultural differences.

## 4.4 Tool for gathering data

As stated in section 3.3.1, the ESM study was carried out on smartphones, and hence, a mobile application was used, where each participant was asked to use their mobile device. Uncertainties on whether or not to develop this application remained well into the thesis. Finally, a decision was made not to develop the application and instead make use of a third-party solution. Considering the time frame and focus of the thesis, developing a mobile app would be time-consuming and out of scope. When deciding on which application to choose, the following criteria had to be fulfilled:

- iOS and Android availability
- Be able to send notifications
- Schedule repeated surveys

As a result, an application called Expimetrics was chosen. After getting familiarised with the application, the authors realised possibility to perform static, non-repeating surveys. Hence, the app was also used for carrying out additional surveys that were included in the data gathering process. More specifically, one on team performance (see section 4.2) and one on participant information.

Four teams participated in the ESM study, and for each team, a project was created in Expimetrics. Further, each project consisted of the following surveys:

- Overall team performance (See section 4.2.2)
- Participant information (See section 4.1)
- ESM morning report (See section 4.5)
- ESM afternoon report (See section 4.5)

Sprint team performance (STP) survey was initially included in the application but had to be omitted later due to technical issues. Google Forms was used instead.

Further, for each project, a schedule was prepared to determine the availability of the surveys. Each project was associated with a unique access code, which was later distributed to the teams. Finally, the participants were asked to download Expimetrics from their respective application store and enter the provided access code.

## 4.5 ESM study design

The ESM study, consisting of a morning and afternoon survey, was scheduled every workday during a 3-week sprint. Notifications were sent out to the participants at ten o'clock in the morning and two o'clock in the afternoon. A survey was available for one hour, and if a participant had not completed one within the first 20 min, a reminder was sent. Each survey looked the same and consisted of assessing your current feelings with SAM (see section 4.3), and assessing your performance (section 4.1) based on a 5-point ordinal scale ranging from *'strongly disagree'* to *'strongly agree'*. Before the actual ESM study began, it was tested during one week to validate that it worked as expected. When the ESM study was initiated, an email was sent out to the participants containing information about the study, and instructions on how to get started (see appendix A).

Responding to self-reports can, after some time, be very disturbing for the participants as it interrupts their daily workday. The sudden disruption of prompting participants to perform a self-report can ultimately affect the data result negatively. To at least mitigate this problem, the maximum limit on the number of daily self-reports was therefore set to two.

## 4.6 Data analysis

The following section presents the methods and tools that have been used to analyse the collected data from the ESM study and the surveys.

### 4.6.1 Team performance

The internal consistency in the OTP survey was measured by calculating Cronbach's alpha. The alpha for STP and MTP surveys were not calculated. Further discussions about the alpha can be found in section 5.2.

The relations between the affects and team performance were analysed in two ways: through manual interpretation and by using Kendall's tau-b correlation (Kendall's correlation). Manual interpretation was made by first calculating a total mean score of the OTP, STP and MTP surveys separately for each team. Second, a composite team performance score (CTPS) was derived, by equally weighting the three surveys and calculating the mean of them. The total mean score for each affect was derived from the ESM study and calculated by team. The CTPS was then compared with the total mean score for valence, arousal and dominance and plotted in a graph.

Kendall's correlation is a non-parametric measure of association between two variables (Colman, 2008) and is suitable for data sets with variables measured in ordinal scale (e.g. a 5-point scale). Kendall's correlation does not strictly assume monotonic relationship between variables (as X increases in value, Y either keeps increasing or decreasing), although it is desirable and it yields better results (Laerd, 2018). A scatter plot for monotonic analysis can be found in figure 5.6. Kendall's correlation

has shown to be a robust, non-parametric correlation measurement, and is in some cases even more robust than Spearman's rank-order correlation (Croux & Dehon, 2010).

To use Kendall's correlation, the collected data had to be managed differently than as opposed to the manual interpretation. Affects were compared with results from OTP survey only, and participants who did not submit survey answers were discarded. The mean scores for each affect and the CTPS from the OTP surveys were calculated and grouped by participants. Kendall's correlation analysis was used to compare the CTPS with the affect dimensions using SPSS 24. Results are presented in figure 5.12.

### 4.6.2 ESM study

The data analysis procedure of the ESM study was done with the statistical tool R. The authors advice to utilise the RStudio software which is an open-source IDE for R that provides a graphical interface for writing and running your code.

As aforementioned, the data from the ESM study consisted of two parts, affects and individual performance. The ESM study was designed so that for each self-report all six questions, three on affects and three on performance, had to be completed before submitting. As a result, it prevented incomplete data to be captured. Still, as the nature of repeated measurements, there were many missed data points. Some participants only answered the morning or afternoon survey and some did not answer at all. The data was structured as one row per observation, with multiple rows per person. The collected data from Expimetrics had to be modified before being imported into R, and the final structure that was used for the analysis procedure is presented in table 4.4.

**Table 4.4:** Variables, types and range used in the final data set.

| **Variable** | **Type** | **Range** |
|---|---|---|
| Time | num | 1-24 |
| Participants | num | 1-28 |
| Team | num | 1-4 |
| Valence | num | 1-5 |
| Arousal | num | 1-5 |
| Dominance | num | 1-5 |
| Efficiency | num | 1-5 |
| Quality Work | num | 1-5 |
| Productivity | num | 1-5 |

To further clarify table 4.4, *Time* starts from 1 which is the morning observation of the first day of the ESM study, and 2 is the afternoon observation of the first day,

and so forth. The affects and performance indicators are measured on an ordinal 5-point scale, where 1 represents *Strongly disagree* and 5 *Strongly agree*.

As presented in table 4.1, three items were used to measure individual performance. These were designed to measure performance as a single unit and there was no interest in working with them individually. Hence, a composite individual performance score (CIPS) was computed for each data entry by calculating the mean score of the three performance items.

As mentioned in section 3.2, there is no common metric for assessing affects, which further entails that there is no stable approach for comparing affects across persons and that an affect score might be perceived differently between two persons. 'A valence score of one for a person may be equal to a score of three for another person. However, a participant scoring two for valence at time $t$ and five at time $t + x$ unquestionably indicates that the participant's valence is increased.' (Graziotin et al., 2014). As a workaround to this problem, Graziotin et al. (2014) suggests the raw data be standardised, i.e. calculate z-score. With the z-score, the mean score of a variable is zero, and each data entry of the same variable represents how many standard deviations above or below the mean it is. The definition of the model for calculating z-score is presented in equation 4.1,

$$Standardscore, z_i = \frac{x_i - \mu}{\sigma}, \tag{4.1}$$

where $\sigma$ is the standard deviation, $\mu$ is the mean of a variable (valence, arousal, dominance, CIPS), and $x_i$ is the data point to be measured. A function called *scale()* in R was used to standardise the raw data.

After having standardise the raw data for each variable, a visual presentation of the relations between the affects and the CIPS was derived. A function called *ggplot* from the *ggplot2* library provided the necessary functionality to produce the presentation.

Given that the ESM study produced repeated measurements of multiple variables, a linear mixed-effects model (LMM) was implemented. LMMs are statistical models for continuous outcome variables in which the residuals are normally distributed but may not be independent or have constant variance (Welch, Galecki, & West, 2014). Welch et al. (2014) further states that LMMs are appropriate for analysing data sets that include longitudinal or repeated measures, where the subjects are repeatedly measured over time or under different conditions. In comparison to linear models, LMMs take into account the effects of both fixed and random factors. Fixed effects are unknown constant parameters associated with either continuous covariates or the levels of categorical factors, whereas random effects are associated with levels of categorical factors sampled from a sample space, such that each particular level is not of intrinsic interest (Welch et al., 2014). Robinson (1991) formulates the model in equation 4.2.

$$y = X\beta + Zu + e, \tag{4.2}$$

where $y$ and $\beta$ are vectors of $n$ observable random variables and of $p$ unknown parameters with fixed values (fixed effects) respectively. $u$ is a vector of random effects, and $X$ and $Z$ are matrices of fixed and random effects respectively. $e$ is an observation error vector. In R, the LMM was implemented with a function called *lmer* from the *lme4* library, which is presented in equation 4.3.

$$Performance \sim (Valence + Arousal + Dominance) * Time + (1|Participants), \tag{4.3}$$

where *Performance* is the dependent variable, *Valence*, *Arousal*, *Dominance*, and *Time* are the fixed effects, and *(1 | Participants)* is the random effect.

After having implemented the model, the *summary()* function in R was used to summarise the results. Also, by importing a library called *lmerTest*, and calling the built-in *anova()* function, the p-value along with additional descriptive statistics was derived.

## 4.7 Deviations considered

During the thesis, different approaches and methods were tested, and this section presents deviations and trade-offs that have been made.

### 4.7.1 Measuring team performance

The initial step that was taken to understand how performance could be measured involved looking at objective measurements. By reviewing literature, it was evident that using objective measurements for performance is a complex task. A study on development activity measurements conducted by Treude, Figueira Filho, and Kulesza (2015) concluded that many of the developers that participated in the study did not believe in the existence of any measure suitable for measuring development activity.

Individual meetings were arranged with a scrum master, manager, and agile coach from VCRS, as well as with a professor from the University of Chalmers (mentioned in section 4.3). From these meetings, it was further concluded that the use of objective measurements would not be feasible. Despite every development team of VCRS conforming to Scrum, evaluations and estimations were done differently, and hence, it was not possible to compare objective measurements between teams. For instance, a unit of velocity was estimated individually for each team, and defect density was dependant on whether or not teams were working with legacy code (or external code).

Further discussions with the agile coach introduced the possibility to use customer value or satisfaction as a performance indicator. However, not all projects were developed directly to customers. Instead, they were developed to internal 'customers', or to improve the system's performance and stability. Furthermore, teams did not

necessarily push out a new release after each sprint completion. Using customer value or satisfaction as a measurement can be useful if performed in many sprints or product releases where it is possible to compare each.

### 4.7.2 Alternative instruments for measuring affects

As aforementioned, the self-assessment manikin (SAM) was used to measure affects, but other methods such as the international short-form of the Positive and Negative Affect Schedule (I-PANAS-SF) and The Scale of Positive and Negative Experience (SPANE) were heavily considered.

As stated in section 3.2, affects can be categorised into two theories, discrete and dimensional. Both I-PANAS-SF and SPANE are favourite measurement instruments for assessing affects with the discrete approach.

I-PANAS-SF is a 10-item self-report measure of positive and negative affect and offers a reliable measure of affects (Thompson, 2007). A study with I-PANAS-SF has been conducted in combination with ESM by Meimann (2016) in which the author concluded that 'I-PANAS-SF is a reliable and valid scale to measure the construct of positive emotions in daily life by means of the experience sampling method.'. Even though I-PANAS-SF is valid, it was considered to be too disturbing and exhaustive for participants to deal with ten items two times a day, for three weeks. Furthermore, adding individual performance items on top of this would not be feasible.

Several researchers have reported problems when using PANAS, a 20-item measurement of affects. Li, Bai, and Wang; Diener et al. (2013; 2010) stated that PANAS was missing core emotions while including items that were not regarded as emotions. Diener et al. (2010) further claimed that PANAS only captured high-arousal feelings in general. To improve PANAS, Diener et al. (2010) developed SPANE which is based on asking participants to report on six positives and six negative feelings concerning their frequency for the past four weeks. The result produces a positive, negative, and balanced score.

As with I-PANAS-SF, a lot of items are used for assessing affect with SPANE, and unlike ESM, SPANE is not used to measure real-time activities. Hence, neither were used in this thesis.

# 5
# Results

This chapter presents results of the collected data of both the surveys and the ESM study. Section 5.1 presents the individual performance results derived from the ESM study and the results derived from the team performance surveys are presented in section 5.2.

## 5.1 Individual performance

This section presents the results on the individual performance and the ESM study. First, data from the ESM study is presented. This is followed by a manual interpretation of the collected data from ESM, and last, descriptive statistics are presented based on the implementation of a linear mixed-effects model.

### 5.1.1 ESM responses

Table 5.1 presents the response rates of the ESM study. The three-week sprint consisted of a total of 15 weekdays, whereas three days were excluded (holidays, etc.). Hence, the ESM study was available for a total of 12 days. Three teams consisted of 8 members and team B had one member working only two days per week (max possible N = 12 for that member). The fourth team consisted of 4 members. The total response rate was calculated to 59.7% which, according to Berkel et al. (2017) is below average (69.6%) for ESM studies. The average time for completing a single ESM survey was 47 seconds.

**Table 5.1:** Response statistics from the ESM study.

| | Tot. possible resp. | Number of responses | Response rate | Avg. response time (sec) |
|---|---|---|---|---|
| Team A | 96 | 74 | 77.1% | 38s |
| Team B | 180 | 123 | 68.3% | 62s |
| Team C | 192 | 120 | 62.5% | 38s |
| Team D | 192 | 77 | 40.1% | 48s |
| **Total** | **660** | **394** | **59.7%** | **47s** |

### 5.1.2 Manual interpretation

The initial goal when analysing the data from the ESM study was to discover if correlations could be found between individual performance and any of the affects. A visual presentation of the relation between each affect and the composite individual performance score (CIPS) is presented in figures 5.1 - 5.4, with each figure representing one of the four teams that participated.

By manually interpreting the graphs in figures 5.1 - 5.4, a lot of conclusion can be made. However, it is still difficult to distinguish between complete and clear patterns that prove a correlation. Also, depending on how meticulous you are, correlations might be perceived differently. To determine whether a relation is a correlation or not, the authors decided to evaluate the 'level of correlation' by introducing **S**trong, **M**oderate, and **W**eak correlation levels. The findings are presented in table 5.2, where fields with '-' indicates no correlation.

**Table 5.2:** Manual interpretation on the correlation between each affect and CPS from figures 5.1 - 5.3

| **Part.** | **Val.** | **Aro.** | **Dom.** | **Part.** | **Val.** | **Aro.** | **Dom.** |
|---|---|---|---|---|---|---|---|
| 1 | W | W | M | 16 | W | - | - |
| 2 | M | W | W | 17 | W | - | - |
| 3 | - | M | M | 18 | - | W | - |
| 5 | M | M | W | 21 | W | W | W |
| 6 | M | M | - | 22 | W | - | W |
| 7 | W | - | W | 23 | S | - | - |
| 8 | - | - | - | 24 | M | W | - |
| 11 | M | M | - | 31 | M | - | - |
| 12 | W | - | - | 32 | W | - | - |
| 13 | S | M | W | 33 | W | W | W |
| 14 | S | W | - | 34 | S | M | W |
| 15 | - | - | M | 35 | - | - | W |
| 36 | - | - | - | | | | |

### 5.1.3 Linear mixed-effects model

Descriptive statistics of the implemented linear mixed-effect model (LMM) is presented in table 5.3. The results were calculated with a sample size of N = 24 as data from a few participants were omitted and showed error when calculating z-score (their SD = 0). A well-known approach for hypothesis testing is based on what is called significance level, or also p-value. A confidence interval of, e.g. 95% entails that a p-value lower than or equal to 0.05 (1 - 0.95) is significant and hence, the

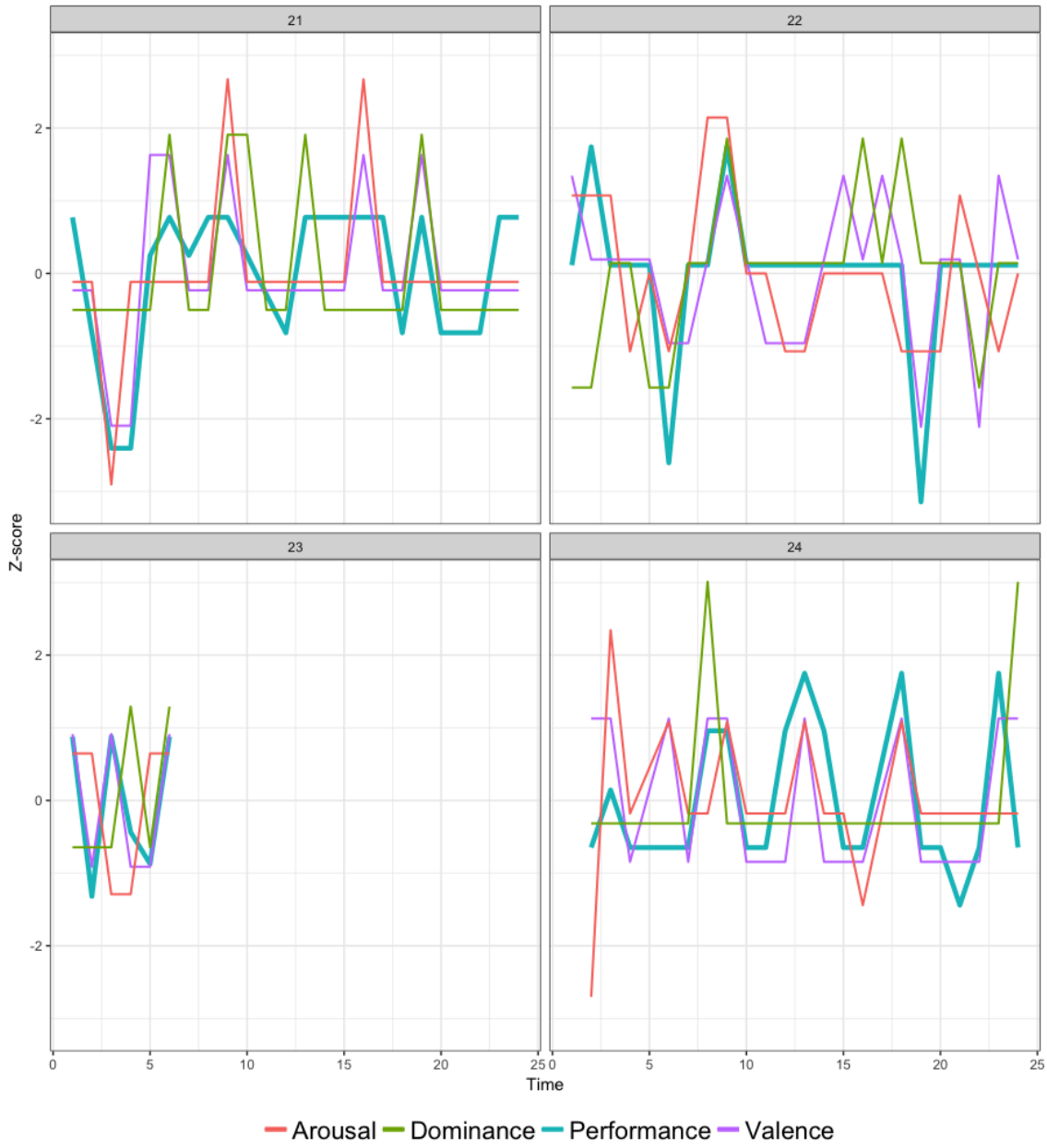


**Figure 5.1:** Relation between each affect and the CPS for team A

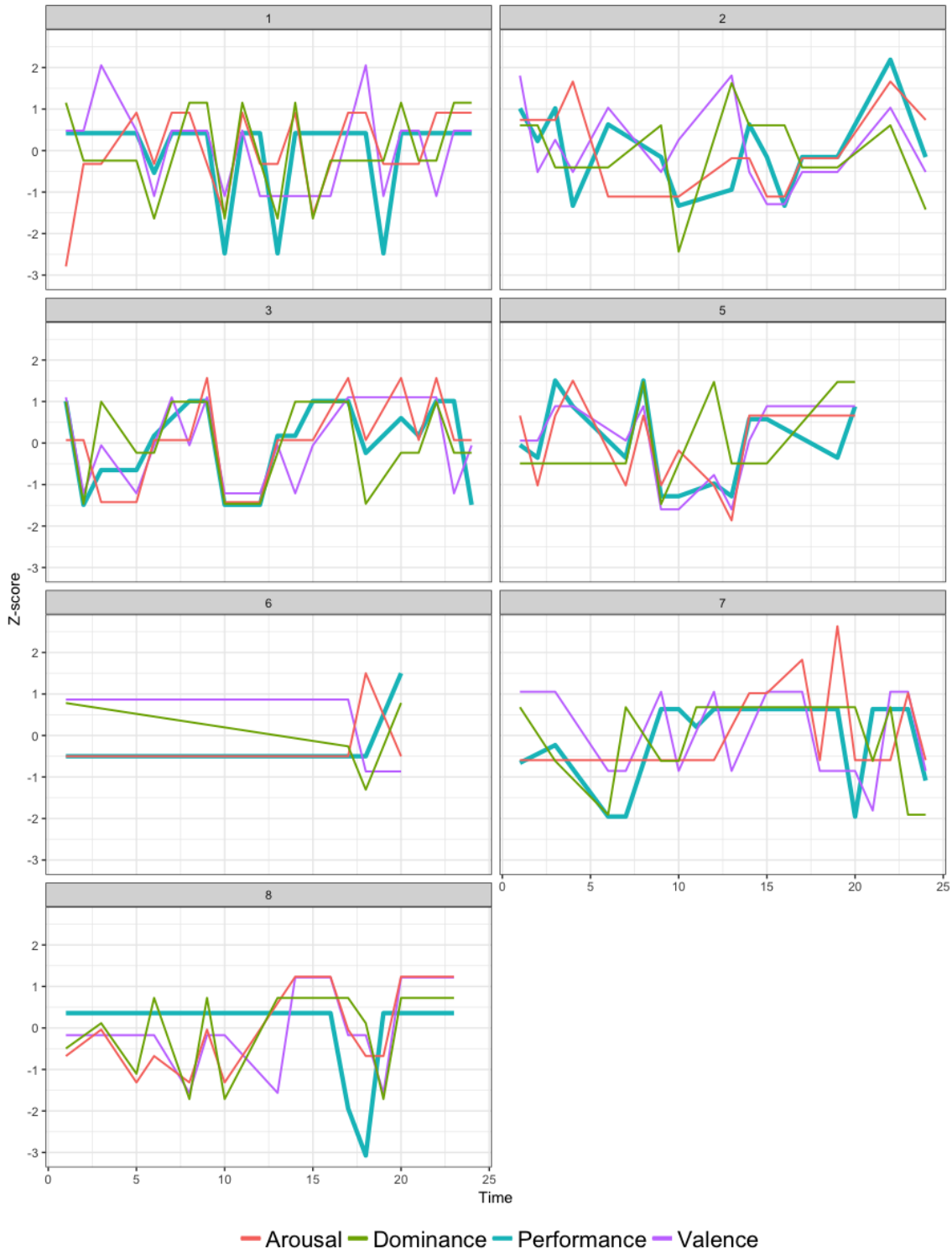


**Figure 5.2:** Relation between each affect and the CPS for team B

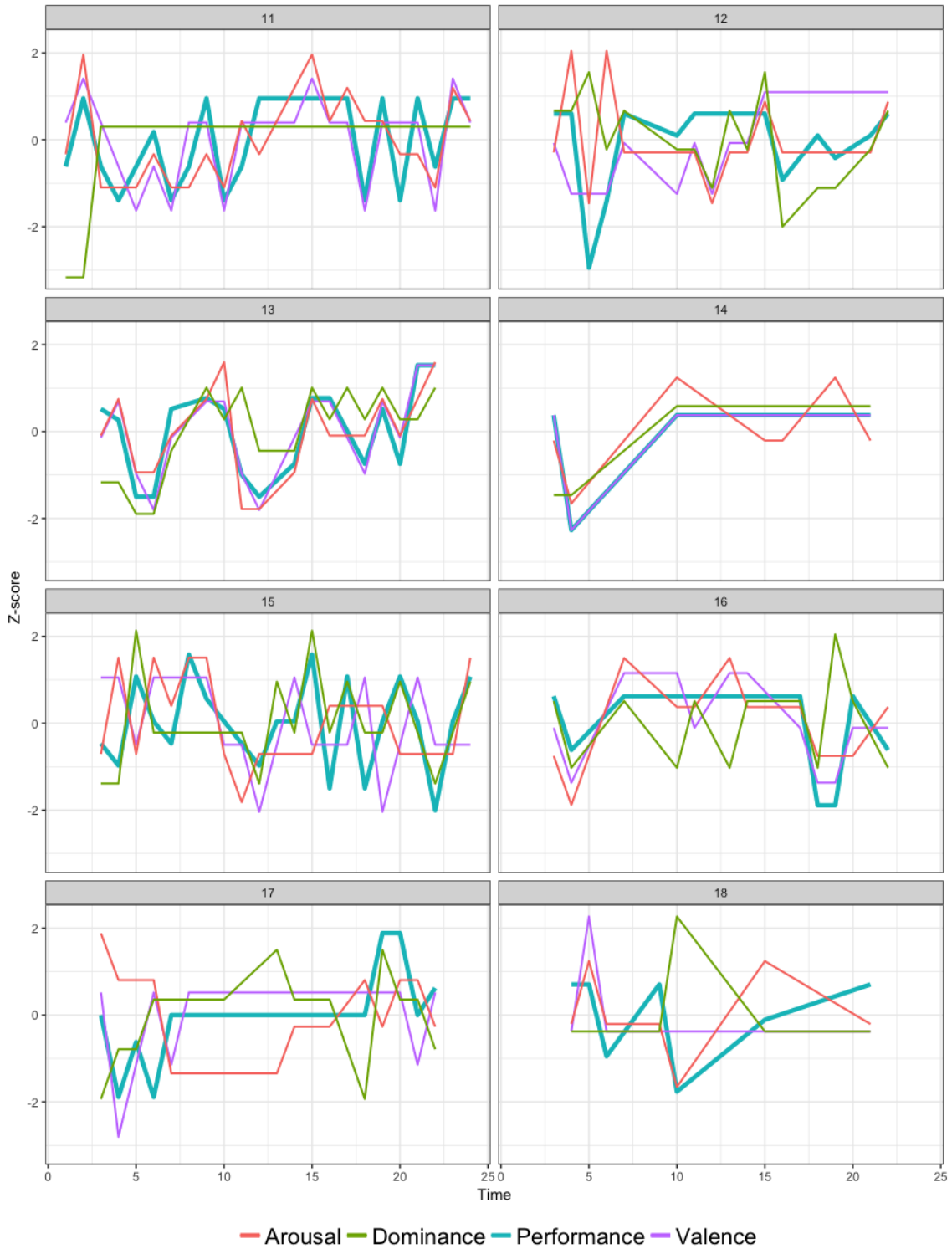


**Figure 5.3:** Relation between each affect and the CPS for team C

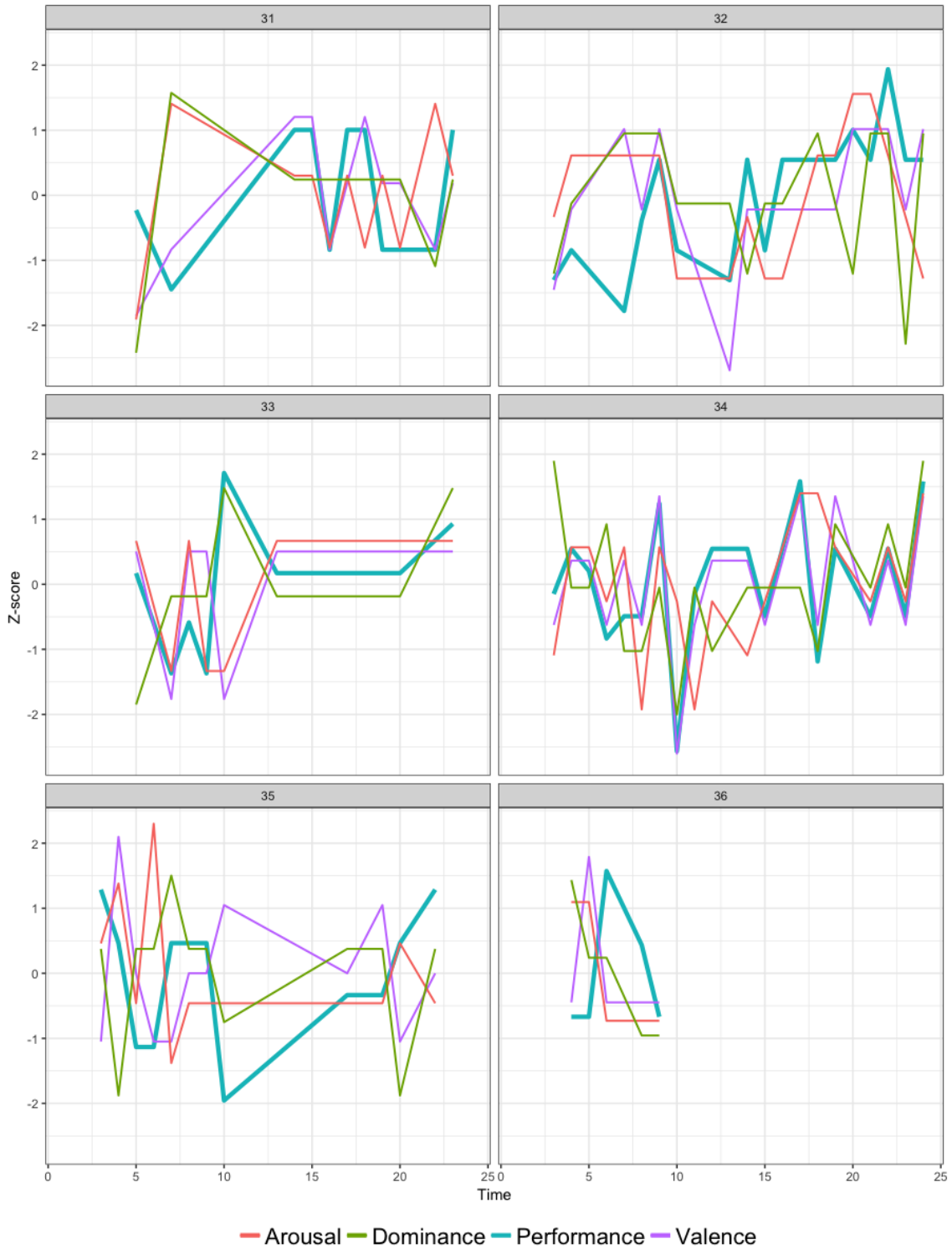


**Figure 5.4:** Relation between each affect and the CPS for team D

null hypothesis can be rejected. In table 5.3, *** and * denotes a 99.9% and a 90% confidence interval respectively.

**Table 5.3:** Fixed effects results from LMM.

| **Fixed effect** | **Sum of square** | **F-value** | **P-value** |
|---|---|---|---|
| Valence | 16.41 | 25.11 | $8.322x10^{-7}$*** |
| Arousal | 0.1905 | 0.29 | 0.5896 |
| Dominance | 1.81 | 2.78 | 0.0966* |
| Time | 2.04 | 3.12 | 0.0780 |
| Valence:Time | 1.07 | 1.64 | 0.2018 |
| Arousal:Time | 1.92 | 2.94 | 0.0875 |
| Dominance:Time | 0.02 | 0.03 | 0.8547 |

The random effects results are presented in table 5.4, which indicate a low variance in the random effect (Participants) with almost all of the variance in the residual group. The table also shows on low standard deviation for both groups.

**Table 5.4:** Random effects results from LMM.

| **Groups** | **Variance** | **Std. Dev.** |
|---|---|---|
| Participants (Intercept) | $7.392x10^{-32}$ | $2.719x10^{-16}$ |
| Residual | 0.6534 | 0.8084 |

## 5.2 Team performance

This section contains an overall presentation of the team performance surveys, followed by results from manual interpretation of the data and Kendall's correlation.

### 5.2.1 Surveys

Cronbach's alpha was used to measure the correlation between the items in the overall team performance (OTP) survey to understand if they measured the same concept (see table 5.5). As the OTP survey is more or less the same as the sprint team performance (STP) and the manager team performance (MTP) survey, it was decided to only calculate the alpha for only one of them. Hence, the OTP survey was selected. The alpha was calculated using data from the OTP survey, containing a total of 8 items, with measurements N = 22 and no excluded values. The alpha was calculated to 0.798 which according to (Tavakol & Dennick, 2011) is considered

acceptable. Table 5.6 shows the alpha value if items were to be removed. As can be seen, eliminating item six and eight would give a slight increase in the alpha value, and removing the other items would decrease.

**Table 5.5:** Cronbach's alpha for the items in the OTP survey.

| | |
|---|---|
| *Cronbach's alpha* | .798 |

**Table 5.6:** Presentation of Overall Team Performance Survey Items and Cronbach's alpha if deleted.

| **Item** | **Cronbach's alpha if deleted** |
|---|---|
| 1. My team produces quality work | .754 |
| 2. My team is able to deliver the expected results | .738 |
| 3. My team delivers on time | .774 |
| 4. My team operates efficiently | .727 |
| 5. My team feels safe to take risks & to be vulnerable in front of each other | .757 |
| 6. My team understands what is, and what is not, acceptable member behaviour | .819 |
| 7. The team communication is good within the team | .776 |
| 8. My team communicates well with other teams | .829 |

Table 5.7 presents an overview of the results from the three surveys, i.e. the team performance indicators. A composite team performance score (CTPS) was derived by calculating the mean score of the eight items.

**Table 5.7:** Mean score of the affects during the ESM study, the team performance surveys (OTP, STP, MTP), and the CTPS.

| | Val | Aro | Dom | OTP | STP | MTP | CTPS |
|---|---|---|---|---|---|---|---|
| Team A | 3.47 | 2.71 | 3.57 | 3.94 | 3.72 | 4.00 | 3.89 |
| Team B | 3.40 | 2.81 | 3.70 | 3.48 | 3.72 | 4.00 | 3.73 |
| Team C | 3.52 | 2.83 | 3.72 | 3.67 | 4.08 | 4.38 | 4.04 |
| Team D | 3.62 | 3.10 | 3.58 | 4.08 | 3.92 | 4.50 | 4.17 |

As aforementioned, four teams participated in the study, and throughout the study, anonymity was promised. Hence, for the remainder of the chapter, each team is

assigned a team name ranging from Team A-D. Out of the 28 participants, 22 submitted the OTP survey (76%).

*Team A*
Team A had four participants in the study (N = 4). All participants submitted complete answers, and with no data excluded. Table 5.8 presents the results from the OTP survey. It is evident that members of team A moderately agree on their performance as the standard deviation (SD) is lower than 1. Table 5.7 shows a total mean score of 3.94 for the OTP survey, and the team members rated their performance in the STP survey slightly lower (mean = 3.72). The MTP survey shows a mean of 4.00. In conclusion, the self-assessed OTP mean score of 3.94 can be perceived as a slightly high when compared to the STP and MTP surveys.

**Table 5.8:** Team A's results from the OTP survey.

| **Item** | **Mean** | **Standard Deviation** |
|---|---|---|
| My team produces quality work | 4.00 | .816 |
| My team is able to deliver the expected results | 3.75 | .500 |
| My team delivers on time | 3.50 | .577 |
| My team operates efficiently | 3.75 | .500 |
| My team feels safe to take risks & to be vulnerable in front of each other | 4.00 | .816 |
| My team understands what is, and what is not, acceptable member behaviour | 4.25 | .500 |
| The team communication is good within the team | 4.25 | .500 |
| My team communicates well with other teams | 4.00 | .816 |

*Team B*
In Team B, eight people participated in the study. However, two participants did not provide answers, hence N = 6. No submitted data were excluded. Table 5.9 presents the results from the OTP survey. SD shows that the team members have a rather diverse opinion about their team performance. Table 5.7 shows a mean score of 3.48 for the OTP survey, whereas the results from STP and MTP surveys suggest the performance to be better, at least during the period of the ESM study.

*Team C*
A total of eight people participated from Team C, whereas two participants did not submit any answers, and hence N = 6. No submitted data were excluded. Table 5.10 presents the results of the OTP survey, and it is evident that the SD vary between the items. The team members fully agree on that they deliver on time (SD = 0). However, they have different opinions on the understanding of acceptable behaviour (SD = 1.033). Table 5.7 presents a mean score of 3.67 for the OTP survey. The

**Table 5.9:** Team B's results from the OTP survey.

| Item | Mean | Standard Deviation |
|---|---|---|
| My team produces quality work | 3.67 | 1.033 |
| My team is able to deliver the expected results | 3.50 | 1.049 |
| My team delivers on time | 2.67 | .816 |
| My team operates efficiently | 3.33 | 1.211 |
| My team feels safe to take risks & to be vulnerable in front of each other | 3.17 | 1.169 |
| My team understands what is, and what is not, acceptable member behaviour | 4.17 | .753 |
| The team communication is good within the team | 3.83 | .753 |
| My team communicates well with other teams | 3.33 | .516 |

STP and MTP surveys suggest team performance be much better, at least during the ESM study period.

**Table 5.10:** Team C's results from the OTP survey.

| Item | Mean | Standard Deviation |
|---|---|---|
| My team produces quality work | 4.50 | .548 |
| My team is able to deliver the expected results | 4.83 | .408 |
| My team delivers on time | 5.00 | .000 |
| My team operates efficiently | 4.33 | .816 |
| My team feels safe to take risks & to be vulnerable in front of each other | 4.17 | .983 |
| My team understands what is, and what is not, acceptable member behaviour | 3.67 | 1.033 |
| The team communication is good within the team | 4.17 | .408 |
| My team communicates well with other teams | 3.67 | .516 |

*Team D*

Team D had nine participants, but only six submitted answers to the OTP survey (N = 6). No submitted data were excluded. Table 5.11 presents the results of the OTP survey, and the SD indicates that the team members mostly agree on their team performance. They fully agree on quality work (SD = 0), but differ in their opinion about understanding acceptable member behaviour (SD = 0.816). Results from table 5.7 shows a mean score of 3.58 for the OTP survey, but STP and MTP surveys suggest their performance for the ESM study period to be much better.

**Table 5.11:** Team D's results from the OTP survey.

| Item | Mean | Standard Deviation |
|---|---|---|
| My team produces quality work | 4.00 | .000 |
| My team is able to deliver the expected results | 4.17 | .408 |
| My team delivers on time | 3.83 | .408 |
| My team operates efficiently | 3.83 | .408 |
| My team feels safe to take risks & to be vulnerable in front of each other | 4.83 | .408 |
| My team understands what is, and what is not, acceptable member behaviour | 4.33 | .816 |
| The team communication is good within the team | 4.33 | .516 |
| My team communicates well with other teams | 3.33 | .516 |

### 5.2.2 Manual interpretation

Figure 5.5 presents the relation between the total mean score of each affect and the CTPS from the team performance surveys (OTP, STP, MTP), grouped by teams. The data has been standardised. Further discussions can be found in section 6.2.

### 5.2.3 Kendall's correlation

Kendall's correlation does not assume monotonic relationships between variables, but such relationships yield better results as Kendall's correlation measures association between two variables. Figure 5.6 shows a scatter plot used to analyse monotonic relationships between variables. A monotonic relationship is found where X increases and Y continues to either increase or decrease. A non-monotonic relationship occurs when X increases and Y first increases and then decreases, or vice versa. The scatter plot showed both monotonic and non-monotonic relationships. For instance, it can be distinguished that Arousal/Performance and Arousal/Valence have non-monotonic relationships. Monotonic relationships can be identified in Performance/Valence and Valence/Dominance. Performance/Dominance can be argued as a monotonic relationship.

Table 5.12 presents the output from SPSS using Kendall's correlation. The sample size for all relationships is N = 23, which was calculated at an alpha level of 0.05. Correlation coefficients range between -1 and 1. A positive correlation means that both values are increasing and a negative correlation shows that one value is increasing as the other one is decreasing. The closer to -1 or 1, the stronger relationship. Results and hypothesis testing are discussed in section 6.1.

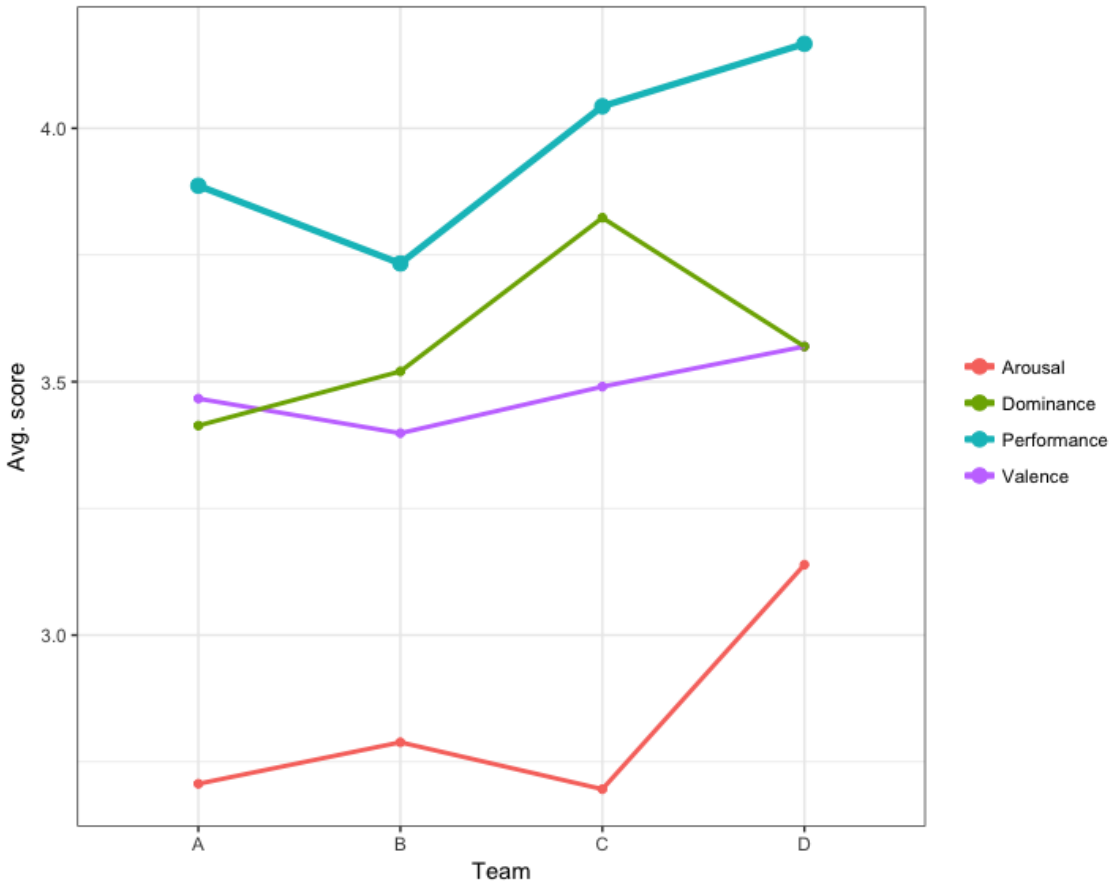


**Figure 5.5:** Relation between each affect and the CTPS for the teams

**Table 5.12:** Results from Kendall's correlation analysis from SPSS 24.

| | | Perf. | Val. | Aro. | Dom. |
|---|---|---|---|---|---|
| **Performance** | Corr. Coeff. | | 0.203 | -0.168 | 0.233 |
| | Sig. (2-tailed) | | 0.184 | 0.275 | 0.130 |
| **Valence** | Corr. Coeff. | 0.203 | | 0.211 | 0.310 |
| | Sig. (2-tailed) | 0.184 | | 0.161 | 0.039 |
| **Arousal** | Corr. Coeff. | -0.168 | 0.211 | | -0.144 |
| | Sig. (2-tailed) | 0.275 | 0.161 | | 0.340 |
| **Dominance** | Corr. Coeff. | 0.233 | 0.310 | -0.144 | |
| | Sig. (2-tailed) | 0.130 | 0.039 | 0.340 | |

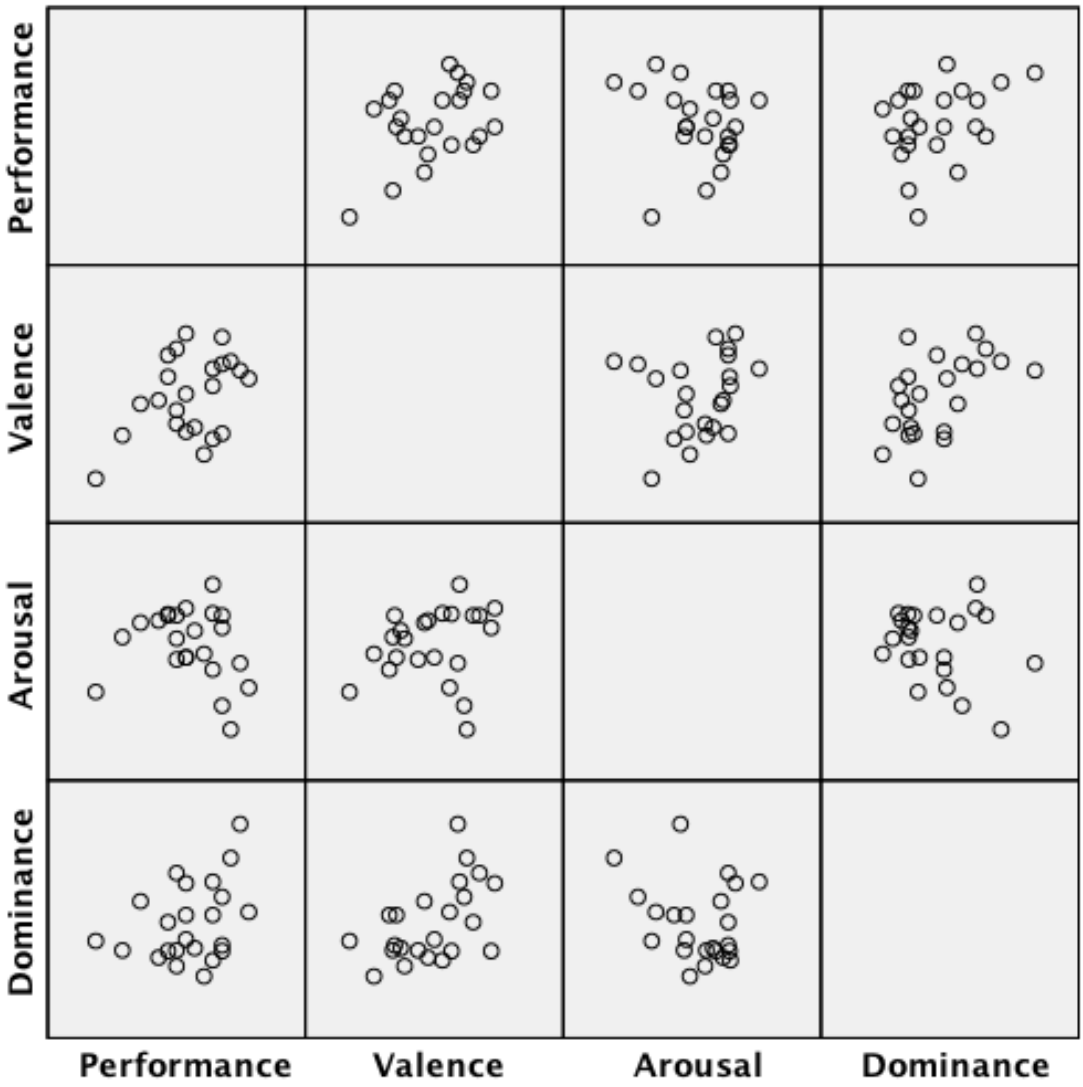


**Figure 5.6:** Scatter plot to analyse monotonic relationships.

# 6
# Discussion

The following sections reflects on and discusses the achieved results. Hypothesis testing is presented in section 6.1 and evaluation of the results is presented in section 6.2. Section 6.3 reflects on the experiences with working with ESM on mobile devices, and the chapter is concluded with a discussion on the threats to validity in section 6.4.

## 6.1 Hypothesis testing

As mentioned in section 5.1.3, the level of significance is a well-known approach for hypothesis testing. Typical significance levels are 0.001, 0.01, 0.05, and 0.10, with an informal interpretation of *very strong evidence*, *strong evidence*, *evidence*, and *weak evidence* respectively (Samuels & Gilchrist, 2014). For this thesis, we decided on the 0.05 level of significance with the motivation that, we believe that proving strong or very strong evidence was not necessarily the main goal. The mail goal was to understand the impact of affects on team performance, and thus, whether we were able to reject the null hypotheses or not was not as important. Needless to say, finding a strong or very strong correlation between an affect and performance would definitely be interesting. With that being said, section 6.1.1 and 6.1.2 performs hypothesis testing with a 0.05 level of significance.

### 6.1.1 Individual performance

Looking at table 5.3, some conclusions can be drawn with regards to the developed hypotheses. We can say that, at the 0.05 significance level, the null hypothesis $H_{02}$ is rejected in favour of the alternative hypothesis $H_{12}$:*The valence affective state of a software developer impacts self-assessed individual performance.* However, no significance could be seen for neither arousal nor dominance, and hence, we fail to reject the null hypotheses for both $H_{03}$ and $H_{05}$. In other words, valence is likely to have a (very) significant impact on the composite individual performance score (CIPS), whereas arousal and dominance did not.

Analysing the results from the random effects (table 5.4), both the residual and the random effect (Participants) indicated a low variance score. Hart (2012) explains that the residual variance tells how much variability there is within the fixed effect(s), and the variance for the random effect of participants tells how much of that within fixed effect(s) variance is explained by participant differences. With that being said, in this thesis we can say that the random effect of *Participants* is at such a low

score that it does not explain much of the variability between *Participants* across the level(s) of CIPS.

### 6.1.2 Team performance

Kendall's correlation was used to test the hypothesis regarding team performance, i.e. $H_{X1}$, $H_{X3}$, $H_{X5}$. Table 5.12 presents the results of Kendall's correlation and figure 5.6 shows a scatter plot of monotonic (and non-monotonic) relationships between variables.

Between the variables performance and valence, we can see a monotonic relationship in the scatter plot. A weak and positive correlation is identified, with a coefficient score of 0.203, but it is not significant. Hence, we fail to reject the null hypothesis $H_{01}$: *The valence affective state of a software developer has no impact on self-assessed team performance.*

The table shows that performance and arousal have a weak and negative correlation, but this interpretation can be questioned as the relationship is non-monotonic. The correlation between performance and arousal is not significant, and we fail to reject the null hypothesis $H_{03}$: *The arousal affective state of a software developer has no impact on self-assessed team performance.*

The relationship between performance and dominance could be argued as monotonic. The correlation is weak and positive, with a coefficient of 0.233, and it is not significant. We fail to reject the null hypothesis $H_{05}$: *The dominance affective state of a software developer has no impact on self-assessed team performance.*

## 6.2 Evaluation of results

In this section, we evaluate and discuss the results of this thesis, which is divided into two parts; individual and team performance.

*Individual performance*
The results from the ESM study proved valence to have an impact on performance. The p-value for dominance was 0.0966, meaning that at a 0.10 significance level it does have an impact on individual performance. However, as aforementioned, the 0.05 level of significance was chosen for this thesis. Also, had a lower level of significance been chosen, such as 0.10, the validity of the findings could be questioned due to the increased risk of making a type I error. Further, looking at the manual interpretation results in table 5.2, it shows zero strong correlations, only two moderate correlations, and thirteen non-correlations (sample size = 25). It should be kept in mind that these results are from manual interpretation, and hence, the outcome will most likely differ from one evaluator to another.

Based on the p-value approach, it is self-explanatory that with a p-value of 0.5896, arousal did not have a significant impact on performance. Further validations can be made by looking at the manual interpretation in table 5.2. Out of twenty-five

participants, five moderate, and seven weak correlations were found, and for twelve participants no correlations at all were found.

As for valence, the p-value of $8.322x10^7$ indicates a strong correlation to performance. It further proves also to be a significance in the 99.9% confidence interval, showing a very strong correlation. The manual interpretation in table 5.2 further validates this with four strong, seven moderate, and nine weak correlations for valence. For seven participants, no correlations were found.

It is interesting to compare the individual performance result from this study to the one made by Graziotin et al. (2014). In that study, dominance proved to have the strongest correlation, with valence showing on a weaker correlation. As with this thesis, no significance was found for arousal. The need for further and future work on the topic is evident. Compared to a study length of 90 minutes in the study by Graziotin et al. (2014), the duration of this study lasted longer (fifteen days), which introduces a new perspective for how to design similar studies.

*Team performance*
Manual interpretation of graphs was not used for testing the hypotheses. However, such analysis was included to get a holistic view of the data regarding team performance. Hypothesis testing for affects and team performance was done using Kendall's correlation (section 6.1.2).

Table 5.7 indicates a close relationship between valence and performance (composite team performance score). For instance, team D has the highest valence score and also the highest CPS, and team B has the lowest valence and CPS. This is further justified in figure 5.5 which indicates that there exists a correlation between valence and performance. It is also evident that the teams have experienced low arousal scores compared to the other affects. Dominance was the affect that all of the teams felt were the strongest during the ESM study. However, the shape of dominance shows a different pattern of performance which may indicate a weak correlation between the variables.

Kendall's correlation (table 5.12) shows that there were no significant correlations between performance and the affect dimensions. It is believed that the study design has a significant impact on the results. The method used to transform data to use Kendall's correlation for analysis between affects and team performance can be a threat to construct validity. It was done by calculating the mean affects for each participant and comparing it with their self-assessed team performance. In other words, the comparison is between the participants average feelings with their perception of their team's performance, e.g. 'how I felt compared to my perception of my team's performance'. Maybe the participants average feelings should be compared to another team performance score that includes different viewpoints. One can question whether this is a valid measurement or not.

It is interesting to see that valence has a significant impact on dominance (at the 0.05 level) as they have a weak and positive relationship (correlation coefficient = 0.310). As valence increases, dominance increases and vice versa. This finding raises

questions about the validity of using self-assessment manikin (SAM) when measuring affects in software engineering. If valence and dominance were strongly correlated, it would be unnecessary to include both dimensions as they may be measuring the same set of human factors. Part of conducting a successful ESM study is to keep the study short and not very time consuming, hence, by omitting one survey item can do much.

## 6.3 Assessment of ESM

The ESM study yielded 394 responses, with a response rate of 59.7% (see table 5.1). According to Berkel et al. (2017), the average response rate is 69.6%, and comparing that to what was achieved in this ESM study, we deem it as acceptable. Thanks to a well-planned and structured survey, the time and effort required for participants were minimised. This is important as it keeps participants motivated and mitigates the feeling of being interrupted. Adams (1963) elaborates on this by stating that participants are more willing to submit responses if their costs concerning time, energy and resources are low.

The use of mobile devices with ESM resulted in a successful data gathering process. A third party mobile application was used which saved a lot of time and effort with regards to having to develop a new ESM application. However, it also had its disadvantages in not having any control over the app, but only access to the data. For instance, some participants had problems with accessing surveys, and there were also problems with controlling the survey notifications which resulted in undesirable mobile notifications. When conducting an ESM study on mobile devices, scheduling surveys and sending notifications is critical and it has to work correctly. Considering that problems occurred in these areas introduced some challenges, however, the overall experience was still good. Data availability, notifications, and scheduling of surveys were the differentiating factors when compared to traditional approaches (e.g. pen & paper).

Participants had no problems with downloading the ESM application on their mobile devices, and there were no problems with setting up accounts and so forth. Having participants to use their device was vital as it does not take away from their natural setting. This can, in turn, mitigate the feeling and thought of actually being in a study, and hence, increase the reliability of the data.

Another benefit of ESM on mobile devices was noticed during the data gathering process in that attendance of the researchers were not as vital as it might have been with traditional approaches. As long as participants were provided with a clear introduction and instructions on the study, and they knew how and who to contact if problems would occur, researchers focus could be aimed at analysing the incoming data. This would not have been possible with traditional approaches as they would demand a lot more effort regarding handing out and collecting paper surveys, and manually importing collected data to statistical tools.

A common problem when carrying out studies like ESM is the threat to validity when gathering data. When participants are asked to provide a self-report, they might alter their answers to one that they believe is right or most suitable, instead of giving an honest answer. The consequences of such behaviour will lead to inaccurate and false data to be gathered and analysed. Unfortunately, this is difficult to prevent, and the reason for such an act can be because of lack of motivation (Scollon et al., 2003). However, a way to mitigate this problem is to apprise the participants by emphasising the importance of the thesis. Anonymously gathering the data is another well-emphasised approach. This way, the participants do not need to worry about being identified or being associated with the collected data.

## 6.4 Threats to validity

This section discusses threats to validity, and is divided into four sections; conclusion validity, internal validity, construct validity and external validity with the definition provided by Wohlin (2012).

*Conclusion validity* is concerned with issues of the relationship between the treatment and the outcome of the study, to be able to draw the correct conclusion. Regarding the team performance data, the small sample size can be a threat to conclusion validity. For instance, in overall team performance (OTP) survey, the sample size was four for one team. However, the threat to validity was mitigated by triangulating the data with two other team performance surveys. Also to be noted is that the STP and MTP surveys had lower sample sizes than the OTP survey. Reliability of measures regarding performance measurements can be questioned as the surveys were never tested on a set of participants before the study began.

*Internal validity* is concerned with the impact of variables that have not been taken into account or are uncontrollable but impacts the results, for instance, paydays, holidays etc. For this thesis, different events such as paydays, holidays or personal events were not be taken into account even if it may have impacted the results.

*Construct validity* is concerned with if a study measures what it intends to. Measuring individual and team performance is a complicated task which has been done in different ways. Researchers have used both objective measurements and self-assessments. In this thesis, individual and team performance is measured with self-assessments which were also done in studies conducted by Graziotin et al.; Lenberg and Feldt (2014; 2018). The self-assessment survey in this thesis is not similar to the other two studies, and there is an uncertainty about whether it is valid performance measurement.

The method that was used to fit the data for Kendall's correlation can be a threat to construct validity because participants average feelings were compared with their perception of their team's performance, and not with a validated performance value (e.g. using different viewpoints to assess performance).

Evaluation apprehension can be a threat to validity. In the ESM study, affects and individual performance is measured, and as there are no right or wrong answers when

answering questions about affects, participants might overestimate their ability to yield better results.

*External validity* is concerned with the ability to generalise the results, e.g. if the results apply outside the study. This thesis was conducted with four teams from VCRS. Hence, the results will apply to teams within VCRS, but it may not apply to teams in other companies. This threat to validity was mitigated by having participating teams to work on different projects. Another threat is the participant gender distribution, only two of twenty-three were female (8.7%). According to StockholmsHandelskammare (2017), 21% of all developers in Stockholm are female. Because of such a low value for female, the results of this study may not be valid for other organisations as it does not reflect reality.

# 7
# Conclusion

Many researchers have raised the need for conducting studies on human factors in software engineering (Graziotin et al., 2014; Lenberg et al., 2014; Khan et al., 2011). This thesis answers the call by contributing to these previous studies and further studying the impacts of affects on performance in an industrial setting, on both individual and team level. Several methods for measuring affects have been discussed, and conclusively, SAM was chosen as it had shown good results in a similar study with ESM. The most suitable approach for measuring performance was with self-assessments, which was used to assess individual performance in the ESM study and team performance in the surveys.

For assessing individual performance, an ESM study was conducted, where data was gathered two times daily for a three week period. The data analysis consisted of manual interpretation and implementation of a linear mixed-effect model (LMM). Three self-assessment surveys were used to gather data on team performance, and together with the affect scores from the ESM study, manual interpretation and Kendall's correlation were used to analyse the data.

The ESM study was carried out on a third-party application, and the overall experience was good with some minor problems. Analysing the results on individual performance was done in R, where the LMM model also was implemented. Kendall's correlation was used in SPSS.

Manual interpretation of the relation between performance and each affect was made on both individual and team level to validate the results from the LMM and Kendall's correlation.

The final results proved a correlation between valence and performance on an individual level. For team performance, no correlations were found for any of the affects. However, the results did present an unexpected finding in which valence showed to have a positive and weak impact on dominance. This raises a question for future studies as to if valence and dominance are measuring the same set of human factors.

# 8

# Future work

Many companies highly depend on the success of their teams, and it is essential for future researchers to study the impact of affects on team performance further. Regarding the process of gathering team performance data, we suggest more frequent measurement points (e.g. once or twice per week) to be included. Also, considering that team performance was assessed through surveys, and affects through ESM (repeated measurements), it was difficult to analyse the correlation using statistical models. We failed to find a solution to this problem since a majority of the performance indicators that we found were static and changed little over time. For instance, factors such as group maturity, norms, and communication usually stay the same over a two or three week period. Hence, to get more valuable results, a study like ours should be conducted over the course of months, or include dynamic performance indicators when measuring performance. The data analysis process for individual performance, i.e. LMM, provided excellent results and should be considered in future studies if the study design consists of repeated measurements.

A study like ours could be conducted across multiple companies to generalise the results. Moreover, carrying out a longitudinal study in which one would apply and analyse the effects of treatments on the participants or teams could provide great findings. For instance, by directly induce various affect dimensions and analyse its impact on performance. We also suggest future studies to try alternative tools, such as I-PANAS-SF or SPANE, for assessing affects.

# References


Acuña, S. T., Gómez, M., & Juristo, N. (2008). Towards understanding the relationship between team climate and software quality—a quasi-experimental study. *Empirical Software Engineering*, *13*, 401–434. doi:10.1007/s10664-008-9074-8

Adams, J. S. (1963). Toward an understanding of inequity. *Journal of abnormal psychology*, *67*, 422.

Agile Manifesto. (2001). http://agilemanifesto.org. Accessed: 2018-04-04.

Altshuller, M. (1923). Byudzhet Vremeni [Time-Budget] perm.

Amrit, C., Daneva, M., & Damian, D. (2014). Human factors in software development: On its underlying theories and the value of learning from related disciplines. a guest editorial introduction to the special issue. *Information and Software Technology*, *56*, 1537–1542. doi:https://doi.org/10.1016/j.infsof.2014.07.006

Baas, M., De Dreu, C. K. W., & Nijstad, B. A. (2008). A meta-analysis of 25 years of mood-creativity research: Hedonic tone, activation, or regulatory focus? *Psychological Bulletin*, *134*, 779–806. doi:10.1037/a0012815

Backs, R. W., da Silva, S. P., & Han, K. (2007). A comparison of younger and older adults' self-assessment manikin ratings of affective pictures. *Experimental Aging Research*, *31*(4), 421–440.

Barrett, L. F. & Barrett, D. J. (2001). An Introduction to Computerized Experience Sampling in Psychology. *Social Science Computer Review*, *19*, 182–183. doi:10.1177/089443930101900204

Berkel, N. V., Ferreira, D., & Kostakos, V. (2017). The Experience Sampling Method on Mobile Devices. *ACM Computing Surveys*, *50*, 1–40. doi:10.1145/3123988

Betella, A. & Verschure, P. F. M. J. (2016). The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLOS ONE*, *11*(2), 1–11. Retrieved from https://doi.org/10.1371/journal.pone.0148037

Bevans, G. E. (1913). *How workingmen spend their time* (Doctoral dissertation). doi:10.1002/job.1803

Bolger, N. & Laurenceau, J.-P. (2013). *Intensive Longitudinal Methods*.

Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering autobiographical questions: The impact of memory and inference on surveys. *Science*, *236*, 157–161.

Bradley, M. M. & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*, 49–59. doi:https://doi.org/10.1016/0005-7916(94)90063-9

Colman, A. M. (2008). Kendall's tau. Retrieved from http://www.oxfordreference.com/view/10.1093/acref/9780199534067.001.0001/acref-9780199534067-e-4437

Croux, C. & Dehon, C. (2010). Influence functions of the spearman and kendall correlation measures. *Statistical Methods & Applications*, *19*, 497–515.

Csikszentmihalyi, M. (1997). *Finding Flow*. Basic Books.

DeMarco, T. & Lister, T. (2013). *Peopleware: Productive projects and teams (3rd edition)*. Addison Wesley.

DeMarco, T. & Lister, T. R. (1987). *Peopleware: Productive projects and teams*. Dorset House Pub. Co.

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D.-w., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*, 143–156.

Downey, S. & Sutherland, J. (2013). Scrum metrics for hyperproductive teams: How they fly like fighter aircraft. In *2013 46th hawaii international conference on system sciences* (pp. 4870–4878). doi:10.1109/HICSS.2013.471

Dutra, A. C. S., Prikladnicki, R., & França, C. (2015). What do we know about high performance teams in software engineering? results from a systematic literature review. (pp. 183–190). doi:10.1109/SEAA.2015.24

Eerola, T. & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*(1), 18–49.

Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. *19*, 207–283.

Ekman, P. (2003). *Emotions Revealed*. St. Martin's Griffin.

Feldt, R., Torkar, R., Angelis, L., & Samuelsson, M. (2008). Towards individualized software engineering: Empirical studies should collect psychometrics. In *Proceedings of the 2008 international workshop on cooperative and human aspects of software engineering* (pp. 49–52). ACM. doi:10.1145/1370114.1370127

Fisher, C. D. & To, M. L. (2012). Using experience sampling methodology in organizational behavior. *Journal of Organizational Behavior*, *33*, 865–877. doi:10.1002/job.1803

Frijda, N. H. (1993). *Moods, emotion episodes, and emotions. In M.Lewis, & J. Haviland (Eds.), Handbook of emotions*. New York: Guilford Press.

Goodman, P., Ravlin, E., & Schminke, M. (1987). Understanding groups in organizations. *Research in Organizational Behavior*, *9*, 127–173.

Google. (2016). Understand team effectiveness. Accessed: 2018-03-27. Retrieved from https://rework.withgoogle.com/guides/understanding-team-effectiveness/steps/introduction

Graziotin, D., Wang, X., & Abrahamsson, P. (2013). Are happy developers more productive? the correlation of affective states of software developers and their self-assessed productivity. *Product-Focused Software Process Improvement*. doi:10.1007/978-3-642-39259-7_7

Graziotin, D., Wang, X., & Abrahamsson, P. (2014). Do feelings matter? on the correlation of affects and the self-assessed productivity in software engineering. *Journal of Software: Evolution and Process*, *27*(7), 467–487. doi:10.1002/smr.1673

Graziotin, D., Wang, X., & Abrahamsson, P. (2015a). Happy software developers solve problems better: Psychological measurements in empirical software engineering. *CoRR*, *abs/1505.00922*. doi:10.7717/peerj.289

Graziotin, D., Wang, X., & Abrahamsson, P. (2015b). Understanding the affect of developers: Theoretical background and guidelines for psychoempirical software engineering. *Proceedings of the 7th International Workshop on Social Software Engineering - SSE 2015*, 25–32. doi:10.1145/2804381.2804386

Gren, L., Torkar, R., & Feldt, R. (2017). Group development and group maturity when building agile teams: A qualitative and quantitative investigation at eight large companies. *Journal of Systems and Software*, *124*, 104–119.

Hackman, J. R. (1987). *The design of work teams*. Prentice-Hall.

Hall, T., Sharp, H., Beecham, S., Baddoo, N., & Robinson, H. (2008). What do we know about developer motivation? *IEEE Software*, *25*, 92–94. doi:10.1109/MS.2008.105

Hariharan, B. & Arpasuteerat, P. (2017). *Combining hard and soft aspects in project performance measurement - a qualitative research undertaken in an agile software development project scenario* (Master's thesis).

Hart, T. (2012, June 2). Making sense of random effects. Retrieved June 2, 2018, from https://www.r-bloggers.com/making-sense-of-random-effects/

Hazzan, O. & Hadar, I. (2008). Why and how can human-related measures support software development processes? *Journal of Systems and Software*, *81*, 1248–1252. doi:https://doi.org/10.1016/j.jss.2008.01.037

Henderson, J. C. & Lee, S. (1992). Managing i/s design teams: A control theories perspective. *Management Science*, *38*, 757–777.

Introduction to Theoretical and Methodological Issues. (1990). In J.-M. Hoc, T. Green, R. Samurçay, & D. Gilmore (Eds.), *Psychology of programming* (pp. 1–7). London: Academic Press. doi:https://doi.org/10.1016/B978-0-12-350772-3.50005-7

Hoegl, M. & Gemuenden, H. G. (2001). Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science*, *12*, 435–449. doi:10.1287/orsc.12.4.435.10635

Huang, M.-H. (2001). The theory of emotions in marketing. *Journal of Business and Psychology*, *16*(2), 239–247.

Imbir, K. K. (2016). Affective norms for 718 polish short texts (anpst): Dataset with affective ratings for valence, arousal, dominance, origin, subjective significance and source dimensions. *Frontiers in psychology*, *7*, 1030.

Izard, C. E. (1977). *Human emotions*. Boston, MA: Springer US.

John, M., Maurer, F., & Tessem, B. (2005). Human and social factors of software engineering: Workshop summary. *SIGSOFT Softw. Eng. Notes*, *30*, 1–6. doi:10.1145/1082983.1083000

Khan, I. A., Brinkman, W.-P., & Hierons, R. M. (2011). Do moods affect programmers' debug performance? *Cognition, Technology & Work*, *13*, 245–258. doi:10.1007/s10111-010-0164-1

Kleinginna, P. R. & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, *5*, 345–379. doi:10.1007/BF00992553

Laerd. (2018). Kendall's tau-b using spss statistics. https://statistics.laerd.com/spss-tutorials/kendalls-tau-b-using-spss-statistics.php. Accessed: 2018-05-30.

Lane, R. D., Chua, P. M.-L., & Dolan, R. J. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia*, *37*, 989–997. doi:https://doi.org/10.1016/S0028-3932(99)00017-2

Lang, P. J., Greenwald, M. K., Bradley, M. M., & Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, *30*, 261–273. doi:10.1111/j.1469-8986.1993.tb03352.x

Larson, R. & Csikszentmihalyi, M. (2014). The experience sampling method, 21–34. doi:10.1007/978-94-017-9088-8_2

Lenberg, P. & Feldt, R. (2018). Psychological Safety and Norm Clarity in Software Engineering Teams. *CHASE'18*, 1–7. doi:10.1145/1235

Lenberg, P., Feldt, R., & Wallgren, L.-G. (2014). Towards a behavioral software engineering. In *Proceedings of the 7th international workshop on cooperative and human aspects of software engineering* (pp. 48–55). CHASE 2014. ACM. doi:10.1145/2593702.2593711

Lenberg, P., Feldt, R., & Wallgren, L.-G. (2015). Human factors related challenges in software engineering – an industrial perspective. In *2015 ieee/acm 8th international workshop on cooperative and human aspects of software engineering* (pp. 43–49). doi:10.1109/CHASE.2015.13

Lewin, K. (1935). *A Dynamic Theory Of Personality*. McGraw Hill Book Company Inc.

Li, F., Bai, X., & Wang, Y. (2013). The scale of positive and negative experience (spane): Psychometric properties and normative data in a large chinese sample. *PLoS One*, *8*, e61137.

Mehrabian, A. & Russell, J. (1974). An approach to environment psychology.

Meimann, A. (2016). *Positive emotions in daily life, measured by means of an experience sampling application.*

Meneghel, I., Salanova, M., & Martínez, I. M. (2016). Feeling good makes us stronger: How team resilience mediates the effect of positive emotions on team performance. *Journal of Happiness Studies*, *17*(1), 239–255. doi:10.1007/s10902-014-9592-6

Newman, J. (1977). The processing of two types of command statement: A contribution to cognitive ergonomics. *IEEE Transactions on Systems, Man, and Cybernetics*, *7*, 871–875.

Pew, R. W., Rollins, A. M., & Williams, G. A. (1976). Generic man-computer dialogue specification: An alternative to dialogue specialists. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *20*, 251–254.

Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. New York: Harper and Row.

Plutchik, R. & Kellerman, H. (1980). *Emotion, theory, research, and experience*. Academic Pr. doi:10.1017/S0033291700053769

Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones. *Sociological Methods & Research*, *37*, 426–454. doi:https://doi.org/10.1177/0049124108330005

Robinson, G. K. (1991). That blup is a good thing: The estimation of random effects. *Statistical Science*, *6*(1), 15–32.

Rudy Ramsey, H. & Atwood, M. (1979). Human factors in computer systems: A review of the literature. *79*, 180.

Safdar, U., Badir, Y. F., & Afsar, B. (2017). Who can i ask? how psychological safety affects knowledge sourcing among new product development team members. *The Journal of High Technology Management Research*, 79–92. doi:https://doi.org/10.1016/j.hitech.2017.04.006

Sajaniemi, J. (2008). Guest editor's introduction: Psychology of programming: Looking into programmers' heads. *Human Technology: An Interdisciplinary Journal on Humans in ICT Environments*, *4*, 4–8.

Samuels, P. & Gilchrist, M. (2014). Statistical hypothesis testing. Retrieved June 4, 2018, from https://www.researchgate.net/publication/275018715_Statistical_Hypothesis_Testing

Sánchez, J. A., Kirschning, I., Palacio, J. C., & Ostróvskaya, Y. (2005). Towards mood-oriented interfaces for synchronous interaction. In *Proceedings of the 2005 latin american conference on human-computer interaction* (pp. 1–7). CLIHC '05. ACM. doi:10.1145/1111360.1111361

Scarnati, J. T. (2001). On becoming a team player. *Team Performance Management: An International Journal*, *7*, 5–10. doi:10.1108/13527590110389501

Schwarz, N. (1990). Feelings as information: Informational and motivational functions of affective states. *2*.

Schwarz, N. & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, *45*, 513–523. doi:10.1037/0022-3514.45.3.513

Scollon, C., Kim-Prieto, C., & Diener, E. (2003). Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies*, *4*(1), 9–27. doi:10.1023/A:1023605205115

StockholmsHandelskammare. (2017). Programmerare – vanligaste yrket i stockholmsregionen. https://www.chamber.se/rapporter/programmerare-vanligaste-yrket-i-stockholmsregione.htm. Accessed: 2018-05-30.

Tavakol, M. & Dennick, R. (2011). Making sense of cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55.

Thompson, E. R. (2007). Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of Cross-Cultural Psychology*, *38*, 227–242. doi:10.1177/0022022106297301

Treude, C., Figueira Filho, F., & Kulesza, U. (2015). Summarizing and measuring development activity, 625–636. doi:10.1145/2786805.2786827

Tuckman, M., B. Jensen. (1977). Stages of small group development.

Watson, D. & Clark, L. A. (1992). Affects separable and inseparable: On the hierarchical arrangement of the negative affects. *Journal of Personality and Social Psychology*, *62*(3), 489–505.

Watson, D. & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin*, *98*(2), 219–235.

Weber, D., Voit, A., Kratzer, P., & Henze, N. (2012). In-situ investigation of notifications in multi-device environments. *Proceedings of the 2016 ACM Interna-*

*tional Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*. doi:10.1145/2971648.2971732

Wegge, J., van Dick, R., Fisher, G. K., West, M. A., & Dawson, J. F. (2006). A test of basic assumptions of affective events theory (aet) in call centre work. *British Journal of Management*, *17*, 237–254. doi:10.1111/j.1467-8551.2006.00489.x

Weinberg, G. M. (1971). *The psychology of computer programming*. Van Nostrand Reinhold.

Weiss, H. M. & Cropanzano, R. (1996). Affective events theory: A theoretical discussion of the structure, cause and consequences of affective experiences at work.

Welch, K. B., Galecki, A. T., & West, B. T. (2014). *Linear mixed models: A practical guide using statistical software, second edition*. London: Chapman and Hall/CRC.

Wohlin, C. (2012). *Experimentation in software engineering*. Springer, Berlin, Heidelberg.

# A

# Mail to participants

The following text were sent out to the participants a couple of days before the ESM study started via email.

Hello participant,

You have probably already been informed by your Scrum master about an upcoming study that you will participate in. This study is conducted by me, Kevin H Griffith, and my thesis partner Erik Nguyen as the last part of the master thesis at Chalmers.

The main goal of the thesis is to understand the impact of human factors on both team and individual performance. The study that you will participate in is conformed of a method called Experience Sampling Method (ESM), which, in short, is used to collect data throughout your daily work by sending out notifications to answer self-reports. More specific, the study will consist of the following:

- Questionnaire on participant data (e.g. age, gender, years of working experience)
- Questionnaire on self-assessed overall team performance
- ESM self-reports twice a day throughout a 3 week sprint
- Questionnaire on self-assessed team performance based on the ESM sprint

**ESM Study**
ESM will be carried out on your mobile device, meaning that you will have to download an application. Follow the steps in the "ESM Guide" at the end of this pdf to correctly setup the app.

We want to emphasize the fact that the study is anonymous, meaning that you as a participant will not be mapped to any of the answers that are collected. Every answer is important and we hope that you will bear with us throughout the upcoming 3 weeks by answering as many self-reports as possible.

Each self-report in the ESM study consists of 6 total questions, 3 on human factors and 3 on individual performance.

Regarding the questions on human factors, they will be presented pictorially where you will make assessments by selecting the figure that correlates to your current feelings the most. Further explanation will be available in the ESM study.

Regarding the questions on performance, they will be presented as regular questions on a scale from 'strongly disagree' to 'strongly agree'. Answer the question based on your current feelings.

There are no right or wrong answers. No answers to any questionnaire can be traced back to you. Please answer as honestly as you can.

**The other questionnaires**
Questionnaires on participant data and self-assessed overall team performance will be sent out at the beginning of the ESM study. The questionnaires will only be open until 27/4 and should only be answered once.

Questionnaire on self-assessed team performance based on the ESM sprint will be sent out at the end of the sprint. You answer this questionnaire only once.

Answer as honestly as you can. Remember, your answers cannot be traced back to you!

**How to reach us**
The easiest way to reach us is by email. We will answer any question you have regarding the study. griffith@student.chalmers.se eriknq@student.chalmers.se

We are also available in the office at the beginning of the study. When entering the building, you will find us on the right side of the building.

**ESM Setup Guide**
1. Download the app "Expimetrics" on your Android or iOS device.
2. Open the app and create a new account. Use a fake name. Use an email that you have access to (for login reasons).
3. Skip the demographic profile form.
4. Enable notifications. This is crucial for the study (Automatically on Android devices).
5. Enter the experience code. (code XXXX).
6. Start the experience.
7. See if any questionnaires are available. Else, wait for notifications to answer.

**Answering the ESM Study**
1. Wait for notification which will come Monday-Friday 10am and 14pm for three weeks.
2. You have one hour to answer before it expires.
3. Simply answer the questionnaire.
4. If the pictures in the ESM questionnaire will not load, please refer to the attached image to view it.

# B

# Participant information survey

This appendix presents the survey for participant information.

With this questionnaire, we are hoping for you to provide us with general information about yourself. Your answer will not be used to identify you or to associate you with answers from the ESM study. Rather, this information will be presented in our final thesis report to provide information about the participants of the study.

1. Your age
2. Your gender
    a. Female
    b. Male
    c. Prefer not to say
    d. Other
3. Years of working experience in IT
4. Team role
    a. Developer
    b. Designer
    c. Tester
    d. Manager
5. Programming language(s)
6. Programming language(s) experience