

UTRECHT UNIVERSITY

MASTER THESIS  
BUSINESS INFORMATICS

---

# Industry-Pushed Ways Of Working In User Story Refinement: How Good Are They Really?

---

*Author:*  
Jasper Berends  
J.Berends2@students.uu.nl  
*Student Number:* 5516080

*Thesis Supervisor:*  
Fabiano Dalpiaz  
F.Dalpiaz@uu.nl

*Daily Supervisor:*  
Beau Verdiesen  
Beau.Verdiesen@infosupport.com

*Thesis Supervisor 2:*  
Sjaak Brinkkemper  
S.Brinkkemper@uu.nl

July 31, 2020

*This thesis is submitted in fulfilment of the requirements  
for the degree of Master of Science  
in the*

Master in Business Informatics  
Graduate School of Natural Sciences



**Utrecht University**





## Acknowledgements

During my thesis, I have received help and support from a lot of people. First and foremost, I want to express my gratitude for all the supervision and guidance that I received from my supervisor dr. Fabiano Dalpiaz. I could not have achieved what I did if it was not for his expertise, frequent feedback and support.

Secondly, I want to thank my daily supervisor at Info Support, Beau Versiesen. His practical insights really helped to improve this work and to validate the feasibility and relevance of the research and he also helped me out a lot by proofreading my work. I also want to thank dr. Sjaak Brinkkemper for taking the time to review my thesis as the second supervisor of this project.

Additionally, my sincere thanks go out to my roommates, friends, fellow graduate interns at Info Support and my family. Their moral support has been invaluable to me and I could not have made my thesis without them.

Finally, I want to thank all students and team members that participated in my research. My thesis results depended heavily on their willingness to participate in my research, for which I will be forever grateful.

Jasper Berends, July 31, 2020





# Contents

<b>1</b>	<b>Introduction</b>	<b>5</b>
<b>2</b>	<b>Research Design</b>	<b>7</b>
2.1	Research Questions . . . . .	7
2.2	Relevance . . . . .	8
2.2.1	Scientific Relevance . . . . .	8
2.2.2	Practical Relevance . . . . .	9
2.3	Research Approach . . . . .	9
2.3.1	Literature Research . . . . .	10
2.3.2	Case studies . . . . .	11
2.3.3	Controlled experiment . . . . .	11
<b>3</b>	<b>Literature study</b>	<b>13</b>
3.1	Requirements Engineering for Software Development . . . . .	13
3.1.1	Challengers in RE . . . . .	14
3.1.2	Agile Software Development . . . . .	14
3.2	Behaviour-Driven Development . . . . .	15
3.2.1	Process-Deliverable Diagram . . . . .	16
3.2.2	BDD Applications and Advantages . . . . .	19
3.2.3	Example Case . . . . .	20
3.3	Three Amigo Sessions in Requirements Engineering . . . . .	21
3.3.1	Example Mapping . . . . .	24
3.3.2	Feature Mapping . . . . .	26
3.4	Shared Understanding . . . . .	29
<b>4</b>	<b>Treatment Design</b>	<b>37</b>
4.1	Case Study Design . . . . .	37
4.1.1	Longitudinal Case Study . . . . .	37
4.1.2	Data Analysis . . . . .	40
4.1.3	Validity . . . . .	40
4.2	Controlled Experiment . . . . .	45
4.2.1	Context . . . . .	45
4.2.2	Planning and Presentation . . . . .	46
4.2.3	Student Handout Package . . . . .	47
4.2.4	Execution of Experiment . . . . .	48
4.2.5	Validity . . . . .	49
4.3	Measuring technique performance . . . . .	52
4.3.1	Questionnaires . . . . .	52

4.3.2	Experimental Output Evaluation . . . . .	55
<b>5</b>	<b>Treatment Validation</b>	<b>57</b>
5.1	Controlled experiment - Requirements Engineering course . . . . .	58
5.1.1	Discussion . . . . .	58
5.1.2	Overall Results . . . . .	59
5.1.3	Results per Aspect . . . . .	62
5.1.4	Results per Group . . . . .	63
5.1.5	Output Analysis . . . . .	64
5.1.6	Conclusion . . . . .	66
5.2	Case Study - Fizzor Example Mapping . . . . .	67
5.2.1	Session Results . . . . .	68
5.2.2	End Evaluation . . . . .	72
5.2.3	Conclusion . . . . .	73
5.3	Case Study - Pension Management Firm Example Mapping . . . . .	74
5.3.1	Evaluation after session 4 . . . . .	78
5.3.2	End Evaluation . . . . .	79
5.3.3	Conclusion . . . . .	79
5.4	Case Study - Pension Management Firm Feature Mapping . . . . .	80
5.4.1	Conclusion . . . . .	81
5.5	Case Study - Example Mapping Online Tryout . . . . .	82
5.5.1	Conclusion . . . . .	83
<b>6</b>	<b>Conclusion</b>	<b>85</b>
6.1	Cross-case Conclusions . . . . .	85
6.2	Answers to Research Questions . . . . .	87
6.2.1	<b>RQ1:</b> What defines a Three Amigo session? . . . . .	87
6.2.2	<b>RQ2:</b> How can the performance of TA sessions be measured? . . . . .	87
6.2.3	<b>RQ3:</b> How do TA sessions perform when used for the first time? . . . . .	88
6.2.4	<b>RQ4:</b> How do TA sessions perform after becoming familiar with the technique? . . . . .	89
6.2.5	<b>MRQ:</b> How do defined Three Amigo session techniques perform for user story refinement? . . . . .	89
<b>7</b>	<b>Discussion</b>	<b>91</b>
7.1	Limitations . . . . .	91
7.2	Future Work . . . . .	92
<b>8</b>	<b>Bibliography</b>	<b>93</b>
<b>A</b>	<b>Questionnaire</b>	<b>99</b>
A.1	Complete Technique Questionnaire . . . . .	99
A.2	TA Session Questionnaire . . . . .	100
A.3	Post-Implementation Questionnaire . . . . .	101
<b>B</b>	<b>Case Study Forms</b>	<b>102</b>
B.1	Informed Consent . . . . .	103

<b>C</b>	<b>Controlled Experiment</b>	<b>105</b>
C.1	User Story Outputs – Example Mapping . . . . .	105
C.2	User Story Outputs – Feature Mapping . . . . .	108
C.3	Lecture Slides . . . . .	110
C.4	Student Handout Package . . . . .	138
C.5	Questionnaire Results . . . . .	150
C.5.1	Results per Technique - Combined . . . . .	150
C.5.2	Results per Aspect . . . . .	151
C.5.3	Results per Group . . . . .	154
C.6	TA Session Outputs . . . . .	160
C.7	Output Analysis . . . . .	167
C.7.1	US1 Aggregation . . . . .	167
C.7.2	US2 Aggregation . . . . .	168
<b>D</b>	<b>Case Study - Fizzor</b>	<b>169</b>
D.1	Correlation Matrices . . . . .	169
D.2	Participant Results . . . . .	170
D.3	Session Results . . . . .	171
<b>E</b>	<b>Case Study - Pension Manager Example Mapping</b>	<b>173</b>
E.1	Correlation Matrices . . . . .	173

### Abstract

New techniques for refining requirements in software development are frequently presented. However, these techniques are often not scientifically validated on their performance. In this study, we aim to validate two refinement techniques, Example Mapping and Feature Mapping. These two techniques are both Three Amigo session techniques, in which people from different work together to refine the software increment. Shared understanding is an essential aspect of these sessions. We present a new measurement tool with which the performance of refinement techniques can be investigated, which also incorporates shared understanding. We investigate the performance of the refinement techniques with a controlled experiment and four case studies. Based on the results, Example Mapping performs well under certain conditions. We observed a learning effect for the technique, which resulted in better performance of sessions after having used the technique several times. For Feature Mapping, results were inconclusive and additional research is required to establish its performance. When conditions are not right for Example Mapping or Feature Mapping, Three Amigo session principles can still be applied to refinements to result in a well-performing session. Further research could be undertaken to generalise this study's findings further, studying the long-term effects of the techniques on the implementation of a user story and studying the performance of the techniques when team members are co-located. Additionally, future research can be conducted to compare the Three Amigo session techniques to other refinement techniques, to further investigate the validity of the performance measurement tool and to study the performance of shorter Three Amigo sessions.

**Keywords:** Three amigo sessions, Example mapping, Feature mapping, Shared understanding, Refinement

# Chapter 1 | Introduction

Software development with the use of Agile methodologies requires a shared understanding of requirements between all stakeholders. Informal and frequent communication has been considered a crucial part of Agile Requirements Engineering (RE) [53]. RE is an integral part of software development and plays a significant role in cost-effective software development [45]. On the other hand, bad requirements are adverse for a software project and often lead to higher costs [31].

For refining requirements, a Three Amigos session can be organised. The main goal of these sessions is to gain a shared understanding and a clear description of the underlying requirements of an increment of work. With a Three Amigos session, at least three people should be present that together have good knowledge of business analysis, software development and quality assurance domains. Examples of what an increment's functionality (such as a user story) should do are often the outcome of these sessions, written in a ubiquitous language that is understandable to everyone [5].

Because Agile environments often require software engineering practises to adapt continually, new techniques for software engineering and RE are presented frequently as well. Examples of these new techniques are Example Mapping (EM) [72], which is presented by Matt Wynne, a well-established figure within the area of Behaviour-Driven Development[15], and Feature Mapping (FM), which is a more structured approach compared to EM. Both EM and FM are techniques that organise Three Amigo sessions in a particular manner.

However, a potential risk with techniques like these is that they are often not scientifically validated on their performance. With established people from this industry "pushing" their presented techniques onto teams without any validation on performance, the risk is that techniques may be implemented in settings where they are no good fit, or that the techniques are not well-performing altogether. Therefore, this research aims to validate how well-performing such industry-pushed techniques actually are by focusing on the Three Amigo session techniques Example Mapping and Feature Mapping.

This thesis is structured as follows. First, in Chapter 2, the research design is described that we followed for this project. Chapter 3 gives a comprehensive overview of the literature study that was conducted in order to get a proper theoretical foundation of the research. The design of the execution phase of the research is presented in Chapter 4. For validating how EM and FM perform, an evaluation method must first be created. As TA session techniques put much emphasis on shared understanding, this will be an essential aspect of the evaluation of the techniques. In Chapter 5, the results of the research shall be presented. The conclusions of the research are elaborated on in Chapter 6. Lastly, the thesis is finalised in Chapter 7, where limitations of this research and future work based on it are discussed.



# Chapter 2 | Research Design

In order to conduct scientific research, a proper research design must first be created. The research design of this thesis is described in this chapter. First, we specify the research questions in Section 2.1. After that, Section 2.2 explains why this research has both scientific and practical relevance. In Section 2.3, the research approach of this thesis is explained.

## 2.1 Research Questions

The Main Research Question (MRQ) is based on the research objectives of researching how well Three Amigo (TA) session techniques work and is defined as follows:

**MRQ:** How do defined Three Amigo session techniques perform for user story refinement?

The hypothesis is that TA session techniques perform well for user story refinement. The concrete examples will help to illustrate the acceptance criteria in a format that is easy to understand, and the session will help increase the shared understanding of team members. In order to answer the **MRQ**, the following underlying Research Questions (RQ's) and corresponding sub-questions are defined:

**RQ1:** What defines a Three Amigo session?

**RQ1.1:** What TA session techniques exist?

**RQ1.2:** Where do TA sessions fit in Requirements Engineering?

**RQ1.3:** Where do TA sessions fit in Software Engineering?

In order to answer the MRQ, we need to understand the basics of TA sessions, what techniques exist, and how TA sessions fit in Requirements Engineering and Software Engineering. RQ1 will be answered with a scientific literature study.

**RQ2:** How can the performance of TA sessions be measured?

**RQ2.1:** How can shared understanding be measured?

**RQ2.2:** How can the performance of a user story refinement technique be measured?

As with RQ1, RQ2 will be answered by a literature study. First, a deeper understanding of what shared understanding means and encompasses will be investigated, followed by ways to test it. Besides that, the overall performance of user story refinement techniques needs to be measured. Together, these two aspects will serve as a way of measuring the performance of a TA session.

**RQ3:** How do TA sessions perform when used for the first time?

**RQ3.1:** How does a first Example Mapping session perform?

**RQ3.2:** How does a first Feature Mapping session perform?

**RQ4:** How do TA sessions perform after becoming familiar with the technique?

**RQ4.1:** How does Example Mapping perform after becoming familiar with the technique?

**RQ4.2:** How does Feature Mapping perform after becoming familiar with the technique?

The execution of the research answers RQ3 and RQ4. RQ3 is answered by conducting a study in a controlled experiment in which the TA session techniques can be evaluated more closely, as well as with the first sessions of all case studies. RQ4 is answered by conducting case studies at real-world software development teams.

The hypothesis is that TA sessions perform well both during the first session, as well as after becoming familiar with the techniques. However, we believe that there is a learning effect that will make the techniques perform better after having used them several times. We expect that both Example Mapping (EM) and Feature Mapping (FM) will perform well. People may, however, strongly favour one over the other. For some types of people or teams either EM or FM may be more suitable, and this will be visible in the results. Altogether, this will hypothetically not result in a big difference between the results of EM and FM, as the number of people that favour EM is the same as the number of people that favour FM.

**RQ5:** Do TA sessions have effects on the implementation of a user story?

Lastly, with RQ5, we aim to investigate the effects of TA sessions in a broader perspective of the software development life cycle. If the case studies last long enough, then user stories that were refined using TA session techniques will be in software products, and the effects of the TA sessions can be evaluated outside of the session itself. The hypothesis here is that the outputs of the TA sessions will positively influence the implementation of a user story due to the concrete examples and due to the shared understanding that team members have gained during the session.

## 2.2 Relevance

### 2.2.1 Scientific Relevance

This study provides a scientific contribution as no research has been done before that analyses Three Amigo sessions, despite the term itself being around for over a decade already [16]. Requirements Engineering (RE) is an integral part of software development and has a big impact on the success of a project [27]. Therefore, it is valuable to research techniques that are used in RE on their performance. This research will present a performance measurement tool to evaluate refinement techniques that focus on shared understanding, which can also later be used and adapted for other techniques.

Besides that, previous publications have also shown the importance of shared understanding in RE and that new approaches are needed that put emphasis on this [8]. Hoffmann et al. share this view and “call for research that explores ways of systematically building mutual and shared understanding in the development process” [26]. As TA session techniques put much emphasis on shared understanding, researching them has significant scientific relevance.



### 2.2.2 Practical Relevance

Shared understanding in software development teams has been shown to improve software quality [2] as well as team performance [52]. As Three Amigo session techniques put emphasis on shared understanding, it is relevant for teams to know whether or not these techniques perform well. Besides shared understanding, knowing if the techniques generally perform well is also of practical relevance. Insight in what makes these techniques good or bad can provide valuable knowledge. Based on this research, practitioners can either recommend or discourage the use of Three Amigo sessions.

## 2.3 Research Approach

In order to answer the MRQ and underlying research questions, the techniques must be evaluated somehow. Firstly, a literature review will be conducted to give insights into the first two research questions and to relate TA sessions to related research. For researching the performance of the techniques in practice, we find the best suitable way for this to be through application of the techniques and evaluating its performance afterwards. Therefore, one important aspect of this research will be case studies. As such, the case study research method is adhered to, as presented by Yin [76].

Yin defines a case study as being two-fold, covering both the scope and features of a case study. Looking at the scope, a case study is an empirical method in which a phenomenon is researched in a real-world context, especially when there are no clear boundaries between the phenomenon being researched and its context. Whereas a controlled experiment separates the artefact that is being researched (in this case, TA sessions), case studies have no such advantage.

As a second part of the definition, Yin describes that a case study “copes with the technically distinctive situation in which there will be many more variables of interest than data points and as one result benefits from the prior development of theoretical propositions to guide design, data collection, and analysis, and as another result relies on multiple sources of evidence, with data needing to converge in a triangulating fashion” [76]. With this second part of the definition, Yin emphasises that a case study is not merely a data collection tactic or a design feature, which he also based on the work of Stoecker [61].

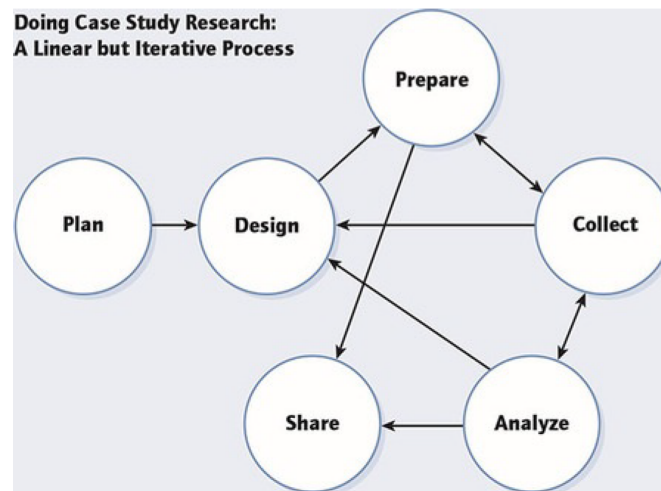


Figure. 2.1: Case Study Method Process [76]

In Figure 2.1, the different steps of the case study method are visualised. As case studies also have a lot of external factors that may influence the research, it will be combined with a controlled experiment. In the controlled experiment, the TA session techniques can be examined in a more isolated environment without possible outside factors influencing the results. As there is a big overlap in the way the techniques are tested in both the case studies and the experiment, the case study processes will be adhered to for both. Linearly, a case study would follow the following sequence of processes:

*Plan → Design → Prepare → Collect → Analyse → Share.*

The first step of the case study method is planning the research. In this phase, the situation is analysed and the research method is chosen. In this research, this step was already performed and the case study method was chosen. Other methods that were considered were design science and action research. However, design science did not fit since no new artefacts are created during this research that are evaluated. Instead, this research examines existing artefacts, the TA session techniques, through a newly created way of measuring performance. Action research also seemed undesirable, as the case studies would involve domains that are unfamiliar to us as researchers. Therefore, participating in the research would make it ineffective and observation is more suitable. This led to the conclusion that the case study method was most suitable for this research.

Secondly, researchers must design the case study. This involves defining the case(s) to be studied, building theory, identifying the type of case study design and testing the design against quality criteria. This step is covered in Chapter 3 and Chapter 4. The third step is also encompassed in Chapter 4, where the case studies are prepared and a protocol is made.

The fourth step in the process is collecting case study evidence, which means performing the case studies, and creating and collecting results from them. Chapter 5 will serve for this purpose. In the fifth step, the collected data is analysed in order to draw conclusions from the case study, which Chapter 6 is for. Lastly, the final step of a case study research is sharing the results. This document serves that purpose.

In Table 2.1, an overview is given on what aspects of the research will cover what RQ's. An exploratory literature review will be performed for answering RQ1 and RQ2. RQ3 is answered by the experiment and by the case studies. Only case studies answer RQ4 and RQ5 as they require longitudinal studies and measuring the longer-term effects of TA sessions in a real-world software implementation, which cannot be verified with our experiment.

	Literature	Case studies	Experiment
RQ1	X		
RQ2	X		
RQ3		X	X
RQ4		X	
RQ5		X	

**Table 2.1:** Sources for answering RQ's

### 2.3.1 Literature Research

Literature research is conducted to answer RQ1 and RQ2. First, the broader fields of Requirements Engineering and Software Development are researched. Shared understanding is also researched in literature considering its claimed importance in Three Amigo sessions. Google

Scholar is primarily used as a search engine, followed by ResearchGate. In the first place, only publications from 2013 and later are considered. From those results, forward and backwards snowballing is used on relevant results in order to find more valuable literature [71]. Forward snowballing refers to finding more recent work from a publication by analysing sources that have cited the publication. Backward snowballing refers to analysing which sources have been cited by the publication that is being reviewed.

Grey literature will be analysed for Three Amigo sessions due to the lack of scientific literature on this topic. This consists of online, non-scientific publications and web pages. Finally, literature that is written or recommended by Utrecht University is also considered.

### **2.3.2 Case studies**

Case studies will be executed by introducing TA session techniques to real development teams. This way, the techniques are tested in a real-world context that closely resembles the intended implementation context. In fact, they will be introduced in a way that perfectly matches the way of implementing the techniques as intended, with the only exception being that they are now implemented scientifically with an additional performance evaluation.

Depending on the duration of the case studies and of the availability of teams, the case studies will be conducted over a period in which the teams do a number TA sessions for each technique, measuring the performance after each session. This way, a possible learning curve of the techniques may also be identified.

### **2.3.3 Controlled experiment**

During the case studies, many outside factors may influence the TA session performance. For example, participating teams may have important deadlines that render them unable to attend TA sessions or to put serious effort into it. It can also happen that other obstacles are preventing a team from working effectively and that those obstacles diminish or remove expected benefits from TA sessions. Therefore, the case studies are complemented with a controlled experiment in which such threats should not exist.



## Chapter 3 | Literature study

In order to create a proper foundation for this thesis, literature research is conducted. A theoretical basis is part of the “Design” step in the case study method and will help strengthen the research and make it easier to generalise the findings [76]. This literature research starts with a section on Requirements Engineering and Agile Software Development in Section 3.1. Following this is Section 3.2 in which Behaviour-Driven Development (BDD) is discussed. BDD is a software development method from which Three Amigo (TA) sessions originate which is why its characteristics are investigated into detail. Following is Section 3.3 in which TA sessions are researched and elaborated on. Defined TA session techniques are also researched into detail in this section. Lastly, literature on shared understanding (SU) is also researched and we present the results in Section 3.4. This chapter answers RQ1 “What defines a Three Amigo session?” and a foundation is laid for answering RQ2 “How can the performance of TA sessions be measured?”

### 3.1 Requirements Engineering for Software Development

Requirements Engineering (RE) is an integral part of software development and has a significant impact on the success of a project [27]. According to Van Lamsweerde, RE can be defined as “a coordinated set of activities for exploring, evaluating, documenting, consolidating, revising and adapting the objectives, capabilities, qualities, constraints and assumptions that the system-to-be should meet based on problems raised by the system-as-is and opportunities provided by new technologies” [68].

In this definition, Van Lamsweerde speaks of a system rather than a software product. This is because Van Lamsweerde considers the tasks of a Requirements Engineer to be broader than just specifying software requirements. Besides software, he instead finds a system also to consider people, devices and existing software. In fact, the services that are implemented to fulfil the stakeholders’ goals are assigned not only to software but also to hardware and people.

Van Lamsweerde identifies four main processes of the RE life cycle, as can be seen in Figure 3.1 [68]. These processes are not sequential but instead go in a spiral. It starts with domain understanding and elicitation, in which the system-as-is is analysed together with objectives. After that, informed decisions must be made in the second phase based on the issues that have arisen during the elicitation process. In the third phase, specification and documentation, a requirements document is created in which all characteristics of the system-to-be are documented. Lastly, the last phase of the RE life cycle is verifying the documented requirements with stakeholders to make sure that they are correct and complete.

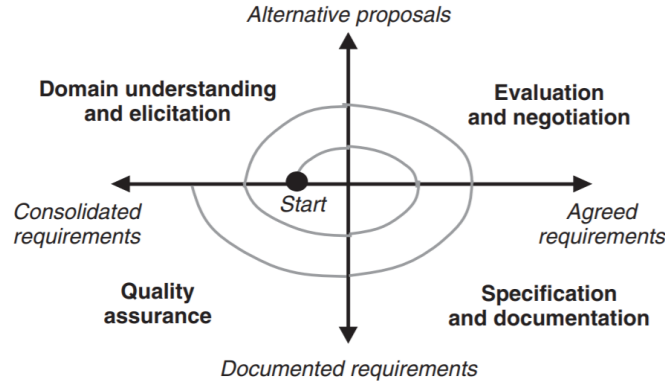


Figure. 3.1: RE Processes [68]

### 3.1.1 Challengers in RE

Inayat et al. performed a systematic literature review on practices and challenges in traditional and Agile RE[29]. One of the challenges they have identified traditional RE is communication issues. Agile RE could solve this with frequent face-to-face meetings in which the customer participates and can interact with the team.

Inayat et al. also mentioned that customers are not always available and that that is a big challenge for Agile RE. Fernández et al. also said this in their research: “Small and agile companies seem to suffer especially from customers not willing to participate with a considerable amount of time” [19]. They obtained data from 228 companies to conduct their research and list the following as the top 10 RE problems as a result:

1. “Incomplete and / or hidden requirements
2. Communication flaws between project team and customer
3. Moving targets (changing goals, business processes and / or requirements)
4. Underspecified requirements that are too abstract
5. Time boxing / Not enough time in general
6. Communication flaws within the project team
7. Stakeholders with difficulties in separating requirements from known solution designs
8. Insufficient support by customer
9. Inconsistent requirements
10. Weak access to customer needs and / or business information” [19].

From this list can be concluded that communication is a pressing issue in RE, both with a customer and within a project team (the second, sixth, seventh and tenth item in the list). Besides that, incomplete or incorrect requirements are also a major issue and is listed several times in the top 10 (the first, fourth and ninth item refer to this). Therefore, research into techniques such as TA sessions is of importance as those could solve such issues.

### 3.1.2 Agile Software Development

Agile Software Development (ASD) methods have emerged rapidly over the past decades as a solution for processes that required heavy up-front planning [1, 25] and had limitations such as “slow speed of delivery, difficulty in handling changing requirements, and formalised

documentation” [56]. To overcome these issues, the Agile Manifesto was presented by Beck et al. in 2001 [7]. As a report from 2019 shows, 54% of all organisations that work in an Agile manner use Scrum, with another 18% using an adapted version of Scrum [69]. Scrum is “A framework within which people can address complex adaptive problems, while productively and creatively delivering products of the highest possible value” [55]. With the use of the Scrum framework, software development teams work on an iterative and incremental product that is releasable after one “Sprint”. During one Sprint, the team works Product Backlog Items (PBIs) that resemble an increment of work. User stories are often used as a template for a PBI and are proven to be effective [37]. A user story often looks as follows:

*As a <role>  
I want to <goal>  
So that <benefit>.*

A user story contains three parts: the role, goal, and reason [34]. The role refers to the person that wants this PBI realised. This can be any stakeholder that is involved in one way or another with the product that is being designed. Secondly, the goal shows what should be realised for this increment of work. Lastly, the desired effects of the PBI will be elaborated on in the benefit.

In Scrum, three roles are defined within a team: the Product Owner (PO), Scrum Master (SM) and Developer [54]. Note that this does not refer to the same *role* that was mentioned with a user story, as user stories can also have (and often do) someone as a role that is not part of the Scrum team. The PO represents the needs of the stakeholders and is responsible for creating and organising PBIs on the Product Backlog. The SM safeguards the Scrum principles and should make sure impediments to the team are removed or minimised as much as possible. A Developer is a person who works on realising the product of the Scrum team. In Scrum, no distinctions exist between, for example, a Software Developer, a Software Tester or Requirements Engineer: they are all considered a Developer. Development teams are “cross-functional, with all the skills as a team necessary to create a product Increment” [55].

Scrum is a framework that can be used by a software development team as part of a broader ASD life cycle. Before the rise of Agile frameworks, the Waterfall model was often used [50], which is a linear approach in which stages were performed sequentially rather than at the same time. One software development method in which Scrum or other Agile frameworks can be applied is Behaviour-Driven Development (BDD). As Three Amigo Sessions originate from BDD, this method will be discussed in-depth in the next section.

## 3.2 Behaviour-Driven Development

Many software development projects suffer from a communication gap between domain experts and software developers [17]. This is often due to the fact that domain experts, business analysts and requirements engineers use jargon that is difficult to understand for software developers, making it difficult for them to translate the needs of domain experts into software requirements and features. Behaviour-Driven Development (BDD) was proposed to fill this gap, among with other advantages, by extending Test-Driven Development (TDD) with ubiquitous language [60].

Test-Driven Development (TDD) was created by Beck [6] and focuses on writing tests before implementing the functionality for which the tests are written. New tests are written that are all supposed to fail, considering that the functionality should not exist yet, after which software

code is written until the tests succeed. Code can then be refactored if necessary, after which the process is complete [65].

Dan North, the creator of BDD, experienced resistance while trying to have programmers write test code instead of dedicated testers after production code had been completed. He experienced problems on three fronts, as he explains in the foreword of the book “BDD in Action” by Smart: “programmers did not want to write tests; testers did not want programmers writing tests; and business stakeholders did not see any value in anything that was not production code” [57]. In an attempt to improve TDD, North introduced the BDD method in 2006 [44].

Whereas TDD focuses on technical aspects in creating software, BDD focuses on user behaviour. It does so by stating that all specifications should be written in a ubiquitous language, which makes them understandable for both domain experts and developers at the same time. In the BDD method, test scripts are written and implemented before the application code itself is written. The main traditional phases (and with that deliverables) of BDD are illustrated in Figure 3.2, adapted from Smart [57].

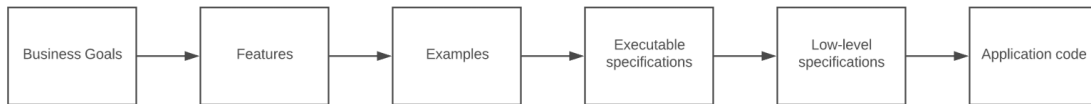


Figure 3.2: BDD phases and deliverables

In the first phase, business goals are defined that the software product should help accomplish. Secondly, features are defined, which mostly are functional software components. In the third phase, examples are written in the form of scenarios. These scenarios are the most important part of BDD around which the method is built. Executable specifications are generated from this in the fourth phase, but nothing is actually functionally tested yet in this phase. In the fifth phase, tests are made functional by writing test code. During the sixth and final phase, the application code is written, which is where the actual business value is generated. An example of how these phases work in practice is given in Section 3.2.3.

Ideally, three roles work closely together in BDD: the product owner, software developer and tester [74]. The “Three Amigos” work closely together in most phases. Together, they create features that can be elaborated on more in user stories [34]. Afterwards, they define examples for each feature, widely referred to as scenarios. Scenarios are implemented in test scripts, which are the deliverables for the next two phases. As the software itself has not been written yet, these tests are built to fail in advance. Therefore, in the last phase of BDD, the application code is the main deliverable. Application code is written until all tests succeed. At this point, assuming that the scenarios from the test script cover all functionality, the application code should cover everything that was intended to be implemented in the way it is supposed to work.

### 3.2.1 Process-Deliverable Diagram

In this section, the Process-Deliverable Diagram (PDD) of the BDD method is presented. PDD’s are introduced by Brinkkemper [10] and are useful for “modelling activities and artefacts of a certain process” [66]. The PDD is presented in Figure 3.3.

The PDD is divided into the same six phases that were mentioned earlier. The first phase is covered by the closed complex activity *Define business goals*, which generates the BUSINESS GOAL concept. As BDD specifies no specific way of doing this, the exact steps one takes to define business goals is considered out of the scope of this research. In the second phase, features



are defined by the three amigos (product owner, software developer and tester) that support the BUSINESS GOAL. A FEATURE has at least a title and possibly USER STORY concepts to elaborate more on the FEATURE functionality.

In the third phase, scenarios are written by the three amigos. First, the SCENARIO with a title is defined, after which the three STEP DEFINITION concepts are defined. Defining the pre-condition, trigger and end state are unordered activities, i.e., the order of execution does not matter. After all three activities are complete, the three amigos can define additional conditions, but are not required to do so. Therefore, every SCENARIO aggregates 3 or more STEP DEFINITION concepts.

In phase 4, executable specifications are generated for the SCENARIO. For each STEP DEFINITION, one SPECIFICATION is generated. This generation can be automated by tools, especially SpecFlow for .NET and Cucumber for Java and other programming languages [57]. In order to keep the PDD organised, these tools are not displayed in the model. A SPECIFICATION has a prefix that helps refer to it and a status, which shows whether the SPECIFICATION is functional. In phase 5, TEST CODE is written that makes the generated SPECIFICATION functional. These activities are again unordered: one is able to choose his or her preferred order for executing this phase.

In the sixth and final phase, APPLICATION CODE is written for the SPECIFICATION. Once the APPLICATION CODE is written, one must run the specification tests in order to validate if the SPECIFICATION is properly implemented. If not, a loop forms and APPLICATION CODE needs to be changed until the tests do pass. Once they pass, the BDD method is complete.

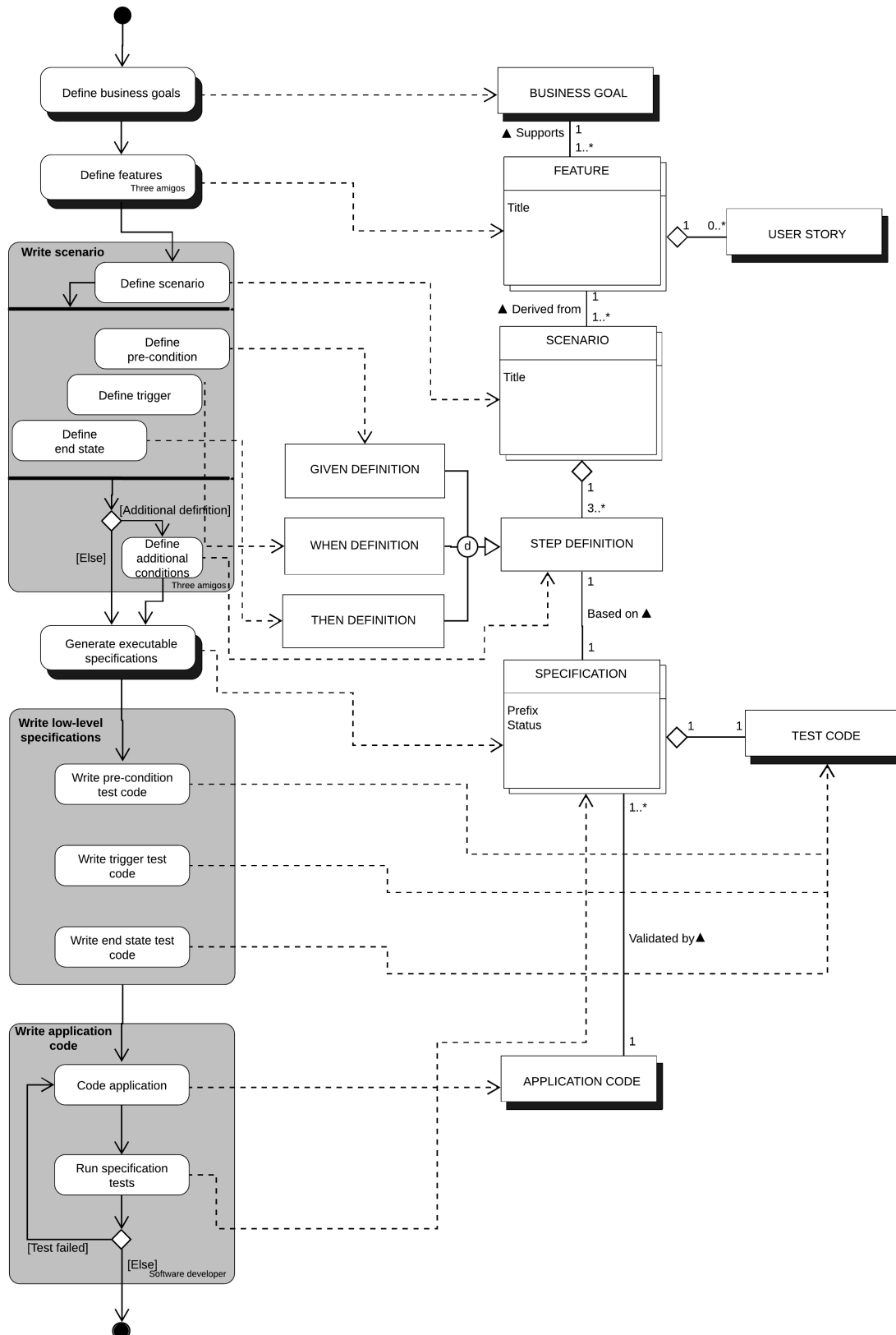


Figure 3.3: Process-Deliverable Diagram of BDD

### 3.2.2 BDD Applications and Advantages

As described earlier in this section, the term Behaviour-Driven Development (BDD) was introduced by Dan North in 2006 [44]. It originates from Test-Driven Development (TDD), which also focuses on writing failing tests before actually implementing application code. Nowadays, the term BDD is closely related to that of similar methods [3]. As Adzic points out, he has heard many names for the same method. In the 50 projects he investigated for his book, he has come across the following names:

- Agile acceptance testing
- Acceptance Test-Driven Development (ATDD)
- Example-Driven Development
- Story testing
- Behaviour-Driven Development (BDD)
- Specification by Example (SbE)

Throughout the literature, these terms are often also used interchangeably, with BDD, SbE and ATDD seemingly being the three most popular ones. In this thesis, the term BDD will be adhered to.

The main advantage of BDD over TDD is that it uses natural language and domain jargon to write and implement functional specifications and tests, while TDD has a technical focus. The fact that it uses natural language and domain jargon makes it possible for both software developers and domain experts to understand the written specifications. A domain-specific language (DSL) that was created for this very purpose is Gherkin [24]. Gherkin defines several keywords for step definitions: *Given*, *When*, *Then*, *And*, and *But*. *Given*, *When*, and *Then* respectively concern the pre-conditions, trigger and end state of one scenario.

Gherkin and BDD focus on writing specific examples rather than writing general, high-level requirements [20]. It is also possible to create an entire list of examples with one scenario in order to test functionality more thoroughly [43]. Nicieja also explains that using this method greatly aids living documentation of software systems. Tools are available to automatically extract all features and scenarios from software code, given that they are written in Gherkin, and output an overview of them as documentation. Augurk is an example of such a “living documentation” system [39].

Besides that, using BDD can also help avoid technical debt in software projects [64]. If intended user behaviour is automatically tested upon introducing new code to a software product, technical debt can be avoided or addressed more easily. Technical debt can negatively impact a product in the future, and the cost of correcting it may grow higher when it is not fixed right away [18]. This shows how BDD may increase the performance of a development team, as also suggested by Trumler and Paulisch [64].

In his book “Specification by Example”, Adzic elaborates on six case studies on BDD [3]. The first case study was uSwitch, a popular comparison website in the UK. Even though they did not plan on implementing BDD specifically, they did implement most of its aspects and optimised their performance significantly. The key advantage that uSwitch experienced from BDD was the ability to run automated tests, something that they first spent a lot of time on. The second case study was RainStor, a “company that builds high-capacity data archiving and management systems”. RainStor concluded that BDD helped them eliminate the need for two different sets of documents as test and requirements documents were merged. The other experienced benefit of BDD was that it helped them maintain a defined scope and that the right product was built and “not waste time developing unnecessary features.” In other cases Adzic

studied, the options of living documentation systems were described as a big advantage. Providing a clear target for what needs to be developed and the ability to continuously validate that target is also mentioned as a driving reason to use BDD.

Case studies performed by others also show positive results of BDD. For example, Park and Maurer find that using practices of BDD “helps communicate domain knowledge required to write software and can help software developers to communicate the status of the software implementation better” [48]. During their case study, the acceptance tests helped software engineers confirm that they understood the requirements, as the tests would not pass until the requirement was properly implemented. BDD has also been proved to help communicate and validate functional business requirements more clearly [38]. However, Melnik et al. do note BDD may have a high learning curve for domain experts who just started using the method.

### 3.2.3 Example Case

An example that may well illustrate the traditional practices of BDD is that of an e-commerce website. For instance, a local book store may want to move their business online by introducing an online store. The online presence is, therefore, the business goal and the first deliverable. A more detailed elaboration on how and why the book store wants to move their business online can be included with this deliverable.

With an online store as the business goal, features must be specified. As almost all online stores have the ability to save items in a *shopping cart*, this is an important feature. Other features can be a payment system, a search function and the ability for users to register on the e-commerce website. User stories can be used to elaborate on a feature more and, in the case of the shopping cart feature, would look as follows:

*As a visitor of the online book store  
I want to save books in a shopping cart  
So that I can buy multiple books at once.*

From this feature and corresponding user story, examples in the form of scenarios can be created. One scenario is adding a book to the shopping cart. A scenario deliverable with three *step definitions* looks as follows:

*Scenario: Adding a book to shopping cart*  
**Given** my shopping cart is empty  
**When** I add a book to my shopping cart  
**Then** my shopping cart should contain one book

A scenario should be concise, testable, understandable, unambiguous and valuable [46]. Using a scenario such as this one, test methods in code can be automatically generated using tools such as SpecFlow for .NET or Cucumber for Java and other languages [57]. Both tools have the same basic functionality with the only difference that other programming languages are targeted. Besides that, Cucumber is open source, while SpecFlow is not. Using a tool reduces the risks of creating duplicate code in the phases of writing executable and low-level specifications. Generating these executable specifications is the only activity of the fourth phase, and has a deliverable that is illustrated in Figure 3.4. These examples are created using SpecFlow and will look different when using other programming languages or tools.

The executable tests generated in the fourth phase do not perform any actions yet. Therefore, during the fifth phase, test methods are coded to make them usable. The deliverable of

```

[Binding]
public class ShoppingCartSteps
{
    [Given(@"my shopping cart is empty")]
    public void GivenMyShoppingCartIsEmpty()
    {
        ScenarioContext.Current.Pending();
    }

    [When(@"I add a book to my shopping cart")]
    public void WhenIAddABookToMyShoppingCart()
    {
        ScenarioContext.Current.Pending();
    }

    [Then(@"my shopping cart should contain one book")]
    public void ThenMyShoppingCartShouldContainOneBook()
    {
        ScenarioContext.Current.Pending();
    }
}

```

Figure. 3.4: Generated executable specifications

that is functional test code, as shown in Figure 3.5. The initial classes and methods needed for the functionality are also created in this phase, albeit still empty classes and methods.

Afterwards, real software code must be written that minimally passes the tests, which is the sixth and final phase of the method. In this case, that implies the feature of adding a book to the shopping cart should be made functional. After the application code is written, the developer must validate that it meets the specified requirements by running the functional tests written in the fifth state. In case the tests do not pass, that means that the application does not meet up to the scenario that was specified in the third phase. The developer must, then, evaluate what went wrong and change the code of the application in order to make the test pass. This is, of course, with the assumption that the specifications were correct and complete. If it turns out that something is amiss with the specifications, the BDD process would have to go back to a previous step. Once all functional tests pass, the process is finished and a shopping cart should be available that meets the functional requirements.

### 3.3 Three Amigo Sessions in Requirements Engineering

Looking at BDD, a common way of refining features and user stories is during a Three Amigo (TA) session. In a TA session, people from different disciplines come together to refine a user story. Originally, this meant having someone present in the workshop from the business, software development and quality assurance (i.e., testing) perspective. Organising TA sessions is expected to result in “a clearer description of an increment of work often in the form of examples, leading to a shared understanding for the team” [5].

Although TA sessions are widely incorporated in BDD as a good way of working, not much literature exists on how exactly to organise these sessions. In fact, no publications could be

```

[Binding]
public class ShoppingCartSteps
{
    ShoppingCart Cart;
    [Given(@"my shopping cart is empty")]
    public void GivenMyShoppingCartIsEmpty()
    {
        Cart = new ShoppingCart();
        Cart.EmptyCart();
    }

    [When(@"I add a book to my shopping cart")]
    public void WhenIAddABookToMyShoppingCart()
    {
        Book BDD = new Book("BDD IN ACTION");
        Cart.AddBook(BDD);
    }

    [Then(@"my shopping cart should contain one book")]
    public void ThenMyShoppingCartShouldContainOneBook()
    {
        Cart.BookCount.Should().Equal(1);
    }
}

```

**Figure. 3.5:** Functional specifications

found at all that gave any proof of TA sessions improving performance. For example, “The Cucumber Book” only provides an explanation of the three roles and their involvement in the TA session, but does not actually explain what a TA session should look like [74].

Adzic mentions TA sessions in his famous book “Specification by Example”, but he does not give any explanation, other than it being a useful technique when the domain in which the software is being developed “requires frequent clarification” [3].

In another book, “50 quick ideas to improve your user stories”, Adzic and Evans give a small explanation of how a TA session could be organised [4]. They describe a session to typically start with a domain expert introducing the user story and explaining some initial scenarios that they think should be included in the user story. Following is the developer, who analyses the scenarios and possible functional gaps and inconsistencies in relation to the existing software. Lastly, the tester will look from his perspective if any scenarios should be added and how the user story should be tested. Even though this activity is called a discussion, the explanation makes it sound like a very procedural session where everyone follows one another, rather than working closely together to refine the user story.

Adzic and Evans also mention that teams should not stick to the number three if other roles can also bring valuable input to the sessions. This view is shared by more practitioners. Tooke, the co-author of the second edition of “The Cucumber Book” [75], has made the same remark in his post on Example Mapping (to be explained later), recommending TA sessions to have between three and five people [63]. Szabo mentions in his book “User Experience Mapping” that there should be a fourth amigo, namely the User Experience Expert, in order to represent

the needs of the users [62]. However, a “common pitfall” of TA sessions is that the whole team gets invited [5], which can turn a TA session into a costly session in which people are not engaged either [30]. Therefore, a good balance must be found between these two.

Looking into TA session techniques, two could be found that are well-defined: Example Mapping and Feature Mapping. Both techniques put a lot of value on gaining shared understanding amongst team members, which is why Section 3.4 will elaborate on shared understanding into detail. Whereas Example Mapping is more of a free-format kind of technique, Feature Mapping is more procedural. Both will be explained separately in the following subsections.

Both EM and FM define a session to take 25-30 minutes. This does not necessarily mean the user story is completely refined by the end of one session. In case the amigos do not find the story “ready” yet, additional TA sessions can be held to refine the user story further. This can be the case, for instance, when many unanswered questions came up during the session. Answers to these questions can be figured out after the session and can be used as input for the next TA session.

When relating TA sessions to the bigger picture of BDD, the techniques fit with the “Define features” activity, which is the second phase in BDD as illustrated in Figure 3.3. The techniques are originally explained to be used for user stories, which can be a part of the feature definition activity.

Looking at Figure 3.1, a TA session can fit in multiple RE process steps: It can fit in domain understanding and elicitation, evaluation and negotiation, and specification and documentation. As a session with people from various disciplines involved, it may often be all three steps during the same TA session. For a developer it may primarily be about domain understanding and elicitation from the domain expert. In contrast, for the domain expert the fact that everything is being specified and documented may be the most important factor. Discussions may arise while examples and rules are being specified, which refers to the step of evaluation and negotiation.

To illustrate EM and FM, we introduce the use case of PremRide, a train booking system. This was inspired by an online session given by Cucumber co-founder Matt Wynne [73] on EM. PremRide already has the ability to book tickets. A passenger can order multiple tickets at once in order to ensure an entire group can get aboard the train. However, this is still very limited to the core basic functionalities. Train conductors have noted that the trains during peak hours are always fully booked and that people who did not reserve a seat are often complaining that they have to wait a very long time before they can enter a train. On the other hand, it also sometimes happens that passengers cancel their booked ticket. It would be preferable to be able to notify people who were told that the train was booked that seats have become available. Therefore, the following two user stories are created:

**US1:**

As a train conductor

I want to make sure no more than 70% of all available seats in the train can be booked  
So that people without a prior reservation also have a chance to get on the train.

**US2:**

As a passenger

I want to sign up for tickets on a waiting list if a train is fully booked  
So that I can still get a ticket if previous bookings get cancelled.

### 3.3.1 Example Mapping

Example Mapping (EM) is a technique for organising TA sessions that was introduced by Matt Wynne in 2015 [72]. In EM, four different types of cards can be used: A story, a rule, an example and a question. First, a user story is picked to refine during the session. If a business representative such as a stakeholder, product owner or business analyst is present, he may introduce the story with some initial information on what it is about if this is not known by the amigos yet. After picking out the user story to refine and possibly giving some initial information, the EM session can start. An EM session itself is very free-format, as can be seen in its PDD in Figure 3.6. As choosing the story to refine or explaining initial details of the user story can already be done at an earlier stage, this is emitted from the PDD. Also, the relations between concepts are hidden in the PDD. This choice was made as concepts can all relate to each other in one way or another, which would make the PDD unclear.

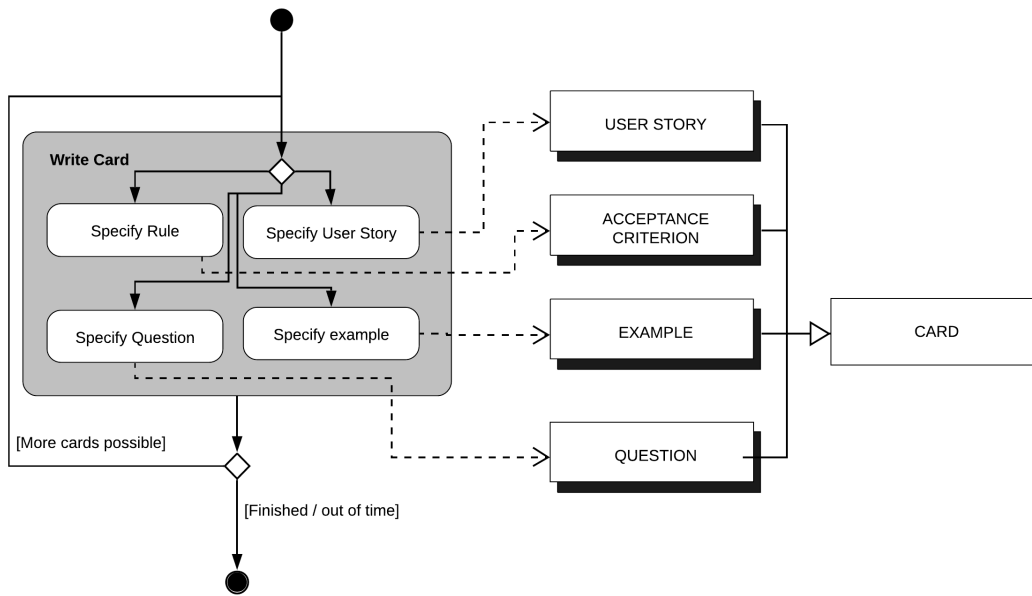


Figure. 3.6: PDD of Example Mapping

Aside from the user story that is already picked, which should be written down on a yellow CARD, three different CARD concepts exist: an ACCEPTANCE CRITERION, EXAMPLE and QUESTION, which respectively represent a rule, question or example. A different colour should be picked for each separate type of CARD, with the recommendation to use a green CARD for an example, a blue CARD for a rule and a red CARD for a question. As EM is very free-format, all the concepts are closed. This means the amigos do not have to stick to a specific format and have the possibility of defining the concepts however they prefer.

With EM, the amigos present in the session can pick their own order of what to write down first. If some rules are already known at the beginning of the session, these can be written down immediately. The same goes with examples: if the amigos are already aware of certain example cases, these can directly be written down at the beginning of the EM session. Rules can be illustrated using examples. In that case, an EXAMPLE is related to an ACCEPTANCE CRITERION. Rules don't necessarily need explaining, for instance if the rule is very straightforward, so an



ACCEPTANCE CRITERION does not necessarily have an EXAMPLE. An EXAMPLE can only be related to one ACCEPTANCE CRITERION, but it may also not immediately associated with a specific rule yet.

Questions may come up that cannot be answered by the participants during the EM session. In that case, the question should be written down on a QUESTION CARD. As a question can be regarding either a rule or an example, but doesn't necessarily need to be, a QUESTION can be linked to an ACCEPTANCE CRITERION, EXAMPLE, or neither.

Lastly, what may happen is that the amigos come up with a new user story during the session. This can, for instance, happen because of a question that comes up, or when too many rules are defined and the amigos decide to split the user story in two. In this case, they add an additional yellow USER STORY CARD.

### Example Case

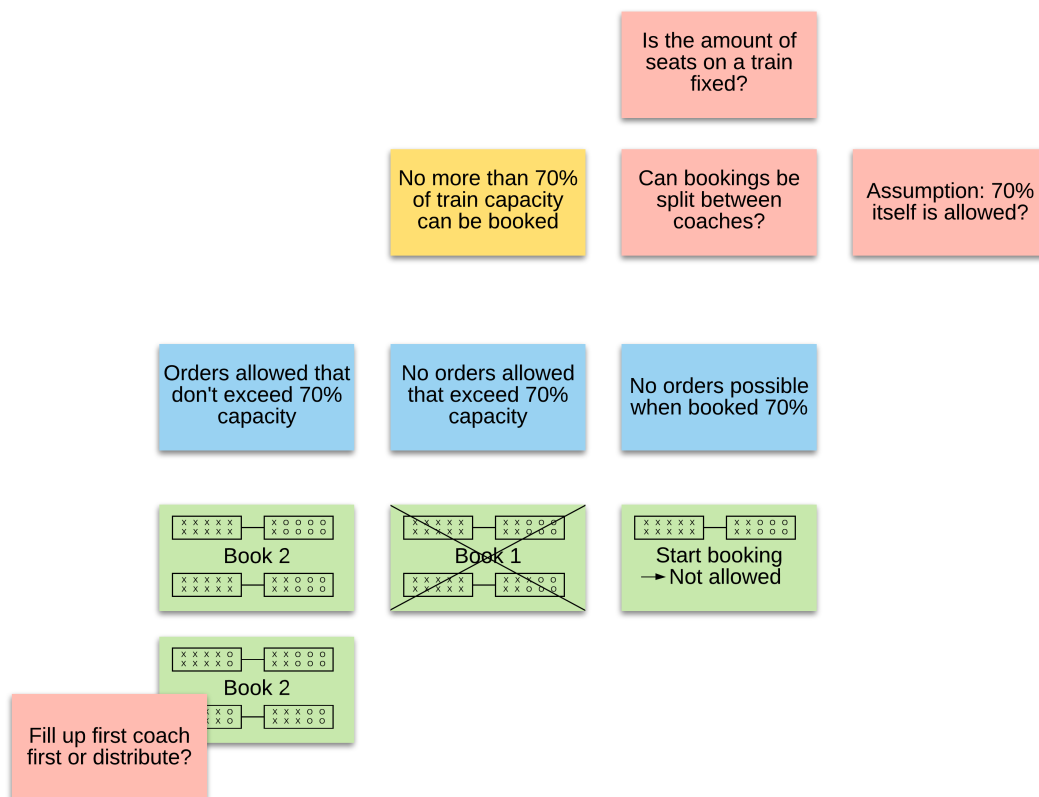


Figure. 3.7: Example Mapping – US1

In order to illustrate EM, we created outputs for both user stories that were mentioned. The outputs for US1 and US2 can be found in Figure 3.7 and Figure 3.8, respectively. Starting with US1, the output has 3 rules, 4 examples, and 4 questions. As can be seen with these examples, different formats are possible for defining them. In these examples, the number of available seats on the train is illustrated by a visual representation of two train coaches. Example Map-

ping does not restrict participants to write down purely text, for example, as is the case with Gherkin scenarios. This way, participants can be creative and specify the examples in a format that is most valuable to them. Three general questions are written down regarding the user story, which are located at the top right. One question refers to a specific example that is given for the first rule. Therefore, this question is placed with that example instead of at the top right with the rest.

The second user story can be found in Figure 3.8 and is added to illustrate the possibilities of creating new user stories during an EM session. This is not mentioned in the original introduction post on EM, but creator Matt Wynne has mentioned it in a later example video [73]. During this EM session, two questions were asked which both led to a new user story: one to introduce a warning system for users once a waiting list is considered “full”, and one that gives people the option to either distribute tickets between coaches or be put on a waiting list.

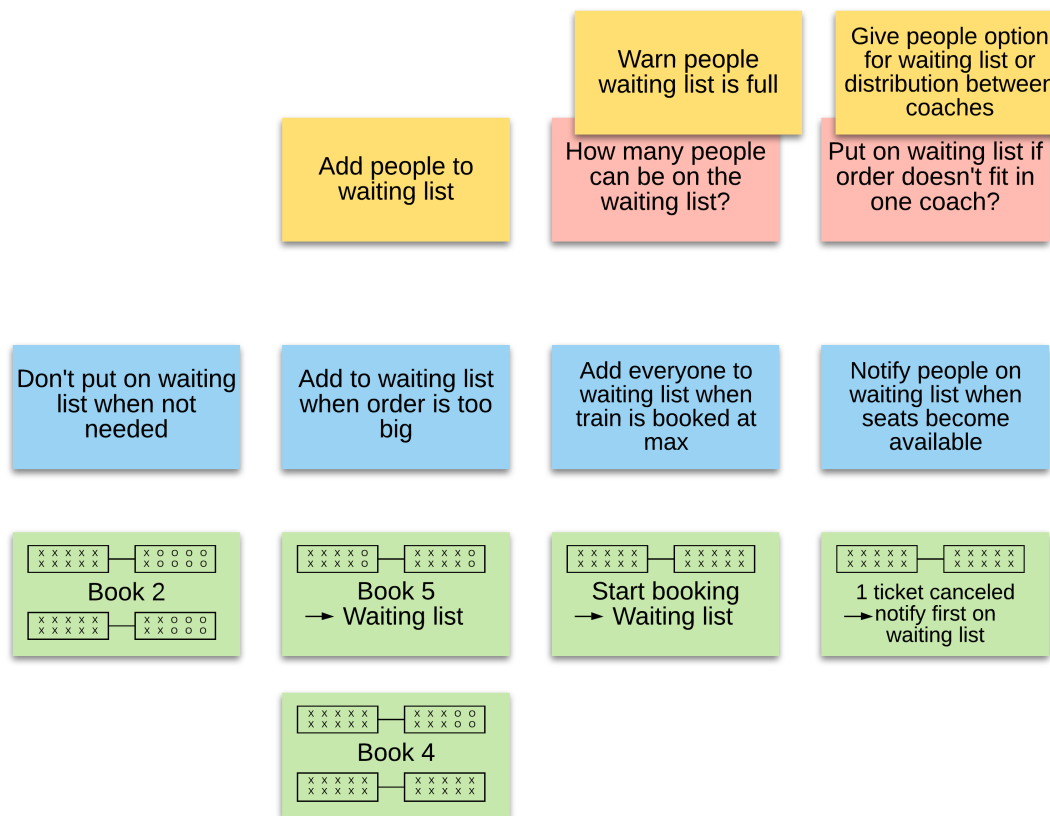


Figure 3.8: Example Mapping – US2

### 3.3.2 Feature Mapping

The second TA technique is Feature Mapping (FM), introduced by John Ferguson Smart in 2017 [58]. Unlike EM, which gives practitioners a lot of freedom on how to order the session, FM is more procedural and has a specific order that must be adhered to. The PDD of FM can be found in Figure 3.9. As with the PDD of EM, the relations between the different concepts are hidden for clarity.

FM has four main activities. The first activity is specifying which actors are involved in the feature or user story. With user stories, one important actor is often already mentioned. This is, however, not always the case, as can be seen in the example user stories US1 and US2 that are explained in Section 3.3. US2 already specifies “As a *passenger*”, which indicates the main actor of the user story. US1 is, however, written from the perspective of the train conductor. Although the train conductor may be a stakeholder in the user story, he does not actually have any role in the functionality that is to be designed.

In the example given in the introductory post, actors are named together with what their involvement is in the user story. The actors are also given names in the example, a recommendation that is also provided for Gherkin scenarios to make them more tangible [43]. It would also be possible to write a complete persona for each actor, which has proven useful for requirements elicitation [11], but this would take up much time and should not be part of the FM session itself. An ACTOR concept will be created for each actor that has a role in the user story’s functionality. This does not necessarily need to be written down on a CARD as it will be used in a later stage when specifying examples.

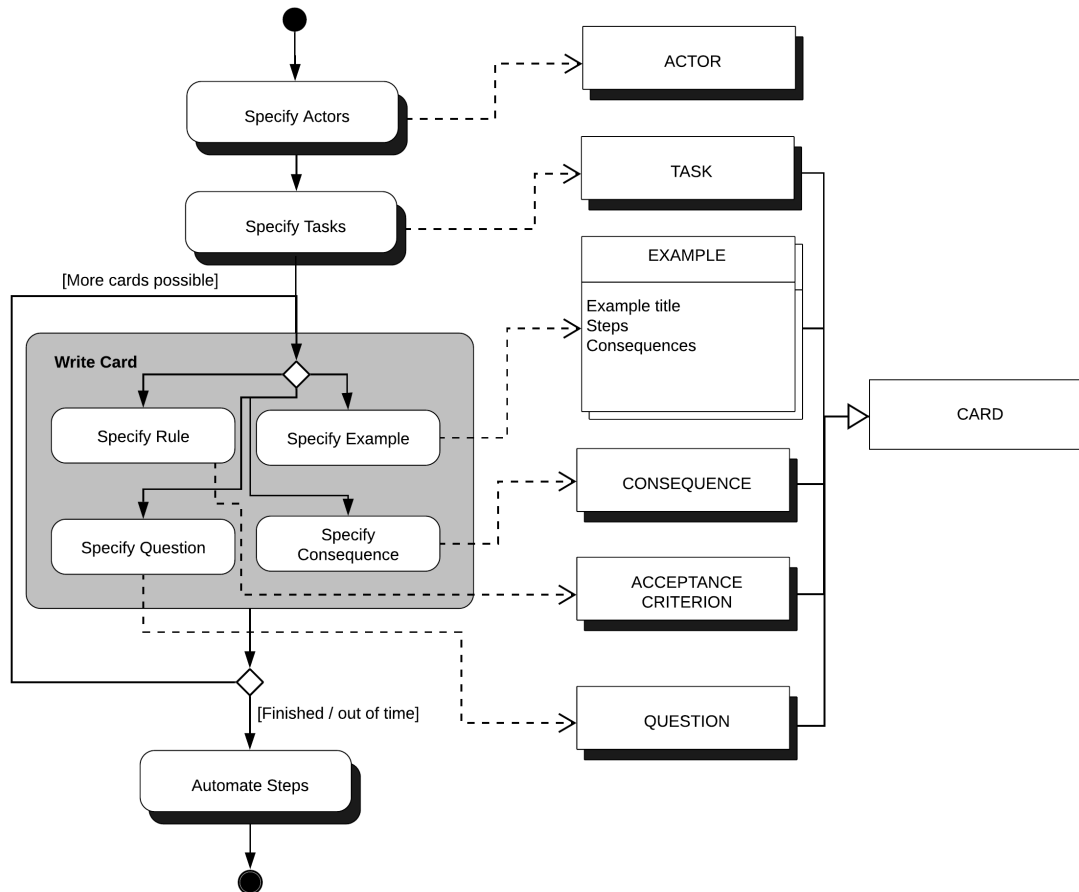


Figure. 3.9: PDD of Feature Mapping

The second activity in FM is defining the tasks. Each TASK is some sort of action that can be performed. This includes all tasks that are possible, not just the happy path (i.e., default

scenario) of the user story. Each TASK CARD can then be put next to each other horizontally. Thirdly, the examples must be specified. This concerns the general scenario titles that can happen, without any specifics of how they work per task.

In the third phase, complete examples are specified that show which tasks are used in which example and how they relate to one another. In this phase, rules and questions can also be added. This makes the fourth phase of FM similar to EM. This makes sense, as Smart explains FM draws partially on EM. The difference between EM and FM in this phase is that with FM, in contrast to EM, examples have a title and specific steps that are written on a separate CARD as well. Besides rules and examples, a CONSEQUENCE can also be added to an example. These consequences show what the explicit result of an example is. A specific column can also be added that gives a different implication for each example on a more general CONSEQUENCE. In the example that Smart gives, a column with consequences is added to show how a grade is calculated from average marks with each example [58].

Even though FM does not explicitly mention writing down an additional USER STORY CARD in case a new story is thought of, there is no reason why this would not be possible with FM. It is not added to the PDD because it is not mentioned in the technique itself, but considering this phase is based on EM, it should be noted that it is also possible.

The final step that FM explains is that of automating the Steps in executable specifications. This is also possible with EM, although it is not specifically mentioned as part of that technique. Each concrete example can be made into one Gherkin scenario that can then be implemented, as explained in Section 3.2.

### Example Case

In order to illustrate the output of refinement using FM, US1 was also refined using this technique. The output of this can be found in Figure 3.10. In this example, the actor specified in the first phase is the passenger, who is called Pete for this session. Secondly, tasks are defined. These are the five yellow cards that are underneath the user story card. The first task that was specified is the trigger for activating this functionality, namely a passenger that wants to order train tickets. This task is one that was not present at the output of EM. One consequence is also specified, namely that Pete is prompted when not enough tickets are left. This is a functionality that was not made visible in the EM refinement.

There are some small differences between the EM and FM output. Whereas EM had a total of 12 cards for US1, FM has 23. As FM is more procedural with extra steps, it makes sense that its output is also more extensive than that of EM.

The last step of FM would be to create executable specifications based on these examples. This is emitted from this example case, as it is a step that would typically be performed outside of the FM session. An example of how this would look like can be found in Section 3.2.3. There are three examples that are specified during the FM session. Therefore, there would be three test scenarios to be automated. In the case of EM, four examples were given for US1, which would translate to four test scenarios.

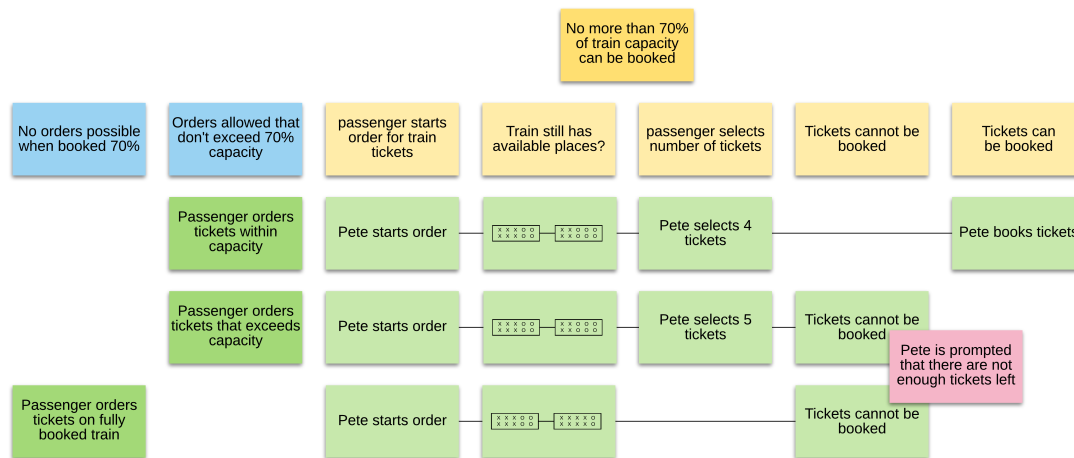


Figure 3.10: Feature Mapping – US1

### 3.4 Shared Understanding

Shared understanding (SU) is a term that has become popular in the field of Agile software development, and previous research has shown that it increases team performance and software quality [2, 52, 77]. Still, there are many variants of this term, and terms often have varying definitions. The term that is most prevailing amongst found literature is team cognition, which by itself has different interpretations as well [70]. As TA session techniques consider shared understanding of the requirements of a user story to be the most important benefit of conducting such sessions, it is important to define what exactly is shared understanding, how it is built and what its enablers and inhibitors are. Therefore, shared understanding is investigated in detail in this section.

According to Cannon-Bowers and Salas, many terms exist for shared understanding, and no clear consensus on definitions for the terms exist [13]. The authors themselves use the term shared cognition, other terms that are named are collective cognition, team knowledge, team mental models, shared knowledge, transactive memory, and shared mental models. They define four broad categories of knowledge that is shared: task-specific knowledge, task-related knowledge, knowledge of teammates and attitudes/beliefs. The first category refers to knowledge regarding the work that is done, e.g., in a software engineering context, this may refer to acceptance criteria of a user story. Task-related knowledge refers to knowledge about processes that are related to the task at hand, but is not necessarily tied to a single task. This includes knowledge about teamwork and dynamics between group members.

Looking at TA sessions, these two categories are mostly affected: implementing a different kind of TA session will alter the way a team works together in order to, hopefully, obtain better task-specific knowledge that is shared amongst everyone involved. Furthermore, Cannon-Bowers and Salas note that measuring shared understanding is a difficult task, but that two aspects should be measured: the content of the shared understanding as well as the way it is shared. However, they believe that there can not be a proper measurement of shared understanding, as long as there is no agreement on definitions and labels used to describe it. Being a publication from 2001, some issues that the authors state are already investigated.

Shared understanding can also be divided into four other categories according to Glinz and Fricker: true implicit shared understanding, true explicit shared understanding, false implicit shared understanding, and false explicit shared understanding [22]. The visualisation that Glinz and Fricker created of their model of SU can be found in Figure 3.11. Where Cannon-Bowers and Salas categorised shared understanding based on the type of knowledge, these categories regard the nature of how information is shared as well as the correctness of it.

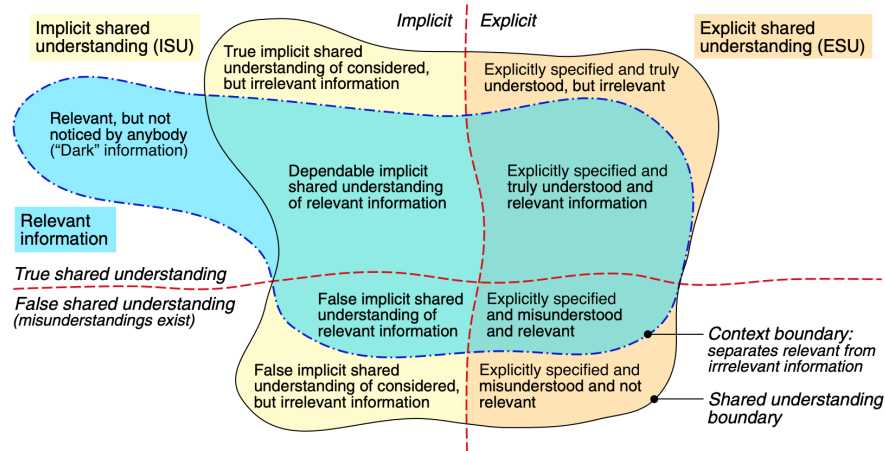


Figure 3.11: Categories of shared understanding [22]

Firstly, there is the difference between implicit and explicit shared understanding. As defined by Glinz and Fricker, “Implicit shared understanding (ISU) denotes the common understanding of non-specified knowledge, assumptions, opinions, and values.” On the other hand, explicit shared understanding (ESU) considers how group members understand explicit information. When looking at software engineering, this is for the most part covered by written specifications.

However, as the authors also note, verbal communication can also be a form of ESU, if remembered by all group members. This is a very volatile form of ESU, however, as it is unrealistic that all group members will remember the explicit information completely over an extended period of time. Looking at TA sessions such as EM and FM, this verbal form of communication may happen when one of the participants explains an example of an acceptance criterion. As examples and acceptance criteria need to be written down in this technique, the information will remain explicit and not fade. EM and FM both facilitate to make shared understanding explicit in an easily attainable way by writing it down on a post-it.

Besides explicit and implicit, Glinz and Fricker define shared understanding as either true or false. True shared understanding is when all parties have the same conception of what is meant by a requirement, whereas false shared understanding implies that all parties believe to have a shared understanding of the requirement but in fact do not. Both parties may have a different interpretation of how a requirement should be implemented, possibly resulting in faulty software. As EM and FM focus on writing examples that illustrate how an acceptance criterion works in practice, shared understanding is validated and the chances for false shared understanding could diminish.

Van den Bossche et al. explore creating shared mental models in teams [67], defining a shared mental model as “team members’ overlapping mental representation of key elements of the team’s task environment.” They consider a shared mental model to be built by three aspects: construction, co-construction and constructive conflict. Construction is when one team

member shares knowledge with others, thus building their mental model. Co-construction is when team members together build a new meaning collaboratively. When the transferred or newly created meaning is agreed upon by team members, a shared mental model is built between them. Constructive conflict occurs when there is no agreement between parties. By negotiation on the disagreed concept, new meanings are created that are part of the shared mental model.

Van den Bossche et al. also note that measurement of shared mental models is problematic. There are techniques that allow for some sort of measurements, but those all have their particular strengths and weaknesses. The authors of this paper have used a cognitive mapping technique themselves for measuring the mental model of participants in their experiment. The authors asked them two questions in order to verify their mental model and the answers were mapped to identify similarities and differences between the answers. Besides that, team effectiveness was measured. Results of their research suggest that constructive conflict is related significantly to building shared mental models. However, the results also indicate that co-construction only works for a shared mental model if parties have an active contribution, it is “not sufficient to simply pay attention and acknowledge a contribution; an active effort to integrate the contribution in the existing representation is needed.” When the shared mental model is higher, this was also related to better team effectiveness and performance.

This means that all parties must work together to increase a proper shared mental model, or shared understanding. Relating the above to TA sessions, team effectiveness and performance may increase due to the fact that the amigos must collaboratively come up with examples of how a user story should be implemented, which creates a bigger shared mental model amongst them.

Bittner and Leimeister give a set of guidelines on how shared understanding is best achieved, as displayed in Table 3.1 [9]. This is based on the aforementioned study by Van den Bossche et al. This contains aspects that are needed to maximise shared understanding, “items”, as well as activity guidelines to achieve them. Bittner and Leimeister define shared understanding as “an ability to coordinate behaviours towards common goals or objectives (“meaning in use” or action perspective) of multiple agents within a group (group level) based on mutual knowledge, beliefs and assumptions (content & structure) on the task, the group, the process or the tools and technologies used (scope/object perspective) which may change through the course of the group work process due to various influence factors and impacts group work processes and outcomes” [9].

According to the authors, mutual agreement is necessary in order to gain shared understanding on a particular perspective. Without mutual agreement, there may be a mutual understanding, but the view on “meaning in use” may differ, which is needed for shared understanding according to the definition given by Bittner and Leimeister. TA sessions facilitate this mutual agreement, as the examples are explicitly written down and shared with other participants. This corresponds with the definition that Glinz and Fricker give on false shared understanding [22].

Bittner and Leimeister also propose a *collaboration design process* for creating shared understanding. This process has a lot of steps and is tested in a 4-hour workshop. It includes each individual having to write down their own description of a task (i.e., requirement), followed by reading the descriptions of all other participants and discussing the similarities and differences between them. This is not how TA sessions such as EM and FM are specified: with these techniques, the focus is on cooperation rather than individual activities. With a smaller time frame of 25 minutes for one session, including these individual assignments may also take up too much time. Therefore, G1 and G2 are not followed in EM and FM. However, all other

Determinant	Item	Design Guideline
Construction	Team members are actively listening to each other	G1: Express individual understandings first
		G2: Encourage members to try to understand each individual perspective
	If something is unclear, we ask each other questions	G3: Ask questions for clarification
Co-construction	Information from team members is complemented with information from other team members	G4: Collect individual descriptions in one shared place
	Team members elaborate on each other's information and ideas	G5: Evaluate understanding and consistency with own perspective
	Team members draw conclusions from the ideas that are discussed in the team	G6: Proceed differences between understandings
Constructive conflict	In this team, I share all relevant information and ideas I have	G7: Encourage sharing of divergent views (parallel and anonymous)
	This team tends to handle differences of opinions by addressing them directly	G8: Address differences in discussion
	Comments on ideas are acted upon	G9: Process every conflicting aspect
	Opinions and ideas of team members are verified by asking other critical questions	G10: Allow clarification and questions and conflict negotiation

**Table 3.1:** Shared Understanding aspects and guidelines – derived from [9]



guidelines fit with the TA sessions due to the focus on working together to create a set of examples that clarify the requirements of a user story. When considering the item that G1 and G2 refer to, "Team members are listening carefully to each other", this is still a condition that can be satisfied by EM and FM. If a participant comes up with an example, he can explain this example to the other participants, thus sharing his individual understanding with them.

Cooke et al. have investigated shared cognition and team cognition in the field of cognitive sciences and believe that focus should be on processes and interactions at a team level instead of on an individual level [14]. They explain the difference between team cognition and shared cognition as the former being about a group as a whole, whereas shared cognition refers to individual cognition. They note several issues regarding the definition of shared cognition in relation to team cognition. Therefore, the authors have proposed the theory of Interactive Team Cognition (ITC) as an alternative theory to shared or team cognition. In their theory of ITC, they define team cognition as an activity rather than a property or product as it is often defined. Team cognition is "an emergent, dynamic activity that is not attributable to any one component of the team, nor the shared cognition of the team members, but to the team members as a whole as it interacts in the face of a changing, uncertain environment." This theory is acknowledged by other research, although the view of team cognition as a property or product is also still prevailing and used more in research than ITC [70, 51].

Team cognition should be measured and studied on a team level, rather than on an individual level, and is always tied to context. Where an assumption of 'traditional' team cognition theories is that the cognition of the team equals the sum of all individuals' shared cognition in that team, ITC does not recognise this assumption: team cognition can be both more or less than that of the sum of the individuals.

Another important implication that Cooke et al. give regarding ITC is that facilitating team member interactions for sharing information in a timely and adaptive manner is more effective than the distribution of content or presenting more information to more team members. This implies that activities such as TA sessions are beneficiary to team cognition.

Wildman et al. have performed a thorough literature review on team cognition across multiple disciplines [70]. They determine five research domains in which team cognition is most often researched: team mental models, transactive memory systems, situation awareness, strategic consensus, and team cognition as interaction. Team mental models are defined by the authors as the similarities of mental models of members of a team and the accuracy of those mental models. The definition of transactive memory systems (TMS) is two-fold: it regards both the knowledge that individuals in a group have, as well as the processes used to "encode, store, and retrieve that knowledge" [49]. The second part of this definition corresponds with the focus on interaction that the aforementioned ITC has.

Whereas research on team mental models and transactive memory systems consider team cognition to be a relatively stable concept, research on team situation awareness generally considers it to be a dynamic construct that changes quickly all the time. However, the concepts of situation awareness are very much overlapping with that of team mental models, according to Wildman et al.

Strategic consensus is most studied in literature on top management teams and is defined by the authors as "team's shared understanding regarding the high-level strategic goals of the team or organisation" [70]. Albeit a very different focus than the other research domains, shared understanding is generally considered as the degree of agreement or sharedness between individuals [32], making its concepts very similar to the other domains. Lastly, team cognition as interaction refers to team cognition as purely the dynamic interactions or processes that occur between team members. This research domain includes the theory on ITC and con-

siders team cognition as communication between team members itself, rather than considering the communication between team members as a process that builds team cognition, as is the case with for example transactive memory systems.

For the purpose of this research, where TA workshop techniques are investigated, both the knowledge of individuals and the interactions to convey information to one another are important. As the technique we investigate is itself an interactive activity, team cognition should include the interaction aspect that is described in theories such as ITC. However, it cannot focus on solely the interaction. Although creating a shared understanding may be the primary goal during a TA session, in the end, what is important is that the user story can be implemented as intended in order to create software of high quality. Therefore, team cognition cannot focus solely on the interaction but should also include the team members' individual knowledge on the subject. The definition of transactive memory systems most closely resembles this, as it also considers both the knowledge itself and processes around it. However, research on transactive memory systems often put focus on the dispersion of knowledge, rather than on knowledge that all team members possess, while TA sessions put emphasise information with team members in order to all get the same understanding. On the other hand, team members that are part of an Agile development team but were not present during the TA session may rely on the specialised knowledge of those that were.

In their research, Wildman et al. have come up with context-dependent recommendations on how to measure team cognition [70], as can be seen in Figure 3.12. In order to get to the appropriate technique to analyse team cognition, the first question that must be answered is if team cognition is conceptualised as the structure of knowledge, or as team interaction. After that, one or two more questions need to be answered, from which a recommended way of data collection is provided.

Lewis has created a list of questions used to measure transactive memory systems [35]. He distinguishes three different dimensions within transactive memory systems: knowledge specialisation, credibility, and coordination. Knowledge specialisation refers to the dispersion of knowledge. This is a general view amongst research on transactive memory systems due to a different interpretation of the term *shared* understanding. *Shared* can mean that the knowledge is known to everyone and is overlapping, or that it is divided amongst team members, as is the case with research on transactive memory systems. With credibility, it is evaluated if people trust the knowledge of others. Coordination refers to the process through which knowledge is shared. In this research, refinement techniques are investigated. Therefore, the coordination section of this research is especially valuable.

Many different views from different disciplines have been discussed in this section. Cannon-Bowers and Salas note that shared understanding must be measured both by its content and the way it is shared. Glinz and Fricker make the distinction between implicit and explicit shared understanding, as well as true and false shared understanding. In research on team cognition, shared understanding is considered either as knowledge structure that individuals possess or as the interaction processes that facilitate knowledge sharing. For the purpose of this research with a focus on TA sessions, shared understanding is defined as the following:

**Shared understanding:** the implicit and explicit knowledge that is shared amongst team members both as a structure and as a process. Besides that, at least two different types of teams exist: the development team as a whole and the team that performs the TA session.

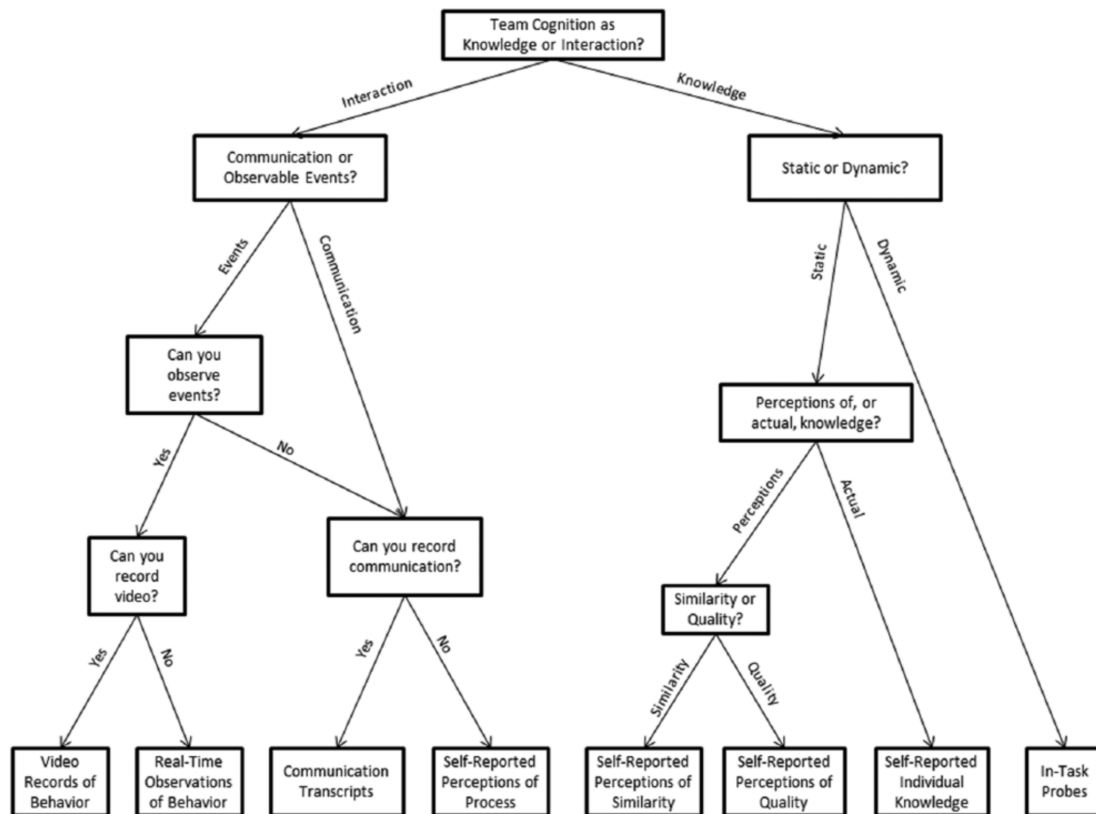


Figure. 3.12: Team cognition data source recommendations [70]



# Chapter 4 | Treatment Design

In this chapter, the execution phase of the research is designed. It concludes the second step of the case study method: “Design”. First, the setup of the case studies is discussed in Section 4.1. Following that section is Section 4.2 in which we discuss the controlled experiment. Lastly, the designed measurement tool of performance for the TA techniques is elaborated on in Section 4.3.

## 4.1 Case Study Design

In order to evaluate the performance of TA session techniques, case studies will be done with software development teams as an important base of data collection. An important choice to make with a case study is whether to make it a single-case or multiple-case design and whether or not the cases that are used are holistic or embedded [76]. A visual representation of this can be seen in Figure 4.1

Looking at holistic or embedded case studies, the difference is that a holistic case study looks at a broader picture and investigates a case as one unit of analysis, while embedded case studies investigate multiple. This research has an embedded case study design due to the fact that several characteristics are being investigated as will be explained in Section 4.3: perceived ease of use, perceived usefulness, intention to use and perceived shared understanding.

The next consideration is whether to have a single-case or a multiple-case design. A multiple-case study is often easier to generalise and can give additional insights that a single-case study does not, but there are also reasons to pick a single-case study design. Yin gives five possible reasons to choose a single-case study design of which one is important to this research, namely the reason to do a longitudinal study: “studying the same single-case at two or more different points in time” [76]. If certain conditions change over time, then a longitudinal study may be a good reason to keep the research a single-case study.

### 4.1.1 Longitudinal Case Study

With Example Mapping and Feature Mapping, there may be a learning curve for participants before they reach the full potential of the techniques for their team. Therefore, looking at teams that perform the technique more than once over an extended period of time is more interesting than a multiple-case study where all teams only use EM or FM only once. As such, a longitudinal case study is best suitable for this research so we can observe any possible changes in technique performance over time. Besides the learning curve, a longitudinal case study also allows us to examine the effects of TA session techniques on other aspects of the software development life cycle. Staying with a team for a longer period means we can also research if there are any effects on the implementation of a user story in software.

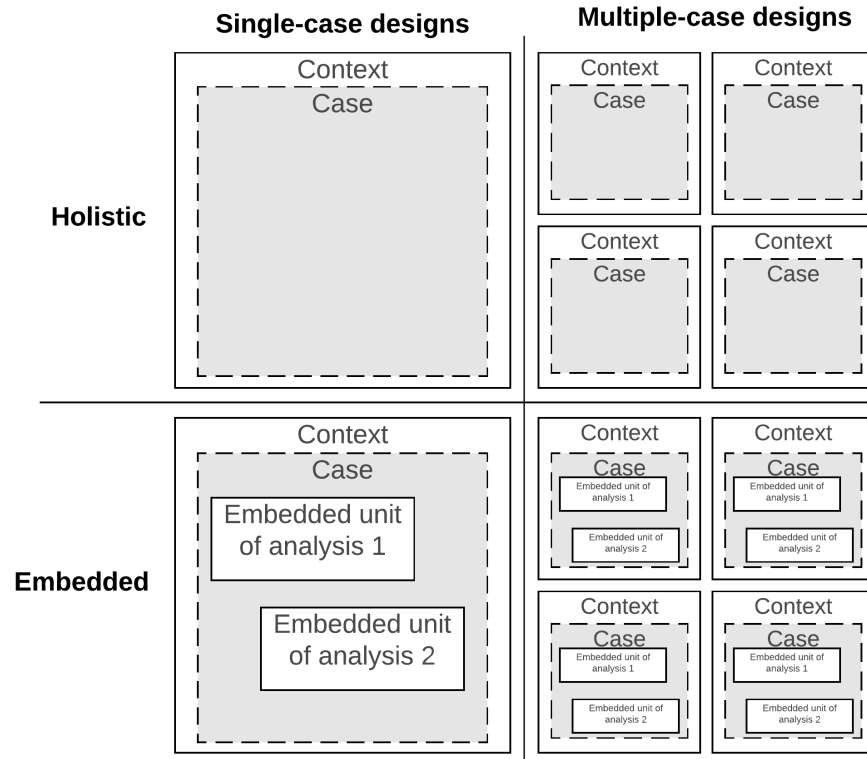


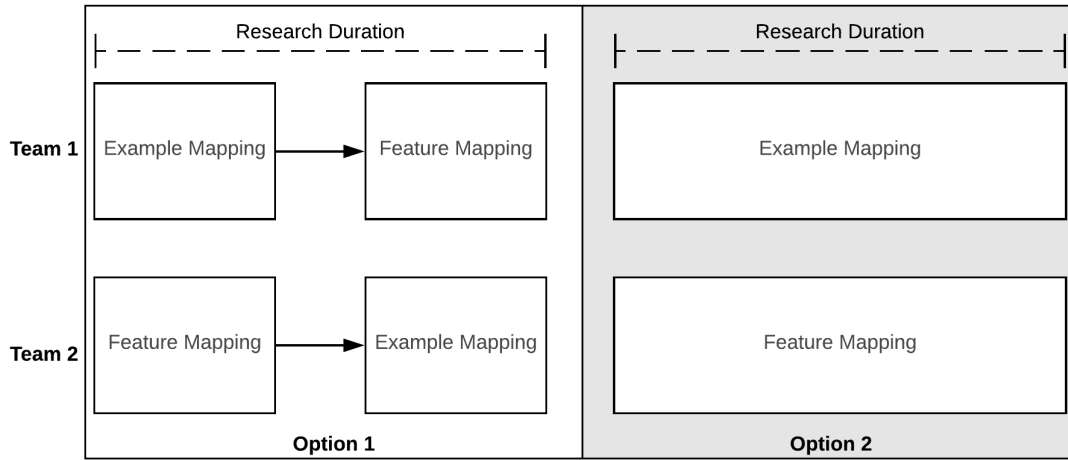
Figure. 4.1: Case study types, adapted from [76]

Nevertheless, ideally, this research will maintain a multiple-case design. If a team is willing to help conduct this research for a longer period of time, they can first do one technique for a while, followed by the remaining technique. Having one team use one technique first and the other technique afterwards may influence the performance of the second technique. Therefore, it would be favourable to have multiple cases (i.e., multiple teams) for the research. That way, techniques can be tested without a team having gained knowledge from another TA session technique first.

A challenge will, however, be to find teams to participate in the research for an extended period of time. In order to illustrate the possible ways of conducting the research, two teams will be taken as the basis for the number of cases. If two teams participate in the study, there are two options that we consider on how to proceed with the implementation, as visualised in Figure 4.2.

As the number of teams that participate may be small, the team characteristics themselves may also be a significant influence on the outcome of the research. The techniques may work well for some teams and work less than desirable for others, depending on their context. Therefore, Option 1 will be chosen. This way, a team will at least have had the chance to test out both techniques. If the team context rendered one technique ineffective, the other one might still prove valuable to them.

If four teams were to participate, a combination of Option 1 and Option 2 would even also be possible. Two teams would then follow the Option 1 setup and test both methods, whereas the other two teams would only test one respective technique for the entire duration of the case study.



**Figure. 4.2:** Case study design options with two teams

For the case studies, one team performing TA sessions will be considered to be one case. If the team has a long enough period to do both EM and FM, it will still be analysed as one case and not as two separate cases. This choice is made because having performed EM for a while may influence the performance of FM and vice versa.

### Time Duration

In order to determine the duration of each case study, several factors are important: time limitations due to the length of this research, company willingness and availability, and the frequency of sessions.

The first factor that influences the duration of the case study is the time limitation of this total thesis research. Eight months are specified for the total thesis. Of these eight months, the first three are used for setting everything up until the research design (i.e., chapters 1–4), and the remaining five months are meant for the execution phase of the research. Of these five months, we plan to have three months for the case studies. In the first month of this phase, we can get in contact with companies and their teams in order to request them to participate in our research. With three months worth of data, the fifth and final month of the execution phase can be spent on analysing all data and concluding the thesis.

The second factor is the willingness of teams to participate and their availability. TA session techniques help refine user stories and should, therefore, not cost much extra time in total to embed in the practices than when they are not used. In fact, if the techniques prove to perform well, they may overall even save time. However, we are still asking teams to change their way of working and to evaluate sessions, which will cost them time and energy.

Lastly, the frequency of the TA sessions may impact the case study duration. Depending on the context of the teams that will participate, they may want to do TA sessions either very frequently, rarely or somewhere in the middle. We want teams to adhere to their usual way of working as much as possible and to disrupt the practices that they are used to as little as possible. Besides that, the number of sessions a team desires may depend on the status of their Product Backlog. This is why we will let teams themselves decide how frequently to do the TA sessions.

We think that two sessions per week is a good assumption for the research design. In order to see some effect of a learning curve, we believe that teams should have at least six sessions using either EM or FM. Combining that with two sessions per week, we need at least three weeks to evaluate a technique longitudinally. This fits in perfectly with the defined three months, which would mean that a team can use one technique for approximately six weeks, followed by another six weeks with the other method.

### 4.1.2 Data Analysis

With multiple cases, the replication approach will be used for data analysis. This procedure for analysis of the obtained data was recommended by Yin and can be observed in Figure 4.3. In this procedure, data of different cases are not accumulated for statistical analysis, but rather analysed individually. This means that for each case study, an individual analysis will be performed on the obtained data. After the individual case reports, cross-case conclusions will be drawn. From those cross-case conclusions, we hope to create a generalised theory on the performance TA session techniques.

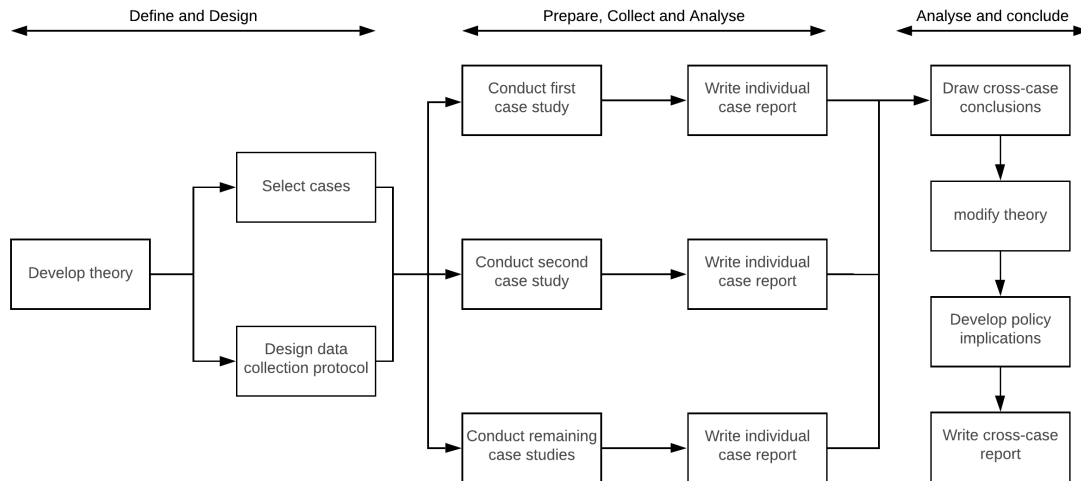


Figure 4.3: Analysis procedure, adapted from [76]

This thesis research also includes a controlled experiment, which will be discussed in Section 4.2. As the case study method is used for this entire thesis, the experiment will be considered as a separate “case” in terms of data analysis. As such, an individual case report will be created for the experiment, and at a later phase, a cross-case report is written in which the experiment is included.

### 4.1.3 Validity

In order to validate case study research on its quality, Yin recommends testing the quality of the case study design on four metrics [76]:

- **Construct validity:** identify if the correct things are measured for the concept that is studied in the research
- **Internal validity:** establishing a causal relationship between treatment and outcome



- **External validity:** generalising the findings of the case study
- **Reliability:** how well the research can be repeated

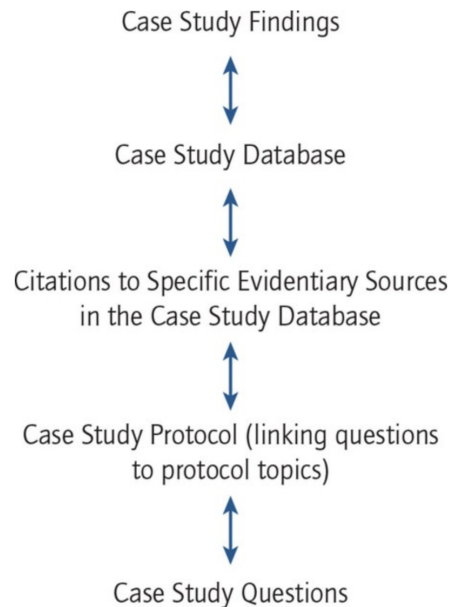
Yin gives several tactics for maintaining quality for individual metrics. Although most of these tactics concern later stages in the research (i.e., data collection and data analysis), Yin notes that it is smart to have already thought of these aspects before proceeding in order to make sure that the quality of the research is high. In the following subsections, these four quality metrics will be discussed individually.

### Construct Validity

For construct validity, three tactics are given. The first is to use multiple sources of evidence [76]. For the case studies, two different sources of evidence will be used: questionnaires and direct observation. Besides that, different questionnaires are used during different phases of the case study (see Section 4.3.1).

The second tactic that is recommended is to have key informants review a draft of the case study report. As this thesis has a supervisor from Utrecht University and one from host company Info Support, the case study report will be reviewed by at least two people that have a scientific and practical view, respectively. Those supervisors can be seen as key informants as their insights and opinions helped shape this research into what it has become. The fact that they have a different perspective will also help ensure that the thesis is of both academic and practical relevance. Participants of the research are offered to get a copy of the thesis, but as we do not want to ask them to spend more of them than necessary, they will not be asked to review a draft.

The third tactic that is given is to maintain a chain of evidence. With this tactic, Yin stresses that it is vital that all aspects of the case study research can be linked back and forth to one another [76], as visualised in Figure 4.4.



**Figure. 4.4:** Chain of evidence [76]

The chain of evidence is meant to “allow the reader of the case study to follow the derivation of any evidence from initial research questions to ultimate case study findings” [76]. To maintain this chain of evidence, several aspects will be incorporated into this research. Firstly, no evidence will be removed from the results. All answers to all questionnaires shall be included in the case study reports. If any results are omitted from the analysis (e.g., in the case of extreme outliers), it will be clearly explained why and those results shall still be included in the case study report.

Another important aspect for the chain of evidence is that the results of the earlier steps should reflect the concepts of newer stages of the case study. For example, the research questions should reflect the research method and findings. If this is not the case, then the findings will not actually give an answer to the research questions, and the wrong concepts have been investigated for this research.

By following a precise order in steps in this thesis (i.e., the chapters) and by often referring to other sections and chapters on how they relate to one another, this chain of evidence is maintained. Once case studies have been performed, then the resulting case study reports will use evidence from the results and refer to them when needed. This way, the later stages in this research also maintain the chain of evidence.

### Internal Validity

Four tactics are given concerning internal validity for high-quality research. The first tactic is pattern matching in which patterns are compared to one another. One way of pattern matching is by comparing patterns in results with the predicted patterns. In this case, the predicted patterns are depicted in Figure 4.5 regarding the different units of analysis, as will be discussed in Section 4.3.1. Once the case studies are done, these patterns will be examined.

The second tactic that is given is explanation building. As the name suggests, this tactic concerns the fact that case study data should be analysed by building explanations about the case. In this research, results and there shall be analysed and explained thoroughly. Scientific literature shall also be referred to whenever possible in order to interpret results.

Thirdly, using logic models can increase the internal validity of research. A logic model “stipulates and operationalises a complex chain of occurrences or events over an extended period of time, trying to show how a complex activity, such as implementing a program, takes place” [76] With logic models, the outcome of one event is the stimulus for the next. This tactic may be possible if case studies take long enough to also evaluate the implementation of a user story for which a TA session technique was used. Then, the outcome of the TA session will be the stimulus for the effectiveness of the implementation of software.

Addressing rival explanations is another valuable tactic for internal validity. Rival explanations concern any reason why the results of the research may have been positive other than the actual treatment itself. Yin notes nine different types of rival explanations [76], which will all be discussed in this section. Of these nine rival explanations, the distinction is made between craft rivals and real-world rivals, where the first regard scientific explanations and the latter regard practical explanations.

Looking at craft rivals, the first rival explanation is that of the null hypothesis, which says that the observed results are purely a coincidence. As no statistical analyses can be made cross-case, as explained in Section 4.1.2, it may be more difficult to reject the null hypothesis. However, as the plan is to conduct multiple longitudinal case studies, together with a controlled experiment, we believe that the overall results will be rigorous enough to give a clear statement about the techniques that are not merely a coincidental result.

The second craft rival explanation is that of threats to validity. Several validity threats are considered:

- **Maturation:** this is a threat when participants react differently because they are repeating the experiment multiple times. With the longitudinal case studies, the maturation (i.e., learning curve) is what we want to study, so this is not a threat to the validity of the research.
- **Selection:** the outcome may be affected by how participants are selected. This threat will be suffered due to the fact that selection of case studies will be based on the connections of the researchers and finding teams that are willing to participate, rather than having a large pool to (randomly) pick a sample from.
- **Mortality:** there is a risk that participants or teams decide to stop participating in the research during the research or even beforehand. As they are participating on a voluntary basis, there will be little to do about this except for trying to convince them to keep with the research.
- **Motivation:** if participants are unmotivated to participate, the results may be negatively influenced by this. We hope to mostly avoid this due to the fact that teams volunteer to participate and are therefore motivated to deliver good results. It may still be possible, however, that there are other factors that occupy the mind of participants (e.g., an important deadline that is coming up) which makes them unmotivated during a particular session.

A third craft rival explanation is that of the researcher bias. If a researcher is hoping to get specific results, he may subconsciously interpret the case study differently and steer the results in the desired direction. Observation will be a part of this research, but as most of the data that is obtained comes from participants themselves, this helps to avoid this threat.

The fourth rival explanation is a real-world rival, namely a direct rival. With a direct rival, it is another intervention that solely accounts for the results rather than the research treatment. This is similar to the fifth rival explanation, the commingled rival, which refers to another intervention contributing to the results together with the research treatment. As only one part of the way of working of development teams will be altered and evaluated (namely the refinement technique), this may occur. During the case study, any cues that may suggest other interventions will be asked about and written down in order to make sure that this is not the case. Adding a controlled experiment as a case also ensures that there is at least one case where these rival explanations do not happen for certain, considering that there is a research environment in which everything is controlled.

A possible rival explanation is the implementation rival which is when the process of implementation is what caused the results, rather than the implementation itself. This explanation may be possible with this research: the fact that a team changes their way of working by introducing a new type of refinement technique, they may become extra aware of refinements and get the benefits because of that, rather than because of the TA session techniques themselves. We hope to avoid this risk by firstly assessing multiple metrics. Secondly, because it is a longitudinal case study, the added awareness will possibly fade after several sessions, thus also diminishing the risk of this rival explanation.

A rival theory is another possible explanation for the research, which says that another theory explains the results. In case if the TA sessions, it may be possible, for example, that merely having the right people together in a session is what is effective and not the TA session itself. That would not be a big issue though: this may mean that the TA session itself is not the artefact increasing team performance, but it did facilitate it. In a way, the TA session would then force people to choose the right participants for a refinement session. Even though it may cloud the results of specific techniques such as EM or FM, it would still mean that TA sessions in general are effective. As such, only sub-questions of the research are at risk by this rival explanation (RQ3 and RQ4), but the MRQ still holds.

The last two rival explanations that Yin gives are super rivals and societal rivals. Super rivals happen when “a force larger than but including the intervention accounts for the results” [76]. Societal rival explanations are societal trends that caused the results, rather than the intervention. As these rival explanations are impossible to predict beforehand, no comments can be made on this regard. Any observed events or trends will be noted down and taken into account when analysing the results.

### External Validity

External validity, as mentioned earlier in this section, concerns the degree to which case studies findings can be generalised. Yin gives two tactics for ensuring external validity [76]. The first of these two tactics is to use theory to in single-case studies. Despite this being a multiple-case study, a lot of theoretical foundation has been laid in Chapter 3 in order to help generalise the findings.

The second tactic is to use replication logic for multiple-case studies. In the section on data analysis, section 4.1.2, it was already explained that this tactic will be applied to this research. Every case will be analysed individually, and only after that, attempts are made to draw cross-case conclusions. This way, external validity will be maintained in this research.

### Reliability

In order to make a research reproducible, the reliability metric is meant to “minimise the errors and biases in a study” [76]. Yin notes that repeating case studies with the exact same conditions rarely occur, but that reliability is still an important quality of case study research.

Two tactics are given for maintaining reliability, of which the first is the use of a case study protocol. A case study protocol contains sections. As this research attempts to implement a standardised method in different contexts (case studies), rather than having fundamentally different implementations with different case studies, no separate protocols are created for each case. Instead, all protocol sections are covered in different sections of this thesis.

The first section should have an overview of the case study with the objectives, issues and relevant readings. An overview of the context of a case study shall be given in Chapter 5, whereas the objectives of the case studies are equal to the research questions that were listed in Section 2.1. As it is possible that companies where the case studies are executed wish to remain anonymous, the extent of the case study overview will depend on the case itself.

The second section has data collection procedures, which includes items such as the sources of data and schedules of data collection activities. The sources of data are elaborated on in Section 4.3.1. The precise timing of data collection activities cannot be made yet as this will depend on the context of the case study. Once known, it will be elaborated on in Chapter 5. The different moments when data will be collected are already roughly known and are also detailed in Section 4.3.1. Another item that is part of this section is an informed consent. Every participant of the case study will sign an informed consent that is displayed in Appendix B.1

The third section contains the protocol questions. This concerns everything that must be answered in order to actually get results from the case study. Section 4.3.1 explains into detail how this shall be done for the case studies. The fourth and last section is an outline for the case study report. The outline for an individual case study will be as follows. First, the context of the case will be described. This includes contextual details on the company and the participating team (if possible), the time frame of the research and any other details that are of importance. Next, an overview of the results will be presented, followed by a detailed analysis on the results.

Creating a case study database is the second tactic that Yin gives in order to get a case study research with good reliability. Results of case studies shall be time-stamped and grouped by case study and also by refinement session. This way, a clear case study database is created that can be referenced to in analyses and that can be traced back to as well in order to maintain the chain of evidence.

## 4.2 Controlled Experiment

Besides the case studies, an experiment will be performed in order to evaluate the TA session techniques in a controlled environment. This way, the techniques are isolated and a good analytical comparison between the two can be made.

For the controlled experiment, the Requirements Engineering (RE) course of Utrecht University is chosen. This course is available to Master students. This year it only has students from the Master in Business Informatics from Utrecht University. This course fits well with the experiment considering that refinement techniques are part of requirements engineering (RE).

### 4.2.1 Context

For this controlled experiment, a train booking system called PremRide shall be designed in the TA sessions. PremRide was explained earlier in Section 3.3 for to illustrate how EM and FM outputs look like. This context was chosen because the experiment will be conducted with students and the assumption is that (almost) all students have experience with travelling by train. This means they will already be familiar with concepts such as having multiple coaches on one train.

However, a difference in the context of this experiment is that train seats do not need to be reserved in the Netherlands with the most used train service provided by Nederlandse Spoorwegen (NS), which is the case in this system. This has a risk of students misunderstanding the assignment. On the other hand, this also opens up the possibilities of students coming up with questions. For example, in Dutch trains, people can stand in the cabins where the doors are and even in the coach itself when the seats are full. In the to-be-designed system, if the maximum amount of seats is booked, perhaps students will come up with the question if standing tickets can be booked.

The same two user stories that were illustrated in Section 3.3 are considered for this experiment, alongside two others:

**US1:**

As a train conductor

I want to make sure no more than 70% of all available seats in the train can be booked  
So that people without a prior reservation also have a chance to get on the train.

**US2:**

As a passenger

I want to sign up for tickets on a waiting list if a train is fully booked  
So that I can still get a ticket if previous bookings get cancelled.

**US3:**

As a train conductor

I want to make sure no more than 70% of all available seats in a coach can be booked  
So that people without a prior reservation also have a chance to get on the train.

**US4:**

As a passenger

I want to cancel my booked tickets

So that I do not have to pay for tickets I cannot use.

From these four user stories, two are selected for the experiment. Outputs are created for all user stories in order to compare possible outcomes and select the best two. All outputs can be found in Appendix C.1 and Appendix C.2 for the outputs of EM and FM, respectively. The two user stories that are chosen are US2 and US3. These user stories are selected because there is some extra complexity in them that can arguably be solved in several ways. For example, with US3, participants can choose to split a booking up between several coaches if it is too big for one or to not allow these bookings at all. By describing to participants that cancelling tickets (US4) is already possible, US2 also has several implementation options. Taking people off the waiting list can be handled by booking the tickets or by sending passengers a notification that spots have become available, or participants may even choose not to include that in this user story at all. Having these different possibilities of how the user stories are refined may give interesting results and, therefore, these two user stories are selected for the experiment.

## 4.2.2 Planning and Presentation

The RE course has two available time slots, of which the one on Mondays will be chosen for this research. This is a four-hour time slot from 13:00 to 17:00, which gives us plenty of time to execute the experiment. The date that was chosen for the research is February 24th, 2020. This was the fourth week of the RE course. Students had already been explained what user stories are and have a general understanding of refining requirements. The planning of the controlled experiment has the following items and time windows for each item:

1. Lecture: 45 minutes
2. Briefing: 15 minutes
3. Experiment execution: 90 minutes
4. Debriefing: 30 minutes

First, a lecture was given to the students. The lecture slides that are used for this can be found in Appendix C.3. During this lecture, students are first introduced with Behaviour-Driven Development (BDD). As this is where TA sessions originate from, it may be useful for students to know the workings of BDD. Also, by introducing students to BDD, they are also introduced to automated tests that work with Gherkin. By showing them the existence of tests that automatically verify functionality, they may also realise a potential benefit of TA sessions is that the produced examples can be used for these automated tests.

After BDD is explained to students, TA sessions are introduced. After a few general slides, EM and FM are individually explained to the students. In order for them to fully understand the techniques, an example is used that shows how they work step by step. Following this is the briefing on the experiment itself. With this briefing, students are first introduced to the case they will be working on, which was explained in detail in Section 4.2.1.

After the case introduction, students are explained what materials are provided to them and what the schedule of the experiment are, which are explained in Section 4.2.3 and Section 4.2.4, respectively. After the execution, the debriefing will take place, which is also explained in Section 4.2.4

### 4.2.3 Student Handout Package

Students are given a handout package that is displayed in Appendix C.4. Each of these items will be elaborated on in the following sub-sections.

#### Instructions

Students receive a page with instructions on how the experiment will take place. On this page, it is instructed which sessions they need to do in which order (to be explained in Section 4.2.4), along with the room they are allocated in. There is also a copy of the schedule that was also in the presentation slides so that they have the schedule at hand. The emoticon of the clock is added to the schedules as students are requested to set a 30-minute timer for each session.

Students are also instructed how they can get materials and what to do when they have questions about the experiment. Next, students are informed on how they need to share their session output with us. They are asked to take a photo of this and paste that in an online Google Presentations document. Lastly, they are asked to return to the lecture room once they are finished with the experiment.

#### Case Description

Second in the student package is the case description. In this description, some information is given on the case and on the user stories that was partially explained during the briefing as well. It provides some initial rules that must be adhered to in the user story. By having a small introduction of each user story, a real-world scenario is imitated a bit more as a domain expert would usually do this during a TA session as well. As students are not aware of the two additional user stories that were considered for this experiment, the user stories they get are labelled US1 and US2.

#### Informed Consent

Next is the informed consent that students need to sign. The informed consent offers protection to both the students and the researchers. Students confirm that they are participating on a voluntary basis which helps for the validity of the research, and they are informed that all results will be anonymous, which helps for the privacy of the students.

#### Questionnaires

The next item in the handout package is the questionnaire. Students will perform two sessions and thus get two copies of the questionnaire. The questionnaire is used to evaluate the performance of the techniques, which will be explained in Section 4.3.1. Besides the questionnaire, students are asked which group they belong to, which user story they refined and what technique they used.

### Technique Overviews

Students are given a small overview of both EM and FM. By having a small example of how the techniques work and a definition of the concepts of the techniques, students are hopefully able to execute the experiment without wasting time on having to look up the full lecture slides.

### Demographics Questionnaire

A small demographics questionnaire is included in the handout package in order to analyse the group that participates in the experiment. The assumption is that there are not enough students to make distinctive conclusions based on the demographics, but it is good to include it anyway as it will take the students merely a few minutes to fill this in. Students are asked for their age, previously obtained degrees and their degree of experience with five concepts: user story refinement, working in an Agile software development environment, Gherkin, Example Mapping, and Feature Mapping.

### Writing materials

Besides the paper handouts, students are also provided with post-its and markers. They are given rectangular post-its of all necessary colours for EM and FM, as can be seen in Figure C.9. Note that the CONSEQUENCE card is made orange instead of purple. Purple was the colour that was used with the introduction of FM, but we concluded that purple post-its were too similar to pink QUESTION cards. Therefore, orange was chosen as an alternative. In the end, what is most important is that there is a clear distinction between the different CARD concepts, which is now the case.

On the first post-it of each colour, the type of CARD concept is written. This way, students have a legend that they are advised to put next to their session map, in order to make sure that it is clear which card refers to what concept. This was especially important due to the fact that the CONSEQUENCE CARD has a different colour from the theory presented in the presentation.

## 4.2.4 Execution of Experiment

In line with previous lectures of the RE course, we expect to have about 22-24 students present during the experiment. Students will be split up in groups of three, meaning there would be about eight groups. If the amount of students present is not a multiple of three (i.e., not everyone can be in a group of three), a fourth person will be added to the first and possible second group. As the TA techniques can also be performed with more than three participants, we believe it to be better to have one or two groups of four students rather than one group with two students.

There are two user stories to be refined and there are two techniques. In order to make sure that there is a minimal impact of a learning curve or of the context of the user stories, groups are executing the refinements in alternation orders, as can be seen in Table 4.1.



	First user story	First technique	Second user story	Second technique
<b>Group 1 &amp; 2</b>	US1	Example Mapping	US2	Feature Mapping
<b>Group 3 &amp; 4</b>	US1	Feature Mapping	US2	Example Mapping
<b>Group 5 &amp; 6</b>	US2	Example Mapping	US2	Feature Mapping
<b>Group 7 &amp; 8</b>	US2	Feature Mapping	US1	Example Mapping

**Table 4.1:** Experiment Execution Order

Students will be divided over six rooms. There is the main lecture room which is a large room, together with five smaller rooms. The lecture room can hold three groups, while the other rooms can hold two groups. In order to make sure students do not get distracted and do not use each other's ideas, the groups will be divided in such a way that this is as hard as possible. The division of groups is given in Table 4.2.

Room	Groups
Room 1	Group 1 & Group 5 & Group 8
Room 2	Group 2
Room 3	Group 3
Room 4	Group 4
Room 5	Group 6
Room 6	Group 7

**Table 4.2:** Experiment Room Division

After the execution of the research, students are asked to return to the lecture room for a debriefing. In this debriefing, students are asked what they thought of the techniques and if they want to present their results. From these presentations, we hope to trigger discussions on other possible ways of executing the techniques or other ways of implementing the user story. This may give us an overview of how the students interpreted the techniques and at the same time gives students the ability to get more insights about the techniques from their peers. Students will also be asked if they preferred EM or FM and what the reason is for their preference.

#### 4.2.5 Validity

As the experiment is considered an individual "case" as part of a multiple-case study, Section 4.1.3 holds tactics that can for a part also be used on the controlled experiment. Besides those, Table 4.3 shows a comprehensive overview of experimental validity threats and how they are mitigated if so. This table is adapted from the work of Panach et al. [47].

**Table 4.3:** Threats to validity of the experiment, adapted from Panach et al. [47]

Type of threat	Status	Threat	Due to	How it is handled/mitigated
Conclusion validity	Avoided	Random heterogeneity of subjects	The background of the subjects can be too heterogeneous due to random selection	Subjects all have a scientific background as they follow an academic Master's course
		Fishing for results	The observers look for a desired result during the research	All available data is analysed and used, and no specific outcome is desired for the observers
		Random irrelevancies in the experimental setting	External elements from the experimental setting can influence the results	Rooms were reserved for students to work without interference
		Reliability of measures	The experiment's validity depends on the reliability of the measures	The treatments were reviewed multiple times for mistakes and students have time to ask questions if anything is misunderstood
	Suffered	Low statistical power	Sample size is too small	Significant results are possible with this sample size, but its power is lower than desired since it is a relatively small sample size
		Reliability of treatment implementation	Deviations from standard procedures	We suffer this threat as we cannot guarantee students do not spend time on other activities during the experiment, such as social media
Construct validity	Avoided	Mono-method bias	Measurement bias can occur due to experiments containing a single type of measure	There are multiple types of measures (questionnaire, output quality)
		Evaluation apprehension	Subjects are anxious about taking part of a research	Students are told that there is no problem if they are not finished after 30 minutes and the results were anonymous.
		Mono-operation bias	A single type of object can lead to measurement bias	This is avoided by using two different cases which are also alternated over the two techniques
	Suffered	Interaction of treatment and testing	Subjects were aware that they were participating in a research	Students could have tried to be more concise for better results than they normally would
		Interaction of different treatments	It is possible that treatments of various studies may interact with each other	This is partially suffered because half of the students used EM first and FM next and for the other half this was the other way around. However, they could have applied the reasoning from the first treatment on the second

Type of threat	Status	Threat	Due to	How it is handled/mitigated
Internal validity	Avoided	Mortality	Subjects leave before completion of the experiment	All students will come back to the lecture room after they are done with the experiment
		Maturation	The subjects react differently to treatments as time passes	Students have to perform both TA session techniques during one time slot
		Resentful demoralisation	Subjects receiving treatments that are not desirable may perform worse	This threat is not present with this experiment as students perform both technique treatments
		Compensatory rivalry	Subjects obtaining less desirable treatments can be motivated to alter the results of the experiment	As students are asked about their perceptions on the techniques, they can actually tell the truth about which technique they desire
	Suffered	Subject motivation	Subjects that are less motivated can potentially achieve worse results than highly motivated subjects	Students have no incentive to perform well, therefore we suffer this threat.
		Selection	The selection of subjects can affect the results of the research	Subjects are selected based on the fact that they follow a specific academic course, so this threat is unavoidable
External validity	Avoided	Interaction of setting and treatment	The objects used in the experimental setting makes it not representative for the real world	This is avoided as the setting resembles that of a real-world TA session: students are in groups of three or four and perform the session on post-its
	Suffered	Interaction of history and treatment	The experiment is conveyed on a particular day or time which potentially affect the results	The experiments took place on a Monday afternoon
		Interaction of selection and treatment	The subject population is not representative to generalise	This threat is partially suffered because the subjects are still students and not professionals/consultants, even though they are considered experts

### 4.3 Measuring technique performance

A way of measuring actual knowledge in a team is through relatedness ratings [70], which is a type of self-reported individual knowledge as was earlier illustrated in Figure 3.12. Relatedness ratings are the most common way of testing individual knowledge, where domain concepts are compared to one another by participants in terms of relatedness, as described by Gorman and Cooke [23]. A technique like this will not work well for the controlled experiment, however, as the two user stories themselves are both related to each other and share the same domain.

Besides that, with user stories being only small increments of a larger product, there are not many concepts that are unique to a single user story. This makes it very difficult to test the knowledge on a specific user story based on the TA session about that user story. Relatedness ratings can be asked regarding an entire domain but seems unsuitable for specific user stories. This problem is also present with the other ways of testing knowledge that are described by Wildman et al. Besides it being a difficult way of testing knowledge, it will also be very time-consuming for participants to do this knowledge test after every session during a longitudinal case study. Therefore, relatedness ratings are unsuitable for testing knowledge of participants in case studies as well. As such, the focus during this research will be on perceptions of knowledge, rather than actual knowledge.

In order to measure the user perception of techniques, the Method Evaluation Model can be used for evaluation [41]. With the Method Evaluation Model (MEM), the perceived ease of use, perceived usefulness, and intention to use are validated for a method or technique, based on a set of questionnaire questions. Shared understanding is an essential aspect of both EM and FM. Therefore, this aspect should also be incorporated in the measurement of the performance of the refinement techniques. A link between SU and effectiveness has been observed in various studies [12, 33, 40]. However, previous research seems inconclusive with regards to how SU influences efficiency [59, 2]. As such, the method evaluation model is expanded to incorporate this. The proposed new model can be observed in Figure 4.5. In the adapted model, SU is only tied to effectiveness and not to efficiency. Perceived SU is also added, as this is what shall be tested, which is linked to perceived usefulness of the technique.

#### 4.3.1 Questionnaires

As mentioned earlier in this section, the Method Evaluation Model (MEM) uses a questionnaire to validate a technique. With the adopted model that includes SU, which is represented in Figure 4.5, an extended questionnaire is designed for this research. The extended questionnaire can be found in Appendix A.1. The questions on perceived ease of use, perceived usefulness and intention to use are adapted from the Method Evaluation Model [41]. SU is built up as a combination of coordination and shared knowledge in the questionnaire. The coordination questions evaluate the interaction between team members and are adapted from the work of Lewis [35]. The shared knowledge questions evaluate if team members perceive they share the same knowledge by the end of the TA session and are adapted from Lim and Klein [36], Schmidt et al. [52], and Gevers et al. [21]. These 11 questions together measure the perceived SU of a participant.

As is the case with the original MEM, the questions are put in randomised order. Moody took this measure in order to make sure that there is not a possible ceiling effect in which monotonous responses are given to questions regarding the same concept [41]. This was based on the earlier work of Hu et al. [28].

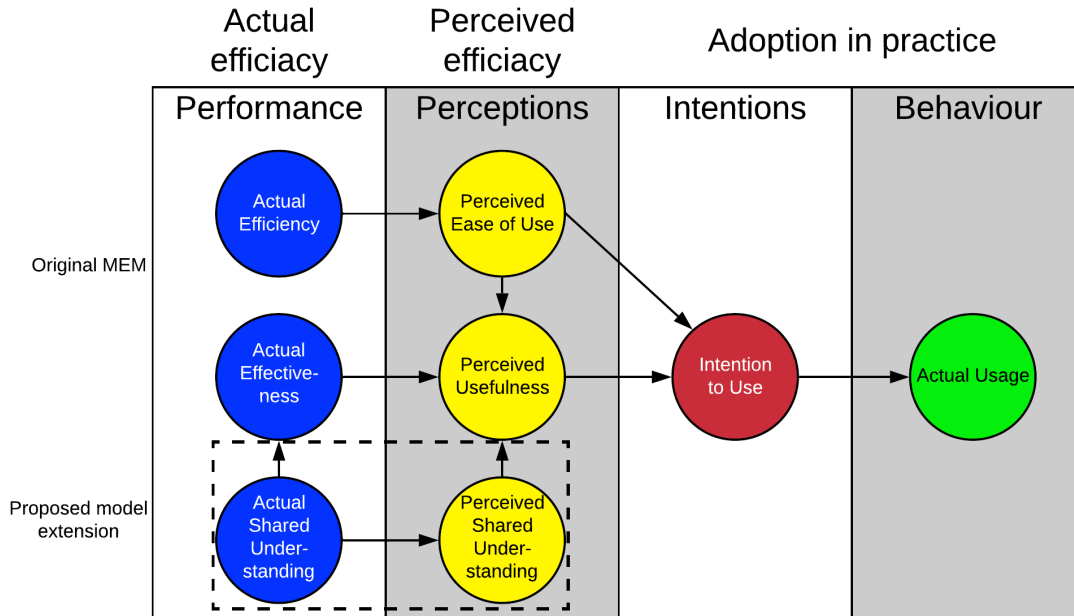


Figure. 4.5: Adaption of the Method Evaluation Model

### Session Questionnaire

This extended survey can be used for both the controlled experiment and the case studies. However, for the case studies, it is undesirable to ask participants to fill in the complete questionnaire after every refinement session. Some participants may be available at all sessions that occur once or twice per week. They may get annoyed with filling in the complete questionnaire, which could influence their willingness to participate and also the quality of their answers. Therefore, a shortened questionnaire is also designed, which can be found in Appendix A.2. The complete questionnaire will be given to them several times during the case study, at minimum the first and last session of one TA technique. The other sessions, the shortened questionnaire is used.

The shortened questionnaire focuses not on a participant's perception of the general TA technique as is the case with the complete questionnaire, but rather on the particular session they just had. This way, the perception of both particular sessions and of the technique in general is evaluated.

This session questionnaire is an adaption of the complete questionnaire and has four out of five of the same categories. Only the intention to use is left out, as this is more about the technique as it is about a particular session. No reversed questions are asked either, due to the fact that it is only a small questionnaire. Besides that, if participants fill in the session questionnaire several times, they will get used to the reverse questions as well. Having reverse questions may also annoy those frequent participants, which could negatively impact the results.

From perceived ease of use, Q4 and Q20 are picked and rewritten. These were best applicable to single sessions rather than a general technique. With perceived usefulness, Q9 and Q23 are selected. These are chosen as they respectively show verification on correctness and communication to stakeholders that are not present during the session. Together, these resemble a good overview of perceived usefulness. For coordination, Q12 and Q19 were chosen as they

show two different aspects of coordination, how the session itself went and if everything was understood well by everyone. Lastly, for shared knowledge, Q3 and Q21 are selected. Q21 is picked as a question on the general agreement, whereas Q3 asks about the understanding of one item, namely the examples. The choice was made explicitly to pick Q3 from Q2, Q3 and Q11, which respectively question the understanding of rules, examples and questions. This choice was made because the representation of the examples is what is most different between EM and FM. Therefore, this is the most interesting aspect to ask participants about in the session questionnaire.

### Session Observation

For the case studies, one of the researchers will be present during as many sessions as possible. This allows us to also observe how a session is going from an outside view. This additional observation could give additional insights that the questionnaires filled in by the participants may not. Therefore, we have come up with the following items that can be observed:

- All the questions of SU coordination
- Perceived involvement of participants
- When it regards an EM session: order of writing cards
- Any other notes (observations) regarding the session

The SU coordination questions can be rated by an observer that is present during the session, whereas all the other categories of the questionnaire cannot. That is why the observer shall also answer this set of questions. Besides that, a question on “perceived involvement” is added. This question was added as the work of Van den Bossche et al. shows that active contribution is necessary to gain a proper SU [67], as mentioned in Section 3.4. Finally, it may be interesting to see in EM sessions if there is a connection between the order of creating cards and the performance of a technique. For example, participants could try and first write down all the rules, write down a lot of examples first, or each time write one rule followed by examples. As the techniques do not prescribe a specific order for this part, it may be interesting to see if patterns can be found in this.

Besides these three observational items, any additional observations will also be noted down. For example, if an outside party disrupts a TA session, this can be noted down as it may have an effect on the performance of said session.

### Questionnaire after implementing a user story

As the case study is planned to take place over a longer period, it would be valuable to know if the outcome of a TA session has had an influence on the realisation of a user story. This is what the questionnaire that can be found in Appendix A.3 is for. Developers that worked extensively on the user story will be asked to fill in this questionnaire. A developer that works on the user story was not necessarily present during the TA session. Therefore, it might be possible to compare two groups with one another, depending on the number of user stories that get realised during the course of the experiment, after the experiment is done. The first group consists of developers that worked on the user story and were present during the TA session of that user story, and the second group consists of developers that were not present during the TA session of the respective user story.

Two categories are depicted: usefulness and shared knowledge. Ease of use, coordination and intention to use refer more specifically to the technique itself and are, therefore, omitted. With usefulness, Q7 and Q9 from the complete questionnaire are adapted. Besides that, one

question is added to evaluate the correctness of the session's output and another one for the completeness.

For shared knowledge, two questions are asked. The first question evaluates if the session output created a shared understanding that helped with the implementation of the user story. The second question is asked to validate the agreement of the session output. The first question looks at examples, whereas the second one is about acceptance criteria. This way, both aspects are evaluated without having too many questions. Nothing is asked about question cards that may have been written during the TA session, as those ought to be answered for the most part before the user story gets implemented anyway. Therefore, the hypothesis is that not many open questions remain at this phase, making a questionnaire question about it obsolete. We also do not want to annoy participants filling in the questionnaire by asking too many questions, causing a big obstruction in their daily work, which is why it was kept to these two questions for shared knowledge.

In order to summarise the division of questions between the different questionnaires, Table 4.4 is created. In this table, the two added aspects of a session observation are included in the "SU - Coordination" column as they regard coordination aspects of a session.

Amount of questions per concept	Perceived Ease of Use	Perceived Usefulness	Intention to Use	SU - Coordination	SU - Shared Knowledge
Extended Questionnaire	7	7	2	5	6
Session Questionnaire	2	2	0	2	2
Session Observation	0	0	0	7	0
Questionnaire After Implementation	4	0	0	0	2

Table 4.4: Overview of different concepts across questionnaires

### 4.3.2 Experimental Output Evaluation

For the controlled experiment, students are all asked to refine the same two user stories. Therefore, it may be possible to have some form of evaluation on their output as well. Some evaluation can be done regarding the correctness and completeness of their session outputs. It must be noted, however, that a less complete output does not mean a less successful session. As it is possible for user stories to go through several TA sessions before they are fully refined, having a less complete output after one session does not have to matter. However, if we can observe that the outputs of either EM or FM are significantly less complete from the other technique, then some conclusions can be drawn from that.

Regarding correctness, it can be observed if students have misinterpreted user stories and refined them the wrong way. Also, they may have included aspects that were out of scope for the user story. Completeness and correctness may be partially subjective, so no hard conclusions can be drawn from just this, but it is a good complement to the questionnaires that students fill in regarding their perceptions of the techniques. Quality of the session output can, perhaps, also be compared to how users perceive the techniques according to their questionnaire answers.





## Chapter 5 | Treatment Validation

In this chapter, we present the results for the experiment and case studies. As explained in Section 4.1.2, results of all cases as well as the experiment will first be presented individually. This means that cases will not be compared to one another yet, as will be done in the cross-case analysis in Chapter 6.

In order to analyse the data, reversed questions from the questionnaire are first transformed. If a participant answered “Strongly disagree” or “Disagree” for a reversed question, this is respectively transformed to “Strongly agree” and “Agree”, and vice versa. Then, in order to analyse the results, responses are grouped per aspect: ease of use, usefulness, intention to use, Shared Understanding (SU) coordination and SU knowledge. We build a correlation matrix first to investigate the relationship between all aspects. This is especially important in order to analyse a possible link between the first three aspects and the two SU aspects. The first three aspects are already established in literature [41], but the SU aspects are not.

For visualising the results, diverging stacked bar charts are created. With these figures, the distribution of responses to the questionnaire is displayed in an organised manner and differences can be easily compared. The way of interpreting these diverging stacked bar charts will be explained when the first figure is presented.

Initially, the plan was to have case studies that would last three months, as explained in Section 4.1.1. Arrangements had also been made to perform a case study with two software development teams for this duration, starting in March 2020. However, by that time the COVID-19 pandemic had escalated, and many companies were going into lockdown and not allowing external guests in their buildings, after which buildings shut down entirely and employees were forced to work from home. Because of this, the case studies that were supposed to start in March could not continue: the teams were both too busy with adapting their ways of working in order to work from home, and did not want to participate in the research anymore, both for their sake and for the integrity of this research.

As such, new case studies had to be arranged. In the end, four case studies were arranged: three for Example Mapping (EM) and one for Feature Mapping (FM). One EM case study consisted of only one single-case session with two teams, the other case studies were longitudinal. Unfortunately, the case studies did not allow us to answer RQ5 regarding long-term effects of TA session techniques on the implementation of a user story. RQ5 is, therefore, not answered in this research. Besides that, case studies did not last long enough for teams to try out both techniques. Therefore, only one technique is used for each individual case study, as opposed to what was desired and explained in Section 4.1.1. This does, however, prevent any cross-technique learning effects that would have occurred if teams performed a second technique after the first. This helps with the validity of the case studies.

This chapter is structured as follows. First, the controlled experiment is analysed in Section 5.1. The first case study was performed at Fizzor using EM, a low-code software development company, which is analysed in Section 5.2. The second and third case studies are con-

ducted at a large pension management firm in the Netherlands, which has requested to remain anonymous. EM and FM are both tested at an individual team and are analysed in Section 5.3 and Section 5.4, respectively. Lastly, in Section 5.5, we present the findings of the single-case EM study that was performed by professionals as a try-out for an online tool. The data of the controlled experiment and case studies is attached to the thesis in separate files.

## 5.1 Controlled experiment - Requirements Engineering course

The controlled experiment at the Requirements Engineering (RE) course took place in the afternoon of Monday, February 24th. In total, 19 students participated in the research. This was a bit less than the expected 22-24 students. As such, we only had six groups rather than eight. All groups were built up with three students, except for Group 1 that had four students. The experiment was conducted as described in Section 4.2.

As there were six rooms, each group got assigned an individual room. This way, groups are not disturbed by each other. The groups performed the techniques for the user stories in the order that is displayed in Table 5.1. There is no Group 5, due to the fact that the division of rooms and execution orders had to match the lower amount of students that were present for the experiment. A copy of the total package (Appendix C.4) was handed to each student. Three coordinators were present and every coordinator got assigned two rooms. Coordinators interacted with the participants as little as possible. Once or twice during a session, they were asked if they had any trouble with the technique. No students asked for help or needed additional clarification.

	First user story	First technique	Second user story	Second technique
<b>Group 1 &amp; 2</b>	US1	Example Mapping	US2	Feature Mapping
<b>Group 3 &amp; 4</b>	US1	Feature Mapping	US2	Example Mapping
<b>Group 6</b>	US2	Example Mapping	US2	Feature Mapping
<b>Group 7</b>	US2	Feature Mapping	US1	Example Mapping

**Table 5.1:** Experiment Execution Order

The students were first given a lecture on the subject. No explicit feedback on the lecture was asked, other than regular inquiries during the lecture if students understood what was said or if they had any questions. Students understood the materials, and one student also commented that it was a clear and easy to follow lecture that showed well how the techniques worked and that it was an intriguing subject because of how practical and “happy” it is (post-its, very agile, and communication-focused). The lecture went a little fast for two students who wanted to take notes, but they also understood everything in the end.

### 5.1.1 Discussion

After the sessions, a short discussion was held to evaluate how the students perceived the experiment and the techniques. Students generally found the first session more complicated than the second one. This can be because of a learning curve towards the techniques, and that using one technique also helps in using the other, as there are similarities between the techniques. It could also be because the student groups had gotten used to working together or that they got used to the domain during the first session. Some groups mentioned they did

not have enough time to finish the full refinement during the session and that they would have created more cards if they had the chance.

One group (Group 1) had a disagreement regarding the rules and examples with EM. They did not reach full agreement on how to organise the examples under the two rules. They let it be as it was at some point and moved on. It was explained to them after hearing this that there are no guidelines on how this ordering should be, they should do whatever feels best to them to understand how the user story works. This was explained beforehand during the lecture, but apparently, they did not fully comprehend this. If someone had facilitated their session constantly, the misunderstanding could have been prevented and this would not have happened. It is part of a learning effect of the techniques, which they did not get because they were not corrected on this until after the sessions. However, it was an explicit choice not to facilitate the sessions, but only ask once or twice per session whether or not they ran into any issues. This choice was made as there were only three coordinators over six groups, and to make sure that no group had additional help that others did not for the sake of research validity. As they did not mention they had any issues, no assistance was given to them.

Some groups came up with a lot of questions (e.g., how to split up between coaches, how the ranking of the waiting list worked) while others did not. The reason that some groups may have had fewer questions or less elaborate results might be because they do not have to implement the user story themselves - nothing has to be built and they don't know the exact domain either. This may have a negative impact on their motivation to think everything through properly during the session.

One group included a category "removed questions" to their board. This way, they showed their thought process. This can be a valuable addition to the techniques, as it can give more insight into the session, which can help for SU amongst team members.

At the end of the discussion, students were asked for their preference between EM and FM. The students were divided amongst this equally: about half preferred EM, and about half preferred FM.

### 5.1.2 Overall Results

Before going into the results themselves, the aspects are compared to one another. For this purpose, a correlation matrix between all concepts can be found in Table 5.2 and shows correlations between all of them. All correlations are significant ( $p < 0.05$ ), albeit some are stronger than others. In Section 4.3.1, it was argued that perceived SU would influence perceived usefulness. What is interesting to see here is that both SU components have an even stronger correlation to perceived ease of use than to usefulness. This suggests that SU plays a more prominent role on perceived ease of use and actual efficiency than it does on perceived usefulness and actual effectiveness in this experiment. This may call for an altered adaption of the MeM model that we presented in Figure 4.5. However, the case studies must also confirm this before making such alterations.

Correlation	Perceived Ease of Use	Perceived Usefulness	Intention to Use	SU - Coordination	SU - Shared Knowledge
<b>Perceived Ease of Use</b>		0.78	0.56	0.76	0.82
<b>Perceived Usefulness</b>	p = 0.0000		0.76	0.54	0.65
<b>Intention to Use</b>	p = 0.0002	p = 0.0000		0.40	0.33
<b>SU - Coordination</b>	p = 0.0000	p = 0.0005	p = 0.0126		0.61
<b>SU - Shared Knowledge</b>	p = 0.0000	p = 0.0000	p = 0.0423	p = 0.0000	

Table 5.2: Pearson's Correlation between aspects

Looking at the results, overall results for the techniques can be found in Figure 5.1. In this figure, all responses are combined into the overall perception of techniques. The chart can be interpreted as follows. The questions are grouped per response category. Right of the 0 percentage line are the positive responses ("Agree" and "Strongly agree"), the negative responses ("Disagree" and "Strongly disagree") are on the left. The "Neutral" category is split in two: half of the responses are mapped to the left, and half to the right. This way, the further the bar chart is to the right, the more positive the overall responses are.

However, the position of the chart is not the only thing that should be considered, as the division between the responses is also important. For example, as can be seen in Figure 5.1, both techniques have been perceived rather positively: 72% and 70% of responses are positive for EM and FM, respectively. 15% and 18% are neutral, and only 13% and 12% are negative. By looking purely at the position of the bar chart, both techniques seem to be rated almost equally positive. However, EM has more "Strongly agree" responses than EM. This indicates that EM is perceived slightly more positive overall than FM.

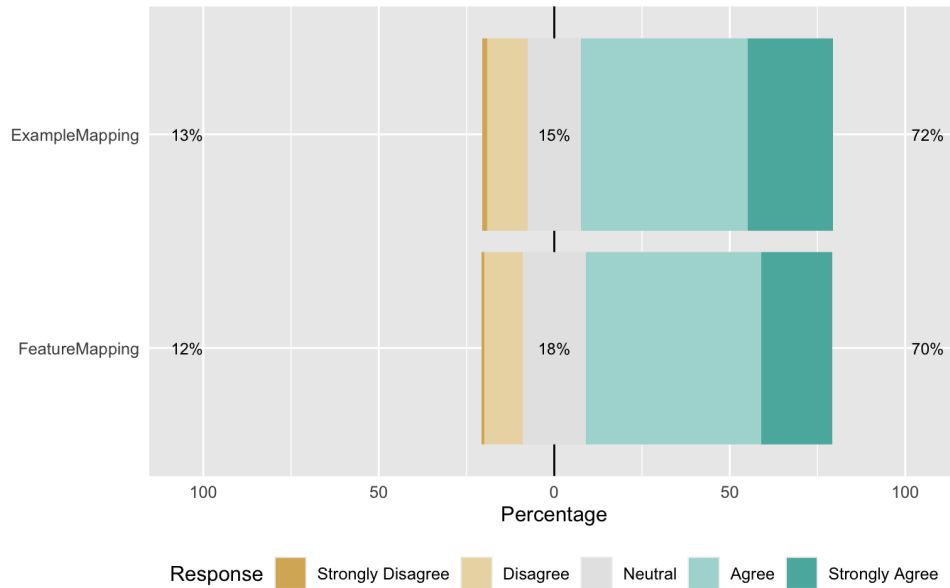


Figure 5.1: Total Responses

Zooming in on the techniques more, we present Figure 5.3 and Figure 5.2, where the ratings of all aspects are visualised per technique. Figure C.10 can be found in Appendix C.5, where the results of each aspect can be more easily compared between techniques. Looking at Example Mapping, the Knowledge aspect scores highest, with 84% positive ratings. This shows that team members perceived that they have good knowledge and agreement on the functionality of the user story. Coordination also scores well with 74% positive responses, 13% neutral and 14% negative. Both also have a good portion of the positive responses in the “Strongly agree” category. From this can be concluded that participants have a high SU regarding the user story after having used EM.

When analysing the Feature Mapping responses, knowledge is also the highest-rated aspect with 81% positive responses. This is slightly less than with EM, but on the other hand, there are also fewer negative responses: 6% for FM as opposed to 10% for EM. This is due to the fact that there are more neutral responses. However, coordination seems to be less with FM than with EM. With FM: only 68% of responses were positive. FM is a lot more structured than EM, which makes it interesting that it scores lower on coordination. This may have to do with the fact that participants are new to the technique and to working together in the chosen teams with their peer students. Being more structured, FM may have a higher learning curve than EM before gaining advantages out of it, which means they have to perform multiple sessions using a technique before getting acquainted with it. This learning curve is not observed during this experiment due to the fact that each technique is only used once.

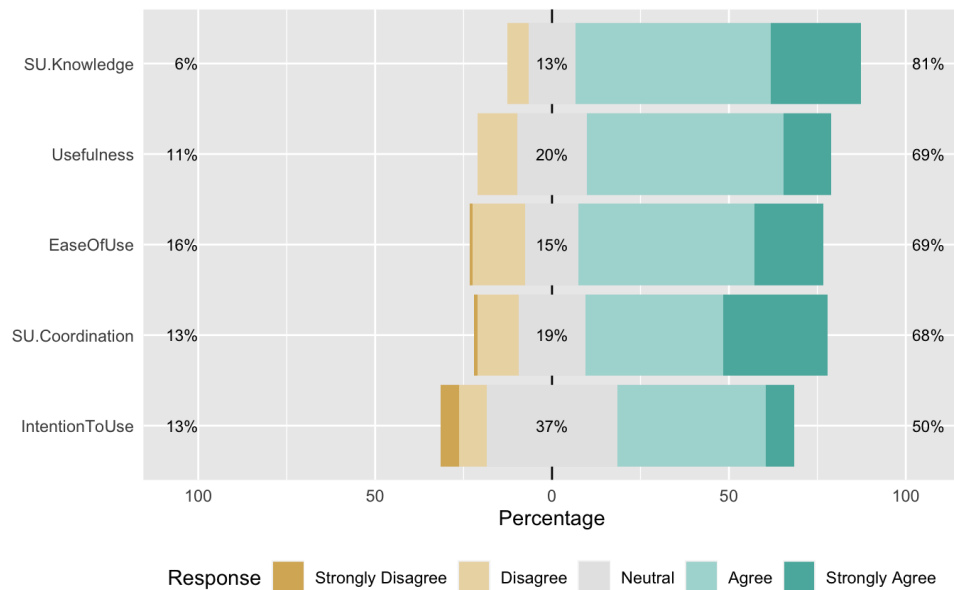
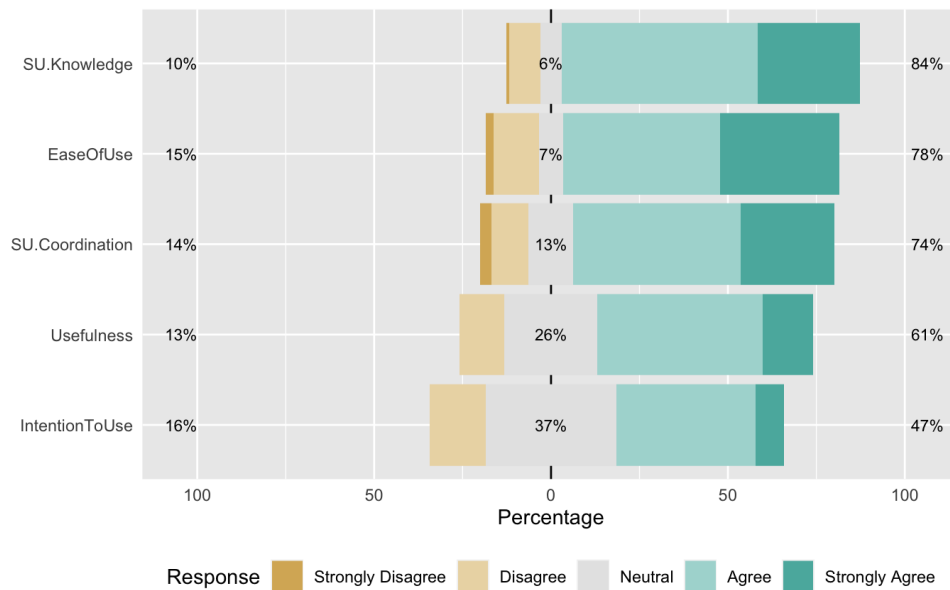


Figure. 5.2: Feature Mapping Responses



**Figure. 5.3:** Example Mapping Responses

This learning curve may also be the reason why ease of Use scores much higher with EM than it does for FM: they are rated 78% and 69%, respectively, with more strongly positive (“Strongly agree”) responses for EM as well. It does not necessarily have to be due to a learning curve, however, as EM may simply be an easier technique to use overall. With less structure, teams may be able to use the technique in a way that suits their preferences and current ways of working more they can with FM.

Usefulness does score higher for FM than it does for EM with 69% instead of 61%. Despite finding the technique more difficult to use and scoring worse on coordination, participants still perceive it to be a bit more useful than EM. The added structure of FM may make a technique seem more useful to participants, despite rating SU and ease of use lower.

Looking at intention to use, both techniques score significantly lower with this than they do with the other aspects: about half the responses are positive and 37% of the responses are neutral. This is overall still a positive rating as only 16% and 13% of the ratings are negative, but it is nonetheless lower than the other aspects. FM is rated slightly higher, but also has more strongly negative responses (“Strongly disagree”) than EM, so there is no clear winner in this aspect. It is unclear why students rated this aspect significantly lower. It may be because they are not familiar enough with the techniques or with the team composition that they do not see the possible advantages of the techniques that they do get according to the other aspects. It is also likely that because they, being students and not professionals, are not familiar with user story refinement at all, and therefore cannot compare the techniques well enough with other techniques that exist.

### 5.1.3 Results per Aspect

In this section, results are presented per aspect and the division of answers can be seen per group. All remaining results from the controlled experiment can be found in Appendix C.5. First, the division of responses per group of all aspects combined can be found in Figure C.11

in Appendix C.5. Results of both EM and FM are combined in the same graphs so that it can be observed if groups rated the two techniques significantly different. Some differences can be observed. For example, Group 1 rated EM a lot better than they did FM. For Group 4, this was the other way around. Group 1 was also the most negative with the ratings for both techniques with the most negative replies. This finding will be used when analysing their output in Section 5.1.5.

Ease of use is mapped out per group in Figure C.12. Similarities are observed between how techniques are rated per group: if a group is more critical for one technique than average, they are also critical of the other. However, except for group 6, all groups rated EM higher than FM.

Usefulness results are visualised in Figure C.13. Two groups rated EM higher on usefulness than FM (Group 1 and Group 2). The other five groups rated FM higher. The biggest deviation between ratings can be found in Group 4: they rated 33% positive and 33% negative with EM and rated 95% positive for FM and 0% negative for FM. Considering the fact that Group 4 performed FM first and EM later, this cannot be accounted to any cross-technique learning effect that was mentioned during the discussion (Section 5.1.1).

Looking at intention to use in Figure C.14, only Group 7 rated both techniques exactly the same. With the other groups, there are big deviations between how the techniques are rated. Clear preferences are observed with Group 1, 4 and 6. For groups 2 and 3, the position of the bar chart indicated a preference towards one technique, but the division between the neutral, positive and strongly positive responses for these groups does not allow for concluding whether they intend to use one technique more than the other. With Group 2, the bar chart indicates a preference towards EM, but there are several strongly positive responses for FM that are not present for EM. With Group 3, FM seems to be the winner in this aspect on first sight, but EM has more positive responses overall and FM is positioned more towards the positive side because of many neutral responses.

With Shared Understanding, knowledge was rated higher than coordination for all groups using all techniques, except for Group 2 in their FM session. They rated coordination at 93% positive and knowledge at 78%. Other than this instance, knowledge scored higher for everyone. This may again be because the teams only used each technique only once and did not master it yet and because the participants are not used to working together in these teams. These aspects arguably have more effect on the coordination than they do on the shared knowledge that is created during a session.

#### 5.1.4 Results per Group

Results per group in practice give the same results that are given in the above section. However, the different view on the data may provide other compelling insights. Therefore, this view is also added to the research. The results of this are found in Figure C.17 up until Figure C.28. What can be seen here, is that most groups give a relatively uniform response, except for intention to use: the ratings of the other four aspects are in most cases really close to one another, but intention to use does not follow that pattern as much. The uniformity of responses within a group could not be observed in the previous view and gives the insight that scores within a group for one aspect often also indicates the score for other aspects (excluding intention to use). The correlation between all aspects suggested that as well, but that was on an overall level rather than within a team. The EM session responses of Group 4 does not contain uniformity in the replies, however, and is the exception to this insight. The different aspects are all rated quite differently from one another, only knowledge and ease of use are close to one another.

### 5.1.5 Output Analysis

All groups worked on the same two user stories, of which the outputs can be found in Appendix C.6. In order to get some objective quality measurement, we analyse the completeness of the outcome. In order to get the completeness, first all rules, examples and questions are aggregated of all of the students' outcomes as well as the example outcomes we generated (Appendix C.1 and Appendix C.2). Tasks and consequences are not analysed objectively but rather as a part of examples and taken into consideration, in order to create uniformity over EM/FM and because many different implementations of tasks and consequences are possible without one being better than the other. Wrong requirements would have resulted in a penalty to their score, but no incorrect results were observed. Groups did write down questions that were out of scope for the user story, but this is not considered wrong.

The result of this aggregation can be found in Appendix C.7. For questions, two categories are defined: important questions that need to be answered to implement this user story properly, and non-pressing questions that are not as important. As the latter category exists of questions that are not often seen in more than one session outcome, these cannot weigh as much as the important questions. The non-pressing questions were disregarded for calculating the percentage of the total, as they are not deemed crucial for successful implementation.

If a group has two rules that could have been put together, this will be considered one rule for the calculations. There is no right or wrong concerning this: a team can either have more rules and fewer examples per rule, or fewer rules and more examples per rule. If these rules are not combined in the calculations, groups with fewer rules will be at a disadvantage for the quality metrics, despite having an output that can be just as good as one with more rules. Two similar questions are also grouped as such.

The results of the completeness analysis can be found in Table 5.3 and Table 5.4 for US1 and US2, respectively.

Group	Technique	Rules	Examples	Questions	Total	Percentage of Total
Group 1	Example Mapping	3	4	2   1	9   1	75%
Group 2	Example Mapping	3	6	3   0	12   0	100%
Group 3	Feature Mapping	1	2	1   0	4   0	33%
Group 4	Feature Mapping	3	4	1   2	8   2	67%
Group 6	Feature Mapping	1	3	0   0	4   0	33%
Group 7	Example Mapping	2	4	0   1	6   1	50%
Total possible		3	6	3   4	12   4	100%
Average		2.17	3.83	1.16   0.5	7.16   0.5	60%

Table 5.3: US1 completeness analysis



Group	Technique	Rules	Examples	Questions	Total	Percentage of Total
Group 1	Feature Mapping	1	3	0   0	4   0	22%
Group 2	Feature Mapping	0	3	1   0	4   0	22%
Group 3	Example Mapping	3	5	1   1	9   1	50%
Group 4	Example Mapping	2	4	4   4	10   4	56%
Group 6	Example Mapping	3	4	1   2	8   2	44%
Group 7	Feature Mapping	1	2	1   0	4   0	22%
Total possible		5	6	7   5	18   5	100%
Average		1.67	3.5	1.33   1.17	6.5   1.17	36%

Table 5.4: US2 completeness analysis

One thing must be noted about this completeness analysis. During the discussion that is elaborated in Section 5.1.1, some groups mentioned not to have enough time. Since these techniques and team composition are both new to the participants, not having complete specifications after one session is possible and should not necessarily be considered bad. It is possible to organise several EM or FM sessions to refine one user story if it is not complete yet after one session. This is an opportunity that was not given to the participants, which is why a lack of specifications cannot play a significant role in determining whether or not a technique performed well.

However, two clear observations can be made from these analyses. Firstly, US1 has overall higher completeness than US2. When analysing all outcomes, this can be explained by a higher level of complexity with US2 in the specifications than with US1. Multiple questions or rules have come up in the outcomes that were not specified in the case description, simply because they were not thought of by us and not present in the example outputs either. This is why US1 only has three rules and eight questions in total (combining important and non-pressing questions), whereas US2 has five rules and twelve questions. Looking at the completeness levels, US1 was possible to complete within one session with Group 2 having 100% completeness, but US2 would have required at least one more session for every group in order to get close to 100%.

Secondly, EM seems to be more complete overall than FM. Looking at the three most complete outputs for each user story, five out of six were made using EM. This can be caused by the fact that EM scored higher on ease of use, as explained in Section 5.1.2. As participants found EM easier to use after one session, it makes sense that they would also produce more output during that session as they struggle less with the technique itself.

As mentioned in Section 5.1.3, Group 1 rated EM a lot higher than FM and Group 4 rated FM much higher than EM. Looking at completeness, it makes sense that Group 1 rated EM higher, as they scored much better on EM completeness (75%) than they did on FM (22%). However, Group 4 actually scored well on both: 67% on FM and 56% on EM. The 56% for EM was with US2, where they actually even scored highest of all groups. Therefore, no clear link between output completeness and participant perception of techniques can be observed.

### 5.1.6 Conclusion

When considering all results that we presented in this section, it can be concluded that both EM and FM performed well, albeit partially in different aspects. Firstly, when looking at the overall perception of the techniques, both scored high, with EM being slightly higher. Shared knowledge scored highest with both techniques and coordination also scored high, which shows that both techniques deliver on their promise of creating SU between team members. EM scored higher on ease of use, whereas FM was deemed more useful by participants. During the discussion, participants indicated that they have a preference over either EM or FM, which we also observed in the questionnaire results of three out of six groups.

Intention to use scored considerably lower overall than the other aspects. This is likely due to the fact that students are inexperienced and do not have knowledge of refinement techniques in order to compare them accurately. This is a limitation of the controlled experiment that was predicted beforehand, which is why case studies with professionals are also part of this research.

EM scored higher in terms of completeness than FM did. Combining this with the fact that EM was perceived easier to use, we believe that EM is a more suitable technique for teams to learn when they have little time to refine requirements. FM was considered more useful and might, therefore, deliver better results eventually, but it has a steeper learning curve which makes it less suitable as a new technique to learn when there is little time available. As the learning curve of either technique is not researched during this controlled experiment, however, no conclusions can be made about that.

## 5.2 Case Study - Fizor Example Mapping

The first case study that took place was at Fizor, a low-code software development company. At Fizor, a total of eight Example Mapping sessions were performed. The first session was held on April 14th 2020 and the last session on May 28th 2020. At Fizor, we used the EM sessions to specify requirements for a new upcoming project. Software implementation of the project was supposed to start during the case study, but the company decided to put this on hold due to circumstances during the case study. Nonetheless, the case study continued so that the requirements would be specified correctly as soon as the project does start implementation.

For the project, a set of user stories was already created, divided amongst several features. For these user stories, no requirements were specified yet. However, an initial effort estimation was created for each user story regarding how much time it would take to implement. User stories were generally rather small: some user stories would take one or two working days to implement, but the effort of many user stories was estimated at 0.5 to 2 hours of work. We therefore decided to group those small user stories together as much as possible for some EM sessions, as it was not deemed valuable by the participants to have a 30-minute EM session for a US that would only take 30 minutes to implement too. The Product Owner would group together user stories that regarded similar functionalities before sessions. The first three sessions regarded one user story, and the remaining five sessions were held by combining 2-9 user stories for each session.

The EM sessions were held online, with the use of Microsoft PowerPoint. Through this platform, all participants could access and work together. One file was used for all sessions, with a new slide for each session. This way, participants could easily switch back and forth between session outputs if they needed to look up something from a previous session. An example of how an output from one of these sessions looks like can be found in Figure 5.4. Cards of all four concepts were created as a template, which could then be easily copy-pasted for the session itself.



Figure. 5.4: Output of EM Session 2

With this case study, the same three participants were present during the sessions: the Product Owner, a Business Analyst and a Developer. In order to preserve privacy, the participants are randomly labelled Person 1, Person 2 and Person 3. In order to keep their anonymity, gender is also not given and all participants are referred to in masculine form (“he” or “him”).

The correlations between the different aspects can be found in Table D.1 and Table D.2 for the long questionnaire and session questionnaire, respectively. We observe that only a few aspects seem to be correlated to one another. With the long questionnaire, a correlation is found only between intention to use and perceived ease of use and between intention to use and shared knowledge with  $p < 0.05$ . With the session questionnaire, the only correlation that is found is between usefulness and ease of use. This is remarkable, as causal correlations have already been proven between ease of use, usefulness and intention to use [41], which are not all present here. We believe the reason for this is that the EM technique was not merely judged purely on the technique itself, but also in regards to the user stories being refined. We made several observations during the case study that can account for no consistent correlations between aspects. Firstly, a session may have been useful with a good outcome despite no great coordination between parties. This can make it so that the technique is perceived useful or easy to use, without scoring high on SU coordination. The other way around applies as well: participants may have created a high SU overall regarding a user story, without rating EM highly on ease of use, usefulness or intention to use. Secondly, EM may not have been very suitable as a technique for certain (sets of) user stories, which could affect any aspect in one way or another, but the session can still be rated positively on other aspects. With Fizzor, this mainly applies to the really small user stories which we have now had to group together. This adaption worked out well, but it does show that EM is less desirable for these individual user stories. We believe this may impact the intention to use negatively, but participants may not necessarily rate the other aspects low as well because of this.

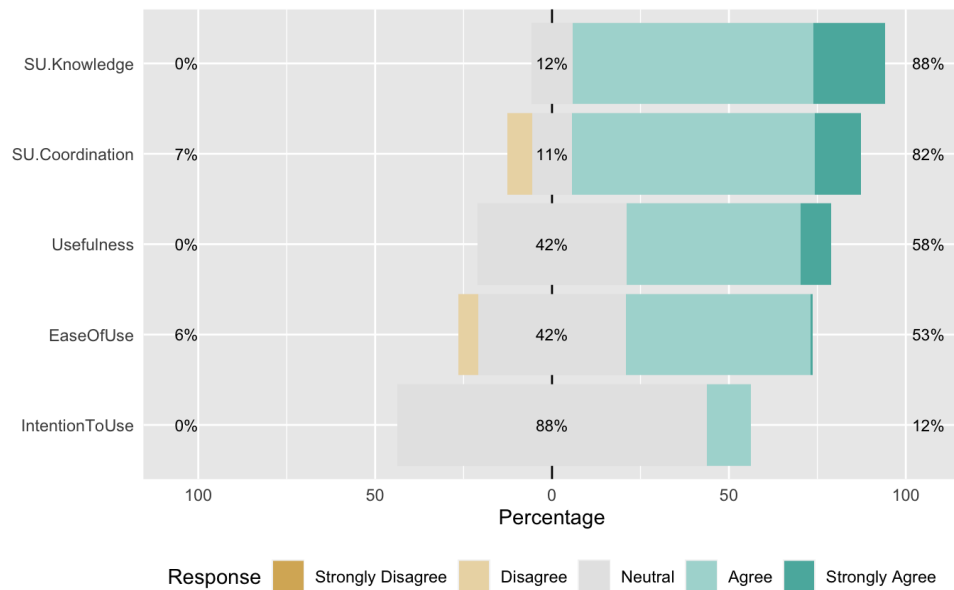
By creating correlation matrices, we hoped to validate our adapted MEM model. From the analysis of the correlation matrix, however, we cannot conclude that the adapted MEM presented in Figure 4.5 is correct. We do still believe that adding the two SU aspects is a valuable addition to the model for refinement techniques and for this research, but the added arrows that show relations between SU and other aspects may not be justified.

### 5.2.1 Session Results

Right before the first session, EM was explained to the participants. After the explanation, unfortunately, Person 3 had to leave unexpectedly right before the session itself started. We agreed to still let the session continue with the other two remaining participants. Because of this, we agreed that both session 1 and session 2 would be evaluated afterwards using the long questionnaire. This way, a good first impression of the technique could still be given by all three participants. Because of the total amount of sessions, only the last session, session 8, was also evaluated using the long questionnaire. The other sessions, session 3 up until session 7, are evaluated using the session questionnaire. This leads to eight results in total for the long questionnaire and fifteen for the session questionnaire.

The overall questionnaire results per aspect of all eight sessions are presented in Figure 5.5. In this figure, we combined the results of the long questionnaire with the session questionnaire in order to get an average score of all eight sessions. This is except for intention to use, which is not evaluated in the session questionnaire. Therefore, intention to use is built up of session 1, 2 and 8.

From the overall results, we can conclude that SU is rated highly by the participants, both on knowledge and coordination. Knowledge has 88% positive ratings and coordination has 82%. Coordination had some negative ratings, but overall the participants were positive.



**Figure. 5.5:** Results per Aspect

In order to analyse the eight sessions more into detail, we created Figure 5.6 for the session ratings from participants, together with Figure 5.7, which shows the observational ratings of the sessions. The session ratings are based on all aspects, except for intention to use when the session questionnaire was used, whereas the observation was purely about the process itself and the coordination between team members. Besides these figures, we also present Figure D.1, Figure D.2 and Figure D.3 in Appendix D, which show the aspect ratings of each participant.

As mentioned before, Person 3 had to leave unexpectedly with the first session, and we decided to still have the session with Person 1 and Person 2. This made for a difficult start, which is mainly observed in the observation results. The participants still had to learn the grips of the technique and how to specify both rules and examples best. It was observed that Person 1 acted very proactively and ended up writing down all the cards, whereas Person 2 was more reactive by responding when something was explicitly asked, without initiating conversation himself. This is why the observation results are rated quite negatively, as participation and coordination could have been much better. The Example Map was not finished entirely within the time of the session, and we decided to finish it and discuss it again next time when Person 3 could also join.

During the second session, the first Example Map was first explained to Person 3. One of the rules was not written down very clearly, and Person 1 and Person 2 had to discuss what it meant again. Afterwards, additional rules, examples and questions were created. The result of this session is shown in Figure 5.4. This was finished rather quickly, and we decided to refine a new user story during this session. Other than the discussion Person 1 and Person 2 had regarding the forgotten rules, however, participation of Person 2 was again not optimal, and neither was that of Person 3. A lot of initiative came from Person 1, and only when prompted would Person 2 or Person 3 intervene. The end result was still a good Example Map that explained the user story well, but coordination could be improved. This may be due to the fact that Person 2 and Person 3 are unfamiliar with the technique and feel objected to trying out something they do not know well. The fact that everyone was working from home and not

co-located at the same location probably did not help either, as it made it easier for them to not participate as actively without it immediately being apparent.

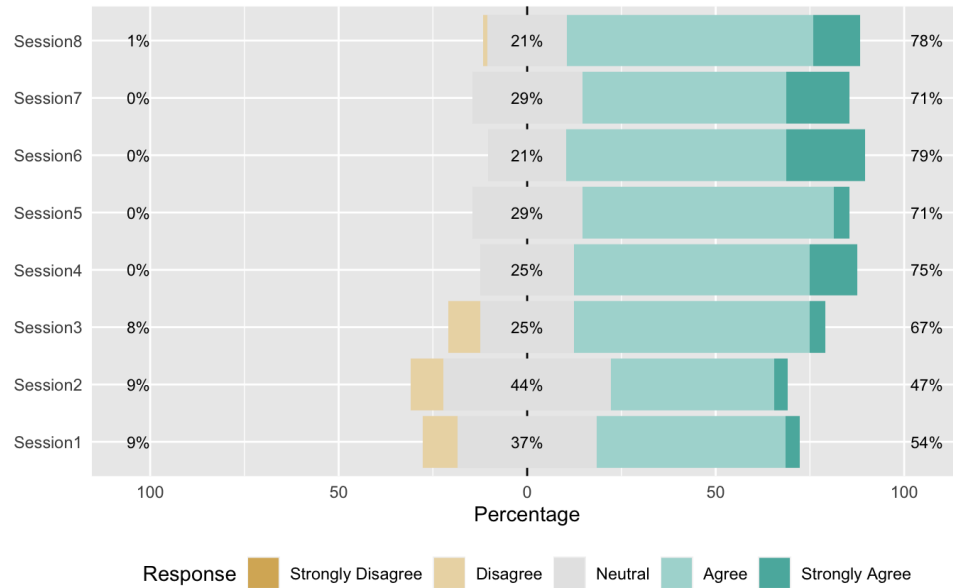


Figure. 5.6: Results per Session (Self-reported Data)

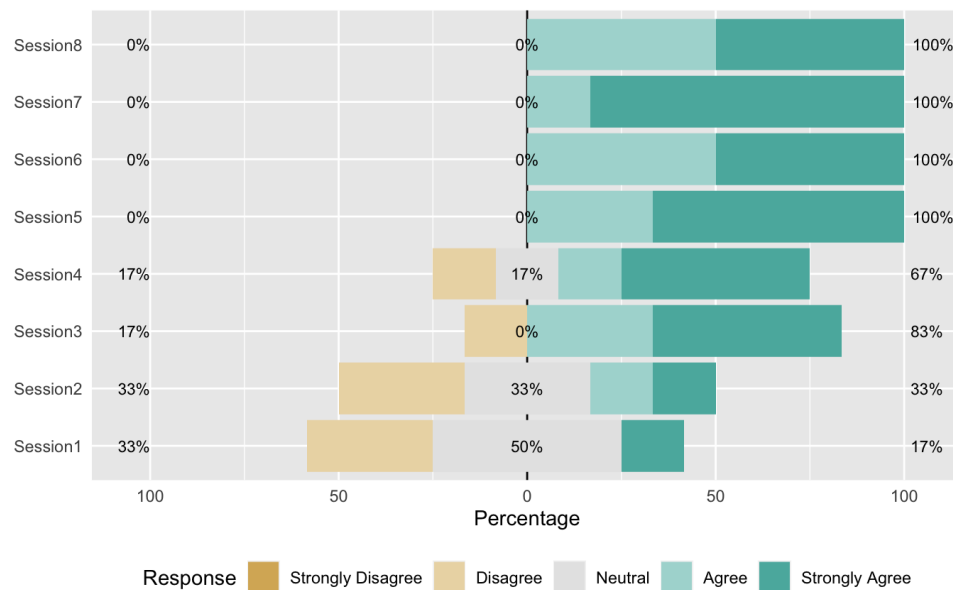


Figure. 5.7: Results per Session (Observation Data)

The next two sessions improved in terms of participation from Person 2, who now more actively discussed the user stories with Person 1. Person 1 was still doing most of the work but did get helpful inputs from Person 2. However, Person 3 was again not actively participating. At the end of the fourth session, it was mentioned that EM helped them to trigger conversations that make them come up with valuable questions that they would not have come up with otherwise. This indicates that, despite the fact that participation should still be improved, EM was still helping the team as a valuable technique. An improvement in the ratings from both the participants and the observation regarding the sessions also confirm this. We believe this is a learning effect of EM: once the team has had a few sessions, they become more confident with the technique, and that may be why Person 2 now involved himself more in the conversation. Person 3 participated in one session less than person 2 did, which may explain why Person 2 was the first to participate more actively.

In order to improve participation of team members, we had a small intervention before session 5 with Person 1 regarding the participation of Person 2 and Person 3. We agreed that he would try and specifically ask the others at the beginning of the session what they thought was a good way to start the session in terms of rules and examples, rather than taking the initiative himself. This small change gave good results: Person 2 and Person 3 both participated much more actively in this session by sharing their thoughts and ideas on the user story and how it could be written down as rules and examples. The fact that we intervened, albeit a small intervention, might be a big factor in the performance of the sessions. Person 2 was already beginning to participate more, so perhaps Person 3 would have done so by himself eventually as well, but it must still be noted that we possibly changed influenced the case study here.

From here on out, Person 2 and Person 3 both participated actively in all three remaining sessions, without Person 1 having to explicitly trigger them to join in on the conversation. We observed that the added inputs from them significantly improved the sessions and outputs. The observation results are, therefore, very positive from session 5 onward. The participant results also show improvement, with session 4 already being much more positive than the previous ones, but session 5 and onward also all have very positive ratings. In the observation results, session 4 was rated significantly lower than session 5, because there were still improvements to be made. However, as participants may have noticed an improvement over the first sessions already, they may have subjectively rated session 4 very high already. This makes a lot of sense because they are very involved in the session, while an observer can keep a more objective perspective.

As big improvements were observed during the course of the longitudinal case study, we created Figure D.4 and Figure D.5 in order to get a closer look at exact differences between the second session, where all three participants were present for the first time, and the last session. In these two figures, we observe great differences. The SU knowledge aspect already had a high score of 94% after session 2, but scored 100% positive ratings after the last session. Coordination also greatly improved, going from 60% positive ratings without any very positive ratings, to 87% positive ratings and also with very positive ratings. Usefulness and ease of use, however, made even more significant improvements: these aspects were first rated positively for 33% and 19%, respectively, with ease of use also having 24% negative ratings, whereas the ratings after session 8 show 71% for usefulness and 76% for ease of use, without any negative ratings.

These improvements really show how the team members have started to master the technique and get more benefits from it as well, which is why they have rated it significantly higher. However, intention to use is still rated quite low and did not improve at all between the second and eighth session. In both sessions, merely 17% was rated positively, and the remaining 83% was all neutral. We believe this can be explained by the context of the project that Fizor is in,

with such small user stories that are sometimes not worth refining thoroughly in their opinion, as that would take more time than implementation itself. We solved this issue by grouping user stories together for EM sessions, which worked out well, but it was not an ideal situation as it required some extra effort. This would be a sensible reason for the team members to not desire EM for all of their refinements, which could be why intention to use scored poorly.

During the case study, we also observed the order in which the EM session was held. In this case study, the team had basically one order that was adhered to for all sessions: during all eight sessions, the team members started by coming up with one or two rules, followed by examples for those rules. After they had come up with those examples, they would think of additional rules that would then be accompanied by additional examples. They could also have started out from examples but did not do so. This way of working during the sessions seemed to work well for them, which is probably why they never considered starting with examples.

### 5.2.2 End Evaluation

After the eighth and final EM session, we have conducted an end evaluation with the participants. The participants were first asked on their general opinion on EM. Person 3 started by explaining that he liked the technique and that the more sessions were held, the more confident he became in them and the more he liked them because of this. He did mention that he has never knowingly used defined refinement techniques before, which made comparing the techniques to others difficult. Nonetheless, he believes that they learned a lot because of the conversations that EM triggered, which would help them to make fewer mistakes when implementing a user story. He also believes that the output of an EM session really helps to discuss a user story with stakeholders outside of the team: having defined rules and underlying examples will make a user story rules more understandable to others. However, he is not sure how much time would be allocated during a new project to them to have these refinement sessions and that they may not gain full advantages of the technique because of that.

Person 2 followed, stating that he thought the sessions have been really useful. Especially because everything is written down clearly and that you can look back at an output of a previous session and immediately know what the requirements of that user story are. Despite mentioning that the user stories they discussed may have been too small or generic to go into much detail sometimes, he liked the technique and thought it was clear. The technique also helps to write down everything you need for a user story in a clear overview.

Person 1 started his elaboration by agreeing with Person 2 and Person 3 regarding their views on EM. He stated that user stories would regularly be pretty straight forward, but that sometimes they also changed a lot during the refinement, and that they would have therefore missed a lot during implementation if it had not been for these EM sessions. He believes that they have come up with requirements that they would otherwise definitely not have gotten if it was not for EM. He mentioned that the grouping of user stories was a good choice, but that they have to be careful not to group too many for one session. In the future, he would use EM by first looking through the user story set and identifying which user stories require refinement, based on estimated effort, complexity, vagueness, and amount of possible implementations. The easier user stories would then not be grouped anymore and would simply be implemented without having EM sessions.

Besides the general opinion, two participants mentioned that the questionnaire asked them about acceptance criteria, while they had the feeling that acceptance criteria were not discussed with this technique. It was explained to them that rules represented the acceptance criteria of a user story, but apparently, this was not conveyed to them properly. Therefore, they say they rated these questions slightly less positive. The questionnaire mentions acceptance criteria



rather than rules, which was done so that we can more easily generalise the questionnaire for other refinement techniques as well where nothing is said about rules. This choice was also made because the term acceptance criteria is a term that professionals are mostly more familiar with, as it is a commonly used term for user story refinement. Retrospectively, however, it may have been better to call them rules in the questionnaire as well, considering this feedback, rather than acceptance criteria.

Participants were also asked about their opinion on the tool that was used for the EM sessions, Microsoft PowerPoint. Person 1 mentioned that PowerPoint was a good solution and that he actually preferred an online tool over having to use this technique in real-life with post-its on the wall. He did mention that one slide may have limited spacing and that if a slide is full it could seem like they were “finished” with refinement. Rules could also be pasted underneath each other rather than beside each other (and have been), however, and this problem therefore did not seem to be present during the case study from an observational point of view. Person 2 mentioned that he liked the fact that one PowerPoint slide gave a good overview of the entire user story.

### 5.2.3 Conclusion

During this longitudinal case study at Fizor, eight EM sessions were held with the same three participants. The first few sessions had some difficulties and participation of team members could be improved, which they did after the first half of the case study. The participants all rated EM high on both SU aspects, which shows that EM delivers on its claimed advantage of gaining shared understanding amongst team members. We have observed a big learning effect in this team, where they started to get much more benefit out of EM over time. Except for intention to use, all other four aspects greatly improved in rating during the course of the case study as seen by comparing session 2 with session 8 into more detail. The observations that we made during the sessions also showed significant improvements over time.

Overall, participants became very positive about the technique. Combining all the above facts, we can conclude that EM performs well as a technique for user story refinement in the context of this longitudinal case study. However, there may be user stories that are so straightforward or small, that EM can still be useful but likely costs more time than is desired. Therefore, the grouping of similar user stories can be considered, although we observed that there are limits to the number of user stories that should be grouped for one session. Another consideration is to have many smaller EM sessions with only one straightforward or small user story, or perhaps of only a couple that are grouped. Instead of a 30-minute session, teams could opt to have EM sessions of only ten minutes for these type of user stories. However, shorter EM sessions are not investigated in this research, so no definitive insight on this can be given without further research.

### 5.3 Case Study - Pension Management Firm Example Mapping

The second case study was at a software development team at a large pension management firm. The first session was held on May 27th and the last session on June 18th, 2020. In total, five sessions were held over the course of this case study. The team that participated in the case study designs and develops the technical links between the pension management firm and third parties. At first, we had our concerns if what they develop would not be too technically-focused rather than creating normal functionality, and therefore maybe not suitable for EM. However, we have found out with this case study that this is not the case, as the results were still positive. The team had to find their way in what they would consider rules and what they would consider examples, which was more difficult due to the technical nature of their products, but they achieved a shared understanding on this after the first or second session.

This team has been working together on their products for a long time already. Many user stories were already created, of which several ones from the upcoming sprints were selected for the case study by the Product Owner and Scrum Master of the team. The Product Owner was present during four out of five sessions, and the Scrum Master was present during all. In total, eight team members participated in the sessions. The first session was held with three, after which all sessions had four or five participants. As with the Fizzor case study, Microsoft PowerPoint was used as an online tool for the case EM sessions.

The Scrum Master mostly took on a facilitating role during the sessions. He actively tried to capture what others were saying and write it down in the form of rules, examples and questions. This helped the rest of the team to continue the discussion uninterruptedly while having to worry less about writing everything down on cards. However, it may also result in the team members paying less attention to the output itself as they are busy discussing the user story with each other. This consequence was also observed after the fourth session of the case study. It may have a negative effect on the SU of the team members as they may not have full knowledge of the output of the session. On the other hand, being able to keep discussing the user story can positively effect on SU as this means they can discuss the user story itself into more detail. This could mean that EM is negatively impacted by this, but that the Three Amigo session principle itself is affected positively. As the team seemed to be content with this way of working during the sessions, we decided not to intervene.

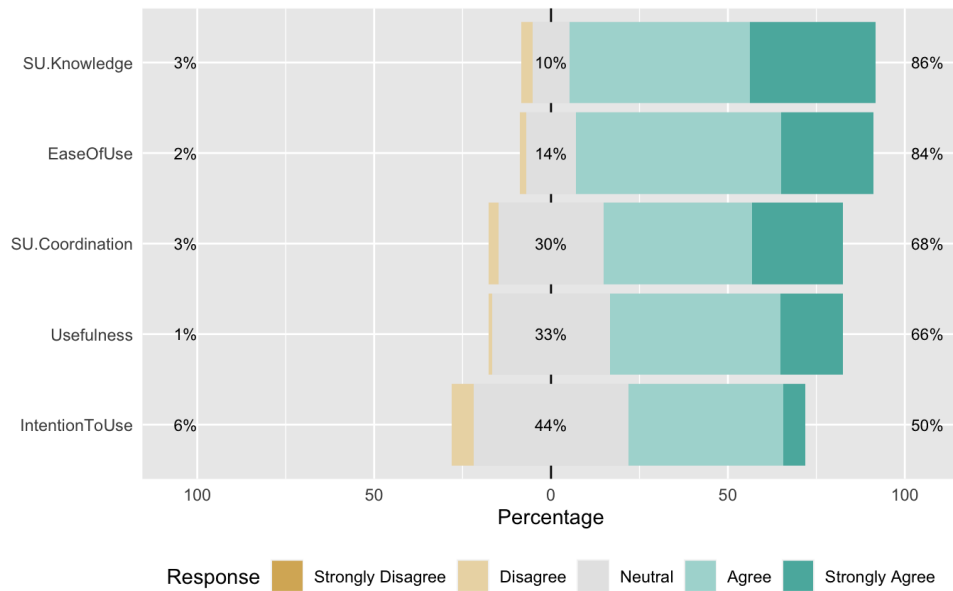
Unfortunately, we did not have a 100% response rate on the questionnaire. As members of this team often had meetings right after the EM session, they were sometimes not able to immediately fill in the questionnaire and then forgot to fill it in later. This is a threat to the validity of the data as it is incomplete. In the end, all eight participants did fill in the long questionnaire. However, the session questionnaire was only filled in nine times. This comes to a total of 17 results, while it should have been 23. This means that we miss six, about 25%, of the responses. We did have an evaluation with after the fourth sessions, together with observations this helps to mitigate this issue. An overview of the number of responses and the number of attendees is presented in Table 5.5. The difference between the number of attendees and responses shows the missing responses per session.

From the gotten responses, we created correlation matrices of the long questionnaire and session questionnaire to see the relation between aspects, which can be found in Table E.1 and Table E.2, respectively. In the long questionnaire, all aspects other than intention to use with both SU aspects have a correlation of more than 0.50. However, only the SU aspects, coordination and knowledge, show significant correlations ( $p < 0.05$ ). As this is a small sample group with only eight responses, getting significant results is difficult. The session questionnaire shows two significant correlations: usefulness with coordination and usefulness with knowledge.

Session	Number of Attendees	Number of Responses
1	4	3
2	5	4
3	5	4
4	4	4
5	5	1

**Table 5.5:** Number of attendees and responses per session

The results per aspect are presented in Figure 5.8. In this figure, we can observe that SU knowledge is rated highest with 86% positive responses, of which also more than one third is also rated very positively. Ease of use is also rated highly with 84% positive responses. It is impressive that ease of use is rated so highly, despite the team's difficulty at the beginning of the case study of defining the difference between rules and examples. This difficulty may, however, have affected on the other three aspects, which are still rated positively, but significantly lower than knowledge and ease of use. Coordination, usefulness and intention to use are rated positively for 68%, 66% and 50% of the ratings, respectively. In order to analyse the sessions

**Figure 5.8:** Results per Aspect

individually, we present Figure 5.9 for the participant ratings of each session and Figure 5.10 for the observation results. What can immediately be observed by comparing the observation results with the total aspect results is that we rated coordination much higher with the observations than the team did themselves. This may be because this team is already very good in this aspect by themselves and are therefore more critical about it. They had seemingly good discussions during the meetings, which is why coordination was observed very positively, but perhaps this is a considered "normal" to them. Also, this case study started around the end of the Fazor case study, where proper coordination was a challenge at first. This may have influenced the subjective observation of this case study, where coordination went a lot better.

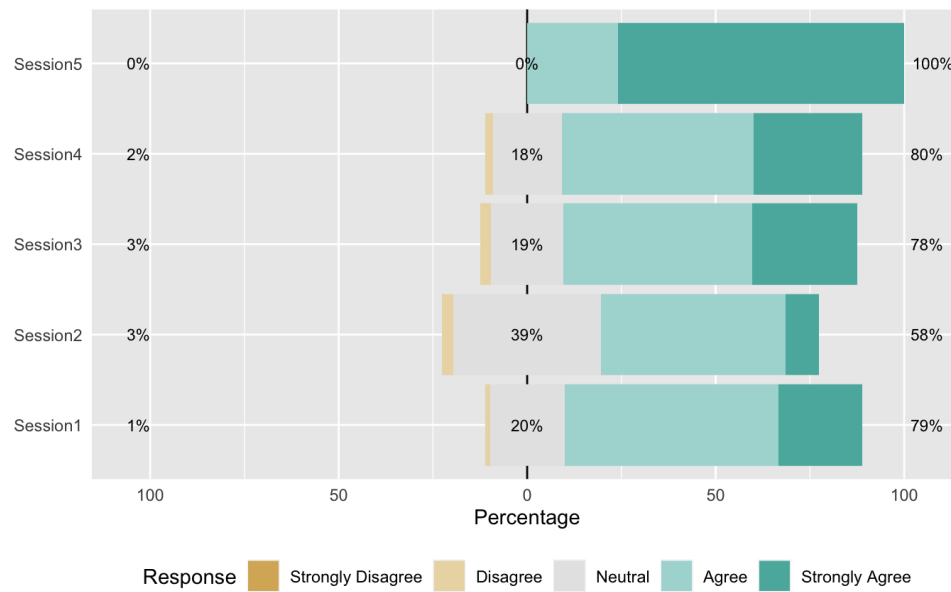
During the first session, EM was not used entirely as intended due to the aforementioned difficulty distinguishing rules from examples. However, despite the technique not being used entirely as intended, EM still had a positive result. As the participants had to think in a particular manner in order to write down rules and examples, they had to stay focused on what was important and write those things down. After the first session, they also noted that they think the technique, and TA sessions in general, offer a good structure and overview. To the participants, discussing the user story with only a few people rather than with the entire team felt like a very effective and efficient way to organise their refinements.

For the second session, a user story was refined that did not seem very suitable for EM. It mostly concerned the naming of a certain component. What we therefore did was discuss what the component was supposed to do in the form of rules and examples, from which names could be derived that represented these rules and examples. Once the team got to the point where they were discussing names, we agreed that it would be best to make the rest of the session a brainstorm about possible names, which could then be written down on example cards. This resulted in a session that still delivered a good overview of everything that was discussed as an output. The session was therefore still valuable to the team and served its purpose. However, it was likely that the TA session principles caused this rather than EM itself. The fact that EM did not match the user story completely is probably also why this session was rated lowest of all sessions, with only 58% positive responses.

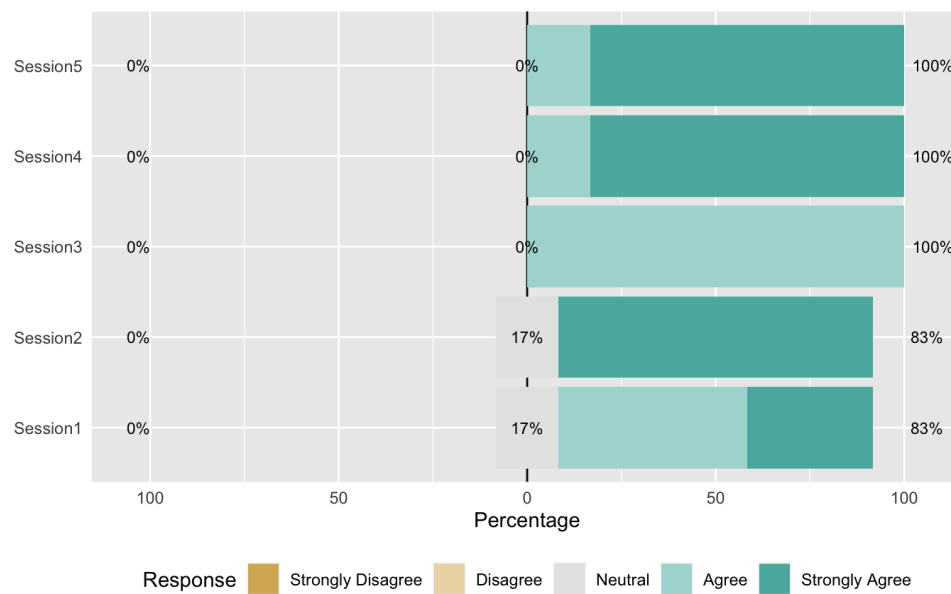
For the third EM session, the Scrum Master had already prepared some initial rules. These were discussed at the beginning of the session, after which new rules were added, following examples. This user story was about testing needs for the team's products. It was therefore not about any new functionality, but rather about testing requirements for future user stories. Despite this very different context, EM seemed very suitable for refining it. We could note, however, that the PowerPoint slide was rather full at the end of the session, which may negatively impact the overview of the rules and requirements.

During the fourth session, EM was again not used entirely as intended when considering the difference between rules and examples. However, the session still went well, and the fact that important things were written down on cards helped the team to still have an overview of the discussion afterwards. At the end of the session, the Product Owner also mentioned that she finds EM very useful for these kinds of user stories, which shows that they were not negatively affected by the fact that EM was not followed correctly: the team's own interpretation of the technique still helped them a lot. The Scrum Master also mentioned that the technique really helps for the team to not keep talking in circles but to be concrete and that it really helps to write down the questions that still need to be answered after which the team can move away from that question and focus on other things.

For the fifth session, the Scrum Master had again prepared some rules in advance, now along with some examples as well. During this session, the team discussed a lot of aspects of the user story. The preparation of the cards helped the team initiate the discussion in a certain direction, but they were not discussed in great detail. This session comes out very positively in Figure 5.9. However, it must be noted that this session misses most responses of all: of the five participants of the session, only one person filled in the questionnaire. Four from the six responses that are missing in the entire case study come from this session. Therefore, the 100% positive ratings are very subjective compared to the other session results, where ratings from several participants are combined.



**Figure. 5.9:** Results per Session (Self-reported Data)



**Figure. 5.10:** Results per Session (Observation Data)

### 5.3.1 Evaluation after session 4

This case study was originally supposed to be around 8-10 sessions. Therefore, we had agreed with the team to have two evaluations: one short evaluation after three or four sessions, and one larger evaluation after the last session. The first evaluation was planned to get some initial opinions about the technique as well as to see if anything could be done to improve the sessions. The evaluation was eventually held after the fourth session. However, the case study came to an unforeseen end after the fifth session due to external reasons. The evaluation after the fourth session was already close to the end of the case study now, so most insights were gained here rather than after the end evaluation

The evaluation was held with the Product Owner and the Scrum Master. As they had both been present for all sessions up until now, while other team members were not, they believed it was not useful for others to join this evaluation. More team members would be joining the end evaluation, which eventually did not happen anymore, unfortunately.

We first asked them about their general opinions of the technique and sessions. The Product Owner mentioned that the sessions have been valuable to them, and also that a big learning effect was observed: with the third session, the Product Owner found that it went a lot better than during the first two.

The Scrum Master agreed with this and added that EM helps the team to have a certain direction in which they need to think in order to write everything down properly, which helps with the conversation. It was also noted that EM provides a good overview of the requirements of a user story.

The Scrum Master also referred back to the first EM session. Since that session, that user story had been discussed with the entire team. He mentioned that the EM session had really helped them to have an overview from which they could determine the necessary approach and actions for that user story. The fact that the output was discussed with the entire team and was also valuable for those who were not present during the sessions indicates that EM also has positive long-term effects regarding the implementation of a user story. This indication is positive for RQ5, but we believe the result of a few user stories is not enough to provide a proper conclusion to the research question.

The participants were asked about their opinion of Microsoft PowerPoint as a tool. They had also noted that the overview had sometimes gotten a bit lost because the slide was full. Switching away from PowerPoint was discussed, but they still wanted to continue using it, as it prohibited scrolling down which could cause them to lose things out of sight. Despite not being the best overview, they deemed this would still be a better overview than when the team would scroll down and not see the first cards anymore.

What was also considered a very valuable aspect of this case study by the participants was the Three Amigo principle: only having one person from each perspective, rather than refining with the entire team. Overall, both the Product Owner and Scrum Master are positive about the technique. The Product Owner sees a lot of potential to keep using EM after the case study is over but did note that maybe not all of their user stories would be suitable for the technique. Especially the really technically-oriented user stories would be less suitable for EM. They will have to look at a user story and assess whether or not they think EM suits it before actually organising a session.

### 5.3.2 End Evaluation

An end evaluation was held with the same people as the first evaluation: the Product Owner and the Scrum Master of the team. This end evaluation was held approximately one month after the fifth and final EM session. Because of this, four out of five user stories that were refined using EM had been discussed with the entire team. From this, the entire team concluded that the EM outputs were valuable as they gave a structured and organised way of visualising the requirements.

The Scrum Master also praised the efficiency of the technique. In just half an hour, a user story would be refined, and a visualisation of the refinement is created as well in that same time. The Product Owner also mentioned again that the fact that only a portion of the team is present with the session (i.e., the TA-principles), that this really helps to stay efficient. EM also forced them to keep a certain structure to the meeting, which helped them to not get off track and discuss things that are not important for this user story.

The Product Owner and Scrum Master agreed that they want to keep EM as a part of their way of working after the case study. Even more so, they want to encourage other development teams in their department to also adapt the method in their ways of working. This indicates that the team is really positive about EM and that it has given them added benefits compared to their previous way of working.

### 5.3.3 Conclusion

During this case study, five EM sessions were held together with one evaluation after the fourth session. Despite the team having some initial challenges to use EM properly, they have gained much from the sessions. The TA principles helped to make the meetings effective and efficient. EM helped the team to aim the conversation in a certain direction and to provide a good overview of a user story's requirements afterwards.

During the evaluations and based on the questionnaire responses, we can observe that participants were generally positive about the case study. It was also mentioned that EM might not be suitable for all types of user stories. With the second session, we also concluded that EM was not really suitable. Nevertheless, good results were achieved due to the TA principles. With the four other sessions, EM did have an added benefit. From this, we can conclude that the TA aspects are very beneficial by themselves already, also when EM is not suitable, and also that EM is conditionally well-performing. It works very well as a technique to give structure to a refinement session, and the resulting output gives a good overview of the requirements of a user story. However, EM will not be suitable for all user stories. For the ones it is not suitable, we would recommend this team and other teams in similar contexts to still have TA sessions, but without EM.

## 5.4 Case Study - Pension Management Firm Feature Mapping

The third case study focused on Feature Mapping (FM) and was held at the same pension management firm as the second case study, but with another team. This team creates a portal that provides pension advisors and employers insights into the company's pension products. This case only consisted of two sessions: the first on June 3rd and the second on June 23rd.

The reason that this case study only had two sessions is that FM did not have a good fit within the context of this team. The team has three types of user stories, depending on the phase they are in, and we tried two types of user stories. First, we tried a user story for which some initial analysis had already been performed. This meant that they already knew what the user needs were and how they would be delivering the insights that that user story would provide. We believed that this would be the right type of user story, as they would know the direction in which the user story would go, but did not have specific acceptance criteria yet. As the team noted that they often missed certain criteria in their current way of working, we believed this would be the right type of user story to apply FM to.

However, the first mismatch between FM with this team was that FM requires participants to write down tasks. As the team's portal delivers insights to users, rather than action-driven functionality, they had great difficulty writing down tasks. We talked about this and came up with the idea to transform tasks into conditional statements regarding an employee's pension contract, which could be used to come up with examples that show how certain conditions alter the provided insights. At this point, another mismatch was observed: when we started with this, the team members explained that they had already analysed this in the previous phase and that using FM like this felt like they were performing double work. They did mention that they liked the structure that FM would provide for writing these things down, but this meant that FM would deliver minimal added value to their current way of working. Therefore, we concluded that this type of user story was unsuitable for FM.

For the second session, we chose a user story that was in an early phase. For this user story, the user needs still had to be come up with, the full analysis still had to be performed and little was known about the user stories. The team explained that during this phase, the implementation of a user story can still go in many different directions. These different directions could be analysed and visualised using FM, which is why we wanted to give this type of user story a try. However, during the session, we came to the conclusion that this type of user story was actually still so vague, that almost nothing could be written down in the format that FM provides. More analysis with people from outside the team was required in order to make any statements for this user story, which meant that they were again not able to use FM properly.

The only other type of user story that the team has is one that is in an even later phase than the first attempted user story, one that regards testing of the implemented functionality. We agreed beforehand that this type of user story would definitely not be suitable for FM, as no new features would be built for those user stories. Therefore, we came to the conclusion that FM was not a suitable method for this team and stopped the case study.

The team members did mention that they really liked the Three Amigo principles that come with FM, and that they want to keep organising sessions like these, but without FM. That way, they have the right people all together in one session, which they believe will be very effective and efficient.

With the two sessions, we received a total of five questionnaire responses. During both sessions, we used the long questionnaire. The aspect results of the case study can be observed in Figure 5.11. In this figure, the observation results are also included, which was done as there were not enough sessions to make a clear analysis of the different sessions. As expected, the figure shows overall negative results. Coordination and intention to use score lowest with only



8% and 10% positive ratings, respectively. The observations have some very positive ratings as well, which is because the team members did all try to work together and participate actively, despite not gaining a benefit from FM.

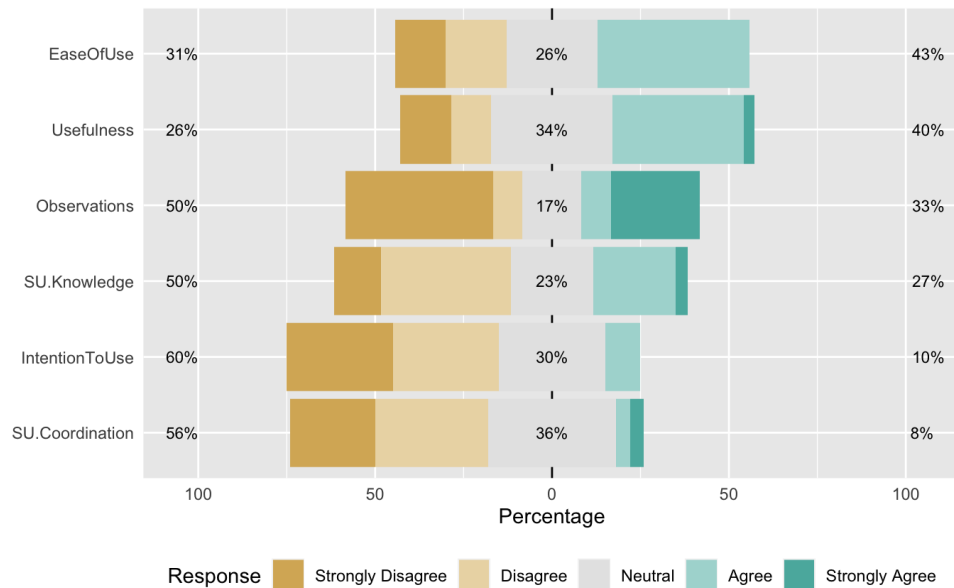


Figure. 5.11: Observation and Aspect Results

### 5.4.1 Conclusion

With this third case study, we tried to implement FM with a team at the pension management firm. To conclude, FM did not perform well in the context of this case study. We posit three possible reasons for this. Firstly, the team may already have a dedicated process in which FM does not fit. The way the team distinguishes different types of user stories, for example, did not match well with FM. One type of user story was still too vague, but the second type of user story was already analysed to the extent that FM felt redundant.

Secondly, FM may be too much action-driven, whereas this team does not develop functionality but rather insights. FM may not work well for an insight-driven platform, due to the focus on actions with the tasks. Thirdly, it may be that the team was not accepting enough of change in their current way of working. Despite participating in this research voluntarily, it is possible that the team is really proud of their current way of working and does not want to change anything about it. As they believe their current processes work well, they may not see a need to reconsider it.

The reason that this case study failed is probably due to a combination of all of the above three reasons. Perhaps Example Mapping would have been a fit for this team, considering it is not focused on user actions as much as FM is. However, there was unfortunately no time left for the research to continue the case study with EM. The team did still find the sessions valuable as they did gain some knowledge of different ways of working and also want to implement TA sessions in theirs, but they will be doing so without Feature Mapping.

## 5.5 Case Study - Example Mapping Online Tryout

The final case study that was performed for this research is a single-case study of EM. This case study was originally a try-out session that explored using EM with online tools. It was set up by a company that focuses on Agile coaching, facilitation and training. Through the social media platform LinkedIn, a call was made for professionals to attend a free online session in which they could explore EM using the online tool Mural. We have requested to be present in the session and to ask attendees to fill in a questionnaire after the session to get their opinion on the technique.

In total, eleven people participated in the session. First, the organiser explained Behaviour-Driven Design (BDD), followed by an explanation of how EM works. This explanation was different from how we have explained EM in the experiment and case studies, which may be a risk of how the participants implement the technique. In general, the explanation that was given during this session was much shorter and did not go into as much detail. One already-finished EM output example was shown to the participants and all concepts were shortly elaborated on. After the technique was explained, the case was presented. Participants had to refine a user story for a mobile scanner app for a supermarket. As this was a familiar context for everyone, we believed it to be a good case.

The eleven people were split up in two groups for the EM session. This means that the groups consisted of five and six people, respectively. In both sessions, a facilitator was present that would help the participants get started. The researcher could unfortunately only be present in one group. It was apparent that the participants did not know each other, which influenced the coordination of the session. For most, it was the first time using the online tool MURAL [42]. Some people were already familiar with EM, but some also were not. This resulted in some people wanting to get started right away, whereas others wanted to first discuss the tool and the case into more detail before actually writing something down. This affects the SU of the participants, as not everyone actively participated during the first part of the session with the refinement itself.

Of the eleven participants, six filled out the questionnaire. The participants were first asked some demographic questions regarding their experience regarding the same five concepts that were asked for the controlled experiment: user story refinement, working in an Agile software development environment, Gherkin, Example Mapping, and Feature Mapping. All of the six people have experience in both user story refinement and working in an Agile software development environment. Three people have prior experience working with Gherkin and EM, while the other three have little to no experience in this area. Two out of six have previous experience with Feature Mapping as well, while the other four have little experience with this. From these demographic results, we can conclude that the six people that filled in this questionnaire are professionals that have a lot of experience working in the field of Agile software development. Despite only having received six responses, this case complements the controlled experiment presented in Section 5.1 well, as those participants did not much prior experience in this field.

The results of the questionnaire are visualised in Figure 5.12. In this figure, we observe that intention to use and usefulness are rated very highly with 100% and 93% positive ratings, respectively. This shows that participants find the technique to be very suitable for online sessions and want to use it more in the future for user story refinement. One other explanation for the high intention to use, however, is that participants of this session joined voluntarily, with the intention to learn more about using the technique in an online environment. Ease of use is also scored well, with 76% of the questions answered positively and only 7% negatively. The SU aspects are rated positively as well, but not as well as the others: knowledge and

coordination score 64% and 57%, respectively. This is probably due to what was explained already regarding the fact that some people wanted to start right away whereas others did not. Also, the people present in this session all have different backgrounds. With the controlled experiment, students may also have paired up with other students that they did not know, but they still had a similar context: they were all students of roughly the same age and were all studying the Business Informatics master's programme. With the case studies, participants all worked together in the same team and already knew each other well. The people in this research all did not know each other as they work at different companies, where they have also established their own way of working. We believe this is why SU was rated much lower than the other aspects.

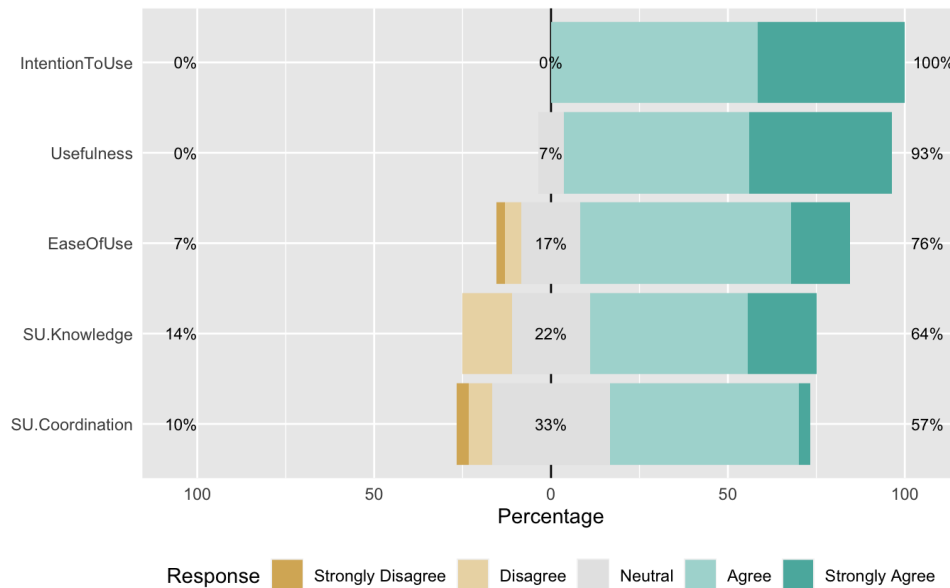


Figure 5.12: Aspect Results

### 5.5.1 Conclusion

With this case study, eleven people tried out EM with an online tool, of which six filled in the questionnaire afterwards. Overall, participants rated the technique quite highly and they all say that they intend to use the technique for future refinements. The context of this research did, however, negatively affect the SU of participants after the refinement session, which is why knowledge and coordination were not rated as positively as the other three aspects.

We believe that a more thorough explanation of EM and of the online tool would have improved the coordination of the session a lot. If an EM example was presented that went through how a session would look step by step, as done with the controlled experiment and the other case studies, participants would all be able to start with the refinement right away, rather than some participants still conceptualising the technique while others had already started. This would greatly improve the coordination aspect of SU, and likely the knowledge aspect as well since participants would be much more involved in the session. Therefore, we conclude that EM performed well within the context of this case study, but that it could have performed even better if there had been more initial agreement regarding EM through a more detailed explanation of the technique, given that our stated reason for the lower SU is correct.



# Chapter 6 | Conclusion

In Chapter 5, we have presented the findings of the experiment and case studies. As explained in Chapter 4, all cases first have to be analysed individually, after which cross-case conclusions can be made. In this chapter, we will first attempt to make those cross-case conclusions in order to generalise the findings. We then proceed to answer the underlying research questions, followed by the main research question.

## 6.1 Cross-case Conclusions

Throughout this research, five separate studies have been conducted: one controlled experiment and four case studies. With the controlled experiment, both Example Mapping (EM) and Feature Mapping (FM) performed well, especially on shared understanding (SU). Despite participants having little experience in user story refinement and working together in the teams that were formed, they still mostly managed to apply the techniques in the context of this experiment and form an opinion about them. FM was rated better than EM did on usefulness, whereas EM scored better on ease of use. Intention to use was relatively low with both EM and FM, probably because participants have too little experience to compare the techniques to other ways to refine user stories. EM outputs were, in general, more complete than FM, which may be because FM has a steeper learning curve.

With the longitudinal case study at Fizzor, we held eight EM sessions. During these eight sessions, we observed a clear learning effect that greatly improved the participation and coordination of the participants. Overall, EM performed well, especially with both SU aspects (knowledge and coordination). Intention to use was not rated well. We believe this is because the project has many very small user stories which made us have to make alterations to the sessions by grouping user stories together. This was a good solution. Another solution would have been to have EM sessions that are shorter, but this was not researched. The participants said EM would be useful for specific user stories that are vague, big or have different possible implementations, and that small user stories would not require EM sessions.

Two case studies were conducted at a pension management firm, one for EM and one for FM. The EM case study consisted of 5 sessions and, overall, the technique performed well. The participants themselves mentioned a learning effect was observed and that they noticed that the sessions went better from the third session. EM was not a great fit for one of the five sessions, but the TA principles still made it a well-performing refinement session. EM helps to give structure to a refinement session and to have a good overview afterwards of what the requirements of a user story are. This means that, just like with Fizzor, in this case study, there were user stories for which EM was suitable and those for which it was not. In this case, participants believed that the very technically-oriented user stories would not be suitable for EM when compared to stories that were focused on functionality that were suitable. As

mentioned in the end evaluation, the outputs of the EM sessions were also valuable to other team members that were not present during the session and the team will also continue to use EM after this case study and will even encourage other teams to adopt it.

The FM case study at the pension management firm did not perform well, unfortunately. We attribute this is due to three possible reasons: FM may not be a good fit in the way of working the team already has, FM may be too action-driven, while this team develops an insight-driven portal, and the team may not have been accepting enough of change. Despite FM not working, the team did say that TA sessions were considered very valuable and that they are going to implement that in their way of working.

Lastly, the EM online tryout was a small case with only six responses. EM was rated well by the participants, but SU scored lower than the other aspects. We believe this was due to the fact that a more elaborate explanation of the technique should have been provided beforehand so that people could have worked together better during the session.

Case	Technique	Ease of Use	Usefulness	Intention to Use	SU Knowledge	SU Coordination
Controlled Experiment	FM	69%	69%	50%	81%	69%
Controlled Experiment	EM	78%	71%	47%	84%	74%
Fizor	EM	53%	58%	12%	88%	82%
Pension Management Firm	EM	84%	86%	50%	86%	68%
Pension Management Firm	FM	43%	40%	10%	27%	8%
Online Try-out Session	EM	76%	93%	100%	64%	57%
Average	EM	73%	77%	52%	81%	70%
Average	FM	56%	55%	30%	54%	39%

**Table 6.1:** Positive Questionnaire Responses per Case

In Table 6.1, an overview is given with the ratings for all individual performance aspects. The percentage depicted here regards all positive replies to the questions in the questionnaire (i.e., when a question is answered “Agree” or “Strongly agree”). The average values are based on all cases having similar weight and must, therefore, be read with caution. For example, the online try-out session case weighed as heavily as the Fizor case study in the average, although the Fizor case study consisted of more total sessions. From this cross-case analysis, we can conclude that EM performed well in all cases. Therefore, we believe that we can generalise the findings for EM and conclude that it is a well-performing technique overall when implemented in teams that have a context similar to those in the different studies. However, we must note that there are some conditions as to the types of user stories for which the most benefits are gained. In the conducted case studies, stories that were very small and straightforward or stories that were very technically-oriented were deemed less suitable for EM. When EM is not suitable for refinement, we have still observed TA sessions in general to perform well. FM

performed well during the controlled experiment, but it did not perform well with the case study. Therefore, the performance of FM requires further investigation before any conclusions can be generalised. Both action-driven and insight-driven products should be studied in order to properly conclude which conditions make FM suitable or not. Even with the FM case study, however, we concluded that TA sessions in general would be a valuable addition to their current way of working.

## 6.2 Answers to Research Questions

This section serves to give an answer the research questions that were constructed in Section 2.1. We will discuss each research question (RQ) separately, followed by the main research question (MRQ). Unfortunately, we could not answer RQ5 (“Do TA sessions have effects on the implementation of a user story?”) due to the context of the case studies that were conducted, as the original case studies were cancelled because of the COVID-19 pandemic as explained in Chapter 5.

### 6.2.1 RQ1: What defines a Three Amigo session?

**RQ1.1:** What TA session techniques exist?

**RQ1.2:** Where do TA sessions fit in Requirements Engineering?

**RQ1.3:** Where do TA sessions fit in Software Engineering?

We performed a literature review to answer the first research question. In Section 3.3, we describe that a Three Amigo (TA) session is a short session in people from different disciplines refine an increment of work together, often in the form of examples. A TA session is claimed to increase the SU of the team. Example Mapping (EM) and Feature Mapping (FM) are two defined TA session techniques, of which EM is rather free-format compared to FM that has a more structured procedure. Both techniques are based on post-it cards of different colours to help provide a structure to the session and create an organised output of the session.

Considering Requirements Engineering (RE), we have categorised TA sessions as part of three out of four RE processes: domain understanding and elicitation, evaluation and negotiation, and specification and documentation. Looking at Software Engineering (SE), we have elaborated on the Behaviour-Driven Development (BDD) method, in which six different phases are defined. TA sessions fit best in the second and third phase, namely defining features and creating examples for those features.

### 6.2.2 RQ2: How can the performance of TA sessions be measured?

**RQ2.1:** How can shared understanding be measured?

**RQ2.2:** How can the performance of a user story refinement technique be measured?

For constructing a performance measurement tool for TA sessions, we researched shared understanding (SU), as a good SU was claimed to be a key benefit of organising TA sessions. In Section 3.4, we have defined shared understanding as the implicit and explicit knowledge that is shared amongst team members both as a structure and as a process. Besides that, at least two different types of teams exist: the development team as a whole and the team that performs the TA session. In order to measure SU, we have created a questionnaire that measures the knowledge and coordination aspects of SU based on self-reported perception, as explained in Section 4.3. These questions are added to the existing Method Evaluation Model by

Moody [41], which together form the core of how to measure the performance of a user story refinement technique. However, additional research is required to find the relation between the aspects of the Method Evaluation Model and SU, as no conclusive correlations were observed during this study. On the other hand, we find that the adapted model is partly validated, as the self-reported SU seemed to be aligned with what we have observed during the case studies. That is also part of the validation, as the individual scores regarding SU are in line with researcher observations. However, finding the exact relations between these aspects in terms of refinement techniques will help validate the adapted model even more.

Besides the questionnaire, we have also performed an output analysis for the controlled experiment. For all case studies, a researcher was present to observe the TA sessions, who answered seven questions regarding coordination of the session. Having had someone observe the refinement session without being involved proves to give valuable additional insights next to the self-reported data.

### 6.2.3 RQ3: How do TA sessions perform when used for the first time?

**RQ3.1:** How does a first Example Mapping session perform?

**RQ3.2:** How does a first Feature Mapping session perform?

We were able to answer RQ3 by the controlled experiment, and the first sessions of all case studies. A controlled experiment was conducted with the students of the Requirements Engineering course at Utrecht University. As explained in Section 5.1, the controlled experiment showed good results for both EM and FM. Both had some different areas where they were rated better than the other: EM was rated higher on ease of use, whereas FM was rated higher on usefulness. SU was rated highly with both, which indicated that the techniques indeed deliver a good SU among team members. Intention to use scored lower for both techniques, which we believe is because participants did not have enough knowledge of other techniques with which they could compare these two. Overall, we conclude that both EM and FM perform well when used for the first time in a controlled environment.

With the case studies, decent results were observed after the initial session with EM. With Fizzor, results were not great after the first EM session due to a lack of participation from participants, but the resulting output was still deemed valuable by the participants. With the pension management firm EM case study, the first session was a lot better, and during the evaluation it was also mentioned that the session output really helped the team move forward with the user story. Lastly, the online try-out session of EM showed good results as well. The SU amongst team members was not as high as with the other cases, but possible solutions to solve this in the future are given to make sure the first session of a team will perform better on those aspects as well.

Summing up all these findings, we can conclude that EM works reasonably well when used for the first time, but that several factors can have a negative influence on initial performance, such as bad participation of team members, and that a proper explanation of the technique is required to gain the most benefits from an initial session. Without active participation or a proper explanation, SU is negatively influenced. These factors are probably applicable to other refinement techniques as well and not just limited to EM.

Unfortunately, the FM case study was not successful. The technique did not fit well within the context of the team, which can be due to several (or a combination of) reasons. We believe that more research into FM must be conducted to be able to provide a substantiated answer to RQ3.2, as the controlled experiment and this case study contradict one another regarding the performance of FM when using it for the first time.



#### 6.2.4 RQ4: How do TA sessions perform after becoming familiar with the technique?

**RQ4.1:** How does Example Mapping perform after becoming familiar with the technique?

**RQ4.2:** How does Feature Mapping perform after becoming familiar with the technique?

For answering RQ4, we have conducted three separate case studies: two in which EM was evaluated and one in which FM was evaluated. As the controlled experiment and the online try-out session only consisted of one TA session (the experiment had two, but only one per technique), these are not considered for RQ4. As we have elaborated in Section 6.1, EM performed well in all cases. There are some conditions that make EM less suitable, such as too technically-oriented user stories or user stories that are very small and straightforward. We do expect that the exact conditions of when a user story will be suitable or not suitable also depends on the context of the team and the team members themselves. Therefore, we withdraw ourselves from more concretely specifying when a user story would be suitable or not. In the cases that EM is not suitable, TA sessions without EM still seem beneficial. Overall, we conclude that EM appears to perform well in a real RE environment after becoming familiar with the technique.

Unfortunately, we do not have substantial results for FM. Only two sessions were organised during the FM case study, which could both be considered an initial session because the context and goal were quite different for both. Therefore, future research will be needed to answer RQ4.2

#### 6.2.5 MRQ: How do defined Three Amigo session techniques perform for user story refinement?

The goal of this research was to evaluate the performance of TA session techniques for user story refinement. Through RQ1 and RQ2, we laid a theoretical foundation to the research by investigating TA sessions and SU into detail and by presenting a performance measurement tool for refinement techniques. With the controlled experiment and four case studies, we investigated the performance of the techniques in both a controlled and real RE environment in order to answer RQ3 and RQ4. Unfortunately, the long-term effects of TA session techniques could not be properly researched during the case studies, which prevented us from answering RQ5. Despite not being able to answer RQ5, we conclude based on this research that TA sessions perform well for user story refinement.



# Chapter 7 | Discussion

In this chapter, we discuss the research and its implications, we state what the limitations of the research are, and we propose ideas for future work. The aim of this research was to evaluate Three Amigo (TA) session techniques on their performance. Requirements Engineering (RE) is considered an integral part of software development with a significant impact on the total overall performance, yet refinement techniques are barely researched on their performance. We have developed a performance measurement tool for refinement techniques with an extra emphasis on shared understanding (SU) and have also conducted case studies to evaluate the performance of two techniques, Example Mapping (EM) and Feature Mapping (FM). By doing so, we have created a valuable contribution to the field of RE with this research. The research also has practical relevance, as we provide insights into the performance of refinement techniques (especially EM) and state several conditions when techniques may be more or less suitable. We have found TA sessions result in a good SU amongst team members, which previous research has shown to improve overall team performance. As such, this research can act as a guidance for teams that are considering different refinement techniques.

## 7.1 Limitations

The biggest limitation of this research is that the longitudinal case studies did not allow us to answer RQ5 as the case studies did not allow us to fully investigate how TA sessions affect long-term aspects. Originally, the plan was to have case studies that would last around three months, which would allow us to answer RQ5, but the COVID-19 pandemic disabled us to execute these initial plans.

Secondly, a limitation is the number of cases that were researched, especially for FM. As this was a rather qualitative research, results are very context-specific, which makes generalising them difficult. We tried to mitigate this threat by having two data sources for each case: the questionnaires and outputs for the controlled experiment, and the questionnaires and observations for the case studies. We believe we can generalise the findings of EM to teams that are in a similar context as the ones we have researched, considering we have had three case studies and the controlled experiment which all showed good results, but additional case studies would help to support this. For FM, additional research is needed. As we only have results from the experiment and from one case study, which are contradictory, there is a need for more research before anything can be concluded about the performance of this technique.

Another limitation is that the COVID-19 pandemic had a significant impact on a team's way of working. As everyone had to suddenly work from home, this itself may have already been a big adjustment for them. A part of this research may have actually been people trying out TA sessions in online tools, rather than researching the techniques themselves. This is because it was new for many people to switch to an all-online work environment. As such, we have likely

also assessed online platforms and refining together remotely, rather than just the TA session techniques. This is a risk to the validity of the research. We think the severity of this is low, as teams already had one or two months to adjust to the new situation before the case studies actually started. Participants had mentioned that they were already used to the new situation for a large part. However, additional research in the future would help to support our findings. This additional research could also be from online sessions, or from co-located sessions, where participants are in the same room together when using the TA session techniques.

## 7.2 Future Work

First and foremost, future research can be conducted to mitigate or even remove the limitations of this research. Having other cases by itself already helps greatly to support generalising the findings. By conducting more case studies, additional insights can be obtained as to which contexts enable or disable TA session techniques to perform well. Also, a longitudinal case study could be conducted in which the refined user stories are also actually implemented during the course of the case study so that RQ5 can be answered. By looking at the entire software development cycle, insights could be found about TA session techniques that we were unable to find.

The third limitation that we stated was that we conducted our research with online sessions due to the COVID-19 pandemic. As stated in the previous section, repeating this research online in the future would already help support the findings, as teams would then already be much more familiar with an online work environment. Besides that, it will also be really valuable to investigate the performance of EM and FM with team members that are co-located rather than working remotely.

Besides overcoming these limitations, future research could also compare TA session techniques to other refinement techniques, such as brainstorming or interviewing. Teams in this case study have said that they really valued the TA session principles because this keeps the group small rather than defining with the entire team. Using a well-defined refinement technique with the whole team, for example, will likely also have its benefits. Therefore, it will be valuable to expand the number of techniques that are evaluated on their performance.

Additionally, research could be conducted to investigate the performance measurement tool in more detail. The questionnaire is partially validated in combination with the observations, but additional research will be valuable to validate or improve the tool. This research can also validate how well the tool adapts to other refinement techniques, as the current tool has a focus on the combination of rules, examples and questions as used in EM and FM.

Finally, during the Fizzor case study, team members found that very small or straightforward user stories would not be worth refining using a 30-minute EM session. Therefore, future researchers could investigate the performance of TA session techniques with an alternating duration per session. For example, depending on the size of the user story, teams could also opt for sessions of 10, 15 or 20 minutes. That way, not much time is spent on these user stories, but a structured output with all the requirements is still produced. If future research shows that short EM sessions are still well-performing, then that would be a fitting solution for these types of short or straightforward user stories.

## Chapter 8 | Bibliography

- [1] Pekka Abrahamsson, Kieran Conboy, and Xiaofeng Wang. 'lots done, more to do': the current state of agile systems development research, 2009.
- [2] Atif Açıkgoz, Ayşe Günsel, Nizamettin Bayyurt, and Cemil Kuzey. Team climate, team cognition, team intuition, and software quality: The moderating role of project complexity. *Group Decision and Negotiation*, 23(5):1145–1176, 2014.
- [3] Gojko Adzic. *Specification by example: how successful teams deliver the right software*. Manning Publications Co., 2011.
- [4] Gojko Adzic and David Evans. *Fifty Quick Ideas to Improve Your User Stories*. Neuri Consulting LLP, 2014.
- [5] Agile Alliance. Three amigos, Sep 2019. URL <https://www.agilealliance.org/glossary/three-amigos/>. Last accessed February 4, 2020.
- [6] Kent Beck. *Test-driven development: by example*. Addison-Wesley Professional, 2003.
- [7] Kent Beck, Mike Beedle, Arie Van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, et al. The agile manifesto, 2001.
- [8] Christoph Becker, Stefanie Betz, Ruzanna Chitchyan, Leticia Duboc, Steve M Easterbrook, Birgit Penzenstadler, Norbet Seyff, and Colin C Venters. Requirements: The key to sustainability. *IEEE Software*, 33(1):56–65, 2015.
- [9] Eva Alice Christiane Bittner and Jan Marco Leimeister. Why shared understanding matters—engineering a collaboration process for shared understanding to improve collaboration effectiveness in heterogeneous teams. In *2013 46th Hawaii International Conference on System Sciences*, pages 106–114. IEEE, 2013.
- [10] Sjaak Brinkkemper. Method engineering: engineering of information systems development methods and tools. *Information and software technology*, 38(4):275–280, 1996.
- [11] Leydi Caballero, Ana M Moreno, and Ahmed Seffah. Persona as a tool to involving human in agile methods: contributions from hci and marketing. In *International Conference on Human-Centred Software Engineering*, pages 283–290. Springer, 2014.
- [12] JA Cannon-Bowers, E Salas, and SA Converse. Cognitive psychology and team training: Shared mental models in complex systems. In *5th Annual Conference of the Society for Industrial and Organizational Psychology, Miami, florida*, 1990.

- [13] Janis A Cannon-Bowers and Eduardo Salas. Reflections on shared cognition. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 22(2):195–202, 2001.
- [14] Nancy J Cooke, Jamie C Gorman, Christopher W Myers, and Jasmine L Duran. Interactive team cognition. *Cognitive science*, 37(2):255–285, 2013.
- [15] Cucumber. Meet the cucumber open team, Dec 2019. URL <https://cucumber.io/tools/cucumber-open/team/>. Last accessed April 14, 2020.
- [16] George Dinwiddie. Effective software development, Jun 2009. URL <http://blog.gdinwiddie.com/2009/06/17/if-you-dont-automate-acceptance-tests/>. Last accessed March 30, 2020.
- [17] Eric Evans. *Domain-driven design: tackling complexity in the heart of software*. Addison-Wesley Professional, 2004.
- [18] Davide Falessi, Michele A Shaw, Forrest Shull, Kathleen Mullen, and Mark Stein Keymind. Practical considerations, challenges, and requirements of tool-support for managing technical debt. In *2013 4th International Workshop on Managing Technical Debt (MTD)*, pages 16–19. IEEE, 2013.
- [19] D Méndez Fernández, Stefan Wagner, Marcos Kalinowski, Michael Felderer, Priscilla Mafra, Antonio Vetrò, Tayana Conte, M-T Christiansson, Desmond Greer, Casper Lassenius, et al. Naming the pain in requirements engineering. *Empirical software engineering*, 22(5):2298–2338, 2017.
- [20] Markus Gärtner. *ATDD by example: a practical guide to acceptance test-driven development*. Addison-Wesley, 2012.
- [21] Josette MP Gevers, Christel G Rutte, and Wendelien Van Eerde. Meeting deadlines in work groups: Implicit and explicit mechanisms. *Applied psychology*, 55(1):52–72, 2006.
- [22] Martin Glinz and Samuel A Fricker. On shared understanding in software engineering: an essay. *Computer Science-Research and Development*, 30(3-4):363–376, 2015.
- [23] Jamie C Gorman and Nancy J Cooke. Changes in team cognition after a retention interval: the benefits of mixing it up. *Journal of Experimental Psychology: Applied*, 17(4):303, 2011.
- [24] Marc Hesenius, Tobias Griebe, and Volker Gruhn. Towards a behavior-oriented specification and testing language for multimodal applications. In *Proceedings of the 2014 ACM SIGCHI symposium on Engineering interactive computing systems*, pages 117–122. ACM, 2014.
- [25] Rashina Hoda, Norsaremah Salleh, and John Grundy. The rise and evolution of agile software development. *IEEE software*, 35(5):58–63, 2018.
- [26] Axel Hoffmann, Eva Alice Christiane Bittner, and Jan Marco Leimeister. The emergence of mutual and shared understanding in the system development process. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*, pages 174–189. Springer, 2013.
- [27] Hubert F Hofmann and Franz Lehner. Requirements engineering as a success factor in software projects. *IEEE software*, (4):58–66, 2001.

- [28] Paul J Hu, Patrick YK Chau, Olivia R Liu Sheng, and Kar Yan Tam. Examining the technology acceptance model using physician acceptance of telemedicine technology. *Journal of management information systems*, 16(2):91–112, 1999.
- [29] Irum Inayat, Siti Salwah Salim, Sabrina Marczak, Maya Daneva, and Shahaboddin Shamshirband. A systematic literature review on agile requirements engineering practices and challenges. *Computers in human behavior*, 51:915–929, 2015.
- [30] Ben Islin. Maximising the outcome of the three amigos agile workshop, Jul 2017. URL <https://inviqa.com/blog/maximising-outcome-three-amigos-agile-workshop>. Last accessed February 6, 2020.
- [31] Mayumi Itakura Kamata and Tetsuo Tamai. How does requirements quality relate to project success or failure? In *15th IEEE International Requirements Engineering Conference (RE 2007)*, pages 69–78. IEEE, 2007.
- [32] Franz W Kellermanns, Jorge Walter, Christoph Lechner, and Steven W Floyd. The lack of consensus about strategic consensus: Advancing theory and research. *Journal of Management*, 31(5):719–737, 2005.
- [33] Richard Klimoski and Susan Mohammed. Team mental model: Construct or metaphor? *Journal of management*, 20(2):403–437, 1994.
- [34] Mathias Landhäußer and Adrian Genaid. Connecting user stories and code for test development. In *Proceedings of the Third International Workshop on Recommendation Systems for Software Engineering*, pages 33–37. IEEE Press, 2012.
- [35] Kyle Lewis. Measuring transactive memory systems in the field: scale development and validation. *Journal of applied psychology*, 88(4):587, 2003.
- [36] Beng-Chong Lim and Katherine J Klein. Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 27(4):403–418, 2006.
- [37] Garm Lucassen, Fabiano Dalpiaz, Jan Martijn EM van der Werf, and Sjaak Brinkkemper. The use and effectiveness of user stories in practice. In *International working conference on requirements engineering: Foundation for software quality*, pages 205–222. Springer, 2016.
- [38] Grigori Melnik, Frank Maurer, and Mike Chiasson. Executable acceptance tests for communicating business requirements: customer perspective. In *Agile Conference, 2006*, pages 12–pp. IEEE, 2006.
- [39] Jonathan Mezach. Augurk, living documentation on your own terms. URL <https://augurk.github.io>. Last accessed February 13, 2020.
- [40] Susan Mohammed, Richard Klimoski, and Joan R Rentsch. The measurement of team mental models: We have no shared schema. *Organizational Research Methods*, 3(2):123–165, 2000.
- [41] Daniel L Moody. The method evaluation model: a theoretical model for validating information systems design methods. *ECIS 2003 proceedings*, page 79, 2003.
- [42] MURAL. Mural, 2020. URL <https://www.mural.co/>. Last accessed July 28, 2020.

- [43] K. Nicieja. *Writing Great Specifications: Using Specification by Example and Gherkin*. Manning Publications, 2017. ISBN 9781617294105.
- [44] Dan North. Introducing bdd, mar 2006. URL <https://dannorth.net/introducing-bdd/>. Last accessed February 4, 2020.
- [45] Bashar Nuseibeh and Steve Easterbrook. Requirements engineering: a roadmap. In *Proceedings of the Conference on the Future of Software Engineering*, pages 35–46, 2000.
- [46] Gabriel Oliveira and Sabrina Marczak. On the understanding of bdd scenarios’ quality: Preliminary practitioners’ opinions. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*, pages 290–296. Springer, 2018.
- [47] Jose Ignacio Panach, Sergio España, Oscar Dieste, Oscar Pastor, and Natalia Juristo. In search of evidence for model-driven development claims: An experiment on quality, effort, productivity and satisfaction. *Information and software technology*, 62:164–186, 2015.
- [48] Shelly Park and Frank Maurer. Communicating domain knowledge in executable acceptance test driven development. In *International Conference on Agile Processes and Extreme Programming in Software Engineering*, pages 23–32. Springer, 2009.
- [49] Yuqing Ren and Linda Argote. Transactive memory systems 1985–2010: An integrative framework of key dimensions, antecedents, and consequences. *Academy of Management Annals*, 5(1):189–229, 2011.
- [50] Nayan B Ruparelia. Software development lifecycle models. *ACM SIGSOFT Software Engineering Notes*, 35(3):8–13, 2010.
- [51] Eduardo Salas, Stephen M Fiore, and Michael P Letsky. Theoretical underpinning of interactive team cognition. In *Theories of team cognition: Cross-disciplinary perspectives*, volume 49, pages 187–207. Routledge, 2013.
- [52] Christoph Schmidt, Thomas Kude, Armin Heinzl, and Sunil Mithas. How agile practices influence the performance of software development teams: The role of shared mental models and backup. *International Conference on Information Systems*, 35, 2014.
- [53] Eva-Maria Schön, Jörg Thomaschewski, and María José Escalona. Agile requirements engineering: A systematic literature review. *Computer Standards & Interfaces*, 49:79–91, 2017.
- [54] Ken Schwaber. *Agile project management with Scrum*. Microsoft press, 2004.
- [55] Ken Schwaber and Jeff Sutherland. The scrum guide™, Nov 2017. URL <https://www.scrumguides.org/scrum-guide.html>. Last accessed March 27, 2020.
- [56] Mali Senapathi and Ananth Srinivasan. An empirical investigation of the factors affecting agile usage. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pages 1–10, 2014.
- [57] John Ferguson Smart. *BDD in Action: Behavior-driven development for the whole software lifecycle*. Manning, 2015.
- [58] John Ferguson Smart. Feature mapping – a simpler path from stories to executable acceptance criteria, Jan 2017. URL <https://johnfergusonsmart.com/feature-mapping-a-simpler-path-from-stories-to-executable-acceptance-criteria/>. Last accessed February 10, 2020.



- [59] Kimberly A Smith-Jentsch, John E Mathieu, and Kurt Kraiger. Investigating linear and interactive effects of shared mental models on safety and efficiency in a field setting. *Journal of applied psychology*, 90(3):523, 2005.
- [60] Carlos Solis and Xiaofeng Wang. A study of the characteristics of behaviour driven development. In *2011 37th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA)*, pages 383–387. IEEE, 2011.
- [61] Randy Stoecker. Evaluating and rethinking the case study. *The sociological review*, 39(1): 88–112, 1991.
- [62] Peter W Szabo. *User experience mapping*. Packt Publishing Ltd, 2017.
- [63] Steve Tooke. Your first example mapping session, May 2018. URL <https://medium.com/@tookky/your-first-example-mapping-session-a1800bf15cef>. Last accessed February 5, 2020.
- [64] Wolfgang Trumler and Frances Paulisch. How “specification by example” and test-driven development help to avoid technical debt. In *2016 IEEE 8th International Workshop on Managing Technical Debt (MTD)*, pages 1–8. IEEE, 2016.
- [65] Burak Turhan, Lucas Layman, Madeline Diep, Hakan Erdogmus, and Forrest Shull. How effective is test-driven development. *Making Software: What Really Works, and Why We Believe It*, pages 207–217, 2010.
- [66] Inge van de Weerd and Sjaak Brinkkemper. Meta-modeling for situational analysis and design methods. In *Handbook of research on modern systems analysis and design technologies and applications*, pages 35–54. IGI Global, 2009.
- [67] Piet Van den Bossche, Wim Gijselaers, Mien Segers, Geert Woltjer, and Paul Kirschner. Team learning: building shared mental models. *Instructional Science*, 39(3):283–301, 2011.
- [68] Axel Van Lamsweerde. *Requirements engineering: From system goals to UML models to software*, volume 10. Chichester, UK: John Wiley & Sons, 2009.
- [69] CollabNet VersionOne. 13th annual state of agile report, May 2019. URL <https://www.stateofagile.com/#ufh-i-521251909-13th-annual-state-of-agile-report/473508>. Last accessed March 27, 2020.
- [70] Jessica L Wildman, Eduardo Salas, and Charles PR Scott. Measuring cognition in teams: A cross-domain review. *Human factors*, 56(5):911–941, 2014.
- [71] Claes Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th international conference on evaluation and assessment in software engineering*, pages 1–10, 2014.
- [72] Matt Wynne. Introducing example mapping, Dec 2015. URL <https://cucumber.io/blog/bdd/example-mapping-introduction/>. Last accessed February 9, 2020.
- [73] Matt Wynne. Introducing example mapping, Feb 2018. URL [https://youtu.be/VwvrGfWmG\\_U](https://youtu.be/VwvrGfWmG_U). Last accessed January 11, 2020.
- [74] Matt Wynne and Aslak Helleoy. *The cucumber book: behaviour-driven development for testers and developers*. Pragmatic Bookshelf, 2012.

- [75] Matt Wynne, Aslak Hellesoy, and Steve Tooke. *The cucumber book: behaviour-driven development for testers and developers*. Pragmatic Bookshelf, 2017.
- [76] Robert K Yin. *Case study research and applications: Design and methods*. Sage publications, 2017.
- [77] Xiaodan Yu and Stacie Petter. Understanding agile software development practices using shared mental models theory. *Information and Software Technology*, 56(8):911–921, 2014.

# Chapter A | Questionnaire

## Appendix A.1 Complete Technique Questionnaire

### Perceived ease of use:

- Q1. I found the technique complex and difficult to use. (reversed)
- Q4. I believe that this technique allows me to express acceptance criteria with little effort.
- Q8. Overall, I found the technique difficult to learn. (reversed)
- Q10. I found the technique easy to learn.
- Q17. I found it difficult to apply the technique in the tasks of the experiment. (reversed)
- Q20. I found the rules of the technique clear and easy to understand.
- Q24. I am not confident that I am now competent to apply this technique in practice. (reversed)

### Perceived usefulness:

- Q7. Acceptance criteria represented using this technique would be easy for participants to understand.
- Q9. This technique would make it easy for participants to verify whether acceptance criteria are correct.
- Q14. Overall, I found the technique to be useful.
- Q16. Using this technique would make it more difficult to maintain the acceptance criteria. (reversed)
- Q22. Overall, I think this technique does not provide an effective solution to the problem of representing acceptance criteria. (reversed)
- Q23. Using this technique would make it easy to communicate acceptance criteria with other stakeholders of a project.
- Q25. Overall, I think this method is an improvement over other user story refinement techniques.

### Intention to use:

- Q18. I would use this technique in the future if I need to express acceptance criteria unambiguously.
- Q27. I would rather use a different way of expressing acceptance criteria if I ever need to define them in the future. (reversed)

### Shared understanding - coordination:

- Q5. We worked together in a well-coordinated fashion.
- Q12. We had very few misunderstandings about what to do.
- Q13. We needed to backtrack and start over a lot. (reversed)
- Q19. We accomplished the task smoothly and efficiently.
- Q26. There was much confusion about how we would accomplish the task. (reversed)

**Shared understanding - shared knowledge:**

- Q2. We have a good understanding of all rules of this user story that were discussed during the session.
- Q3. We have a good understanding of all examples of this user story that were discussed during the session.
- Q11. We understand what questions still need to be answered before we can proceed with implementing this user story.
- Q21. We have agreement on the overall output of this session.
- Q6. We could not reach agreement on certain rules. (reversed)
- Q15. We did not have a good understanding of the examples of this user story by the end of the session. (reversed)

## Appendix A.2 TA Session Questionnaire

**Perceived ease of use:**

- Q1. This session has allowed me to express acceptance criteria with little effort.
- Q7. I found the rules of the technique clear and easy to understand during this session.

**Perceived usefulness:**

- Q3. This session makes it easy to verify whether acceptance criteria are correct.
- Q5. This session makes it easy to communicate acceptance criteria with other stakeholders of a project.

**Shared understanding - coordination:**

- Q4. We had very few misunderstandings about what to do.
- Q8. The session went smoothly and efficiently.

**Shared understanding - shared knowledge:**

- Q2. We have a good understanding of all examples of this user story that were discussed during the session.
- Q6. We have agreement on the overall output of this session.

## Appendix A.3 Post-Implementation Questionnaire

### Usefulness:

- Q1. Acceptance criteria of this user story were easy to understand.
- Q4. The examples helped to verify whether or not acceptance criteria were implemented correctly.
- Q2. The acceptance criteria and examples were correct.
- Q5. No acceptance criteria or examples were missing.

### Shared understanding - shared knowledge:

- Q3. The examples helped to get an understanding on how the user story had to be implemented.
- Q6. There were few disagreements about the acceptance criteria of the user story.



# Chapter B | Case Study Forms

## Appendix B.1 Informed Consent

### Performance of Three Amigo Session Techniques

#### *Information sheet (10-03-2020)*

Thank you for considering to participate in this study. This information sheet outlines the purpose of the study and provides a description of your involvement and rights as a participant, if you agree to take part.

#### **Purpose of the research**

The purpose of this research is to investigate the performance of Three Amigo session techniques for user story refinement.

#### **Participation**

You are free to decide whether you want to participate or not. Your participation is entirely voluntary, meaning that you will not directly benefit from your participation. However, the results may be shared with you afterwards, if you would be interested in this. The researchers involved in this project do not foresee that there are any risks associated with your participation. If you do decide to participate, you are asked to sign a consent form which you will sign and return prior to the experiment.

#### **Withdrawal procedure**

You can withdraw from the study at any point whilst it is ongoing, without providing any reason for this. Withdrawing from the study will have no effect on you. If you withdraw from the study, we will not retain the information you have given thus far, unless you give us permission to retain the information.

#### **Usage of the data**

Corporate and personal information collected during the research will be anonymised. The anonymised collected information will be analysed for research studies by the researcher and may be used in publications and other research outputs.

Personal data will be processed on the legal basis of consent and is gathered for obtaining informed consent and for contacting you. Any personal information you communicate will be treated confidentially. All (personal) data will be treated as confidential as possible. Only the researchers involved in the project will have access to the personal data. Your data will be anonymised in research outputs, meaning that your name will not be used and any details which may reveal your identity will be disguised in any report or publication resulting from the study. You have the right to request access to and rectification or erasure of personal data while it is in storage. All personally identifying information collected will be destroyed once it is no longer needed for the project. Data will be kept in locked files that only the researchers involved in this study can open.

#### **Contact details**

If you have any questions or complaints regarding this study please contact dr. Fabiano Dalpiaz: [f.dalpiaz@uu.nl](mailto:f.dalpiaz@uu.nl).

## Performance of Three Amigo Session Techniques

### *Informed consent form*

Research Investigator: Fabiano Dalpiaz  
Institution name: Utrecht University

**By signing this document, you agree to all the following statements**

#### **Taking Part**

I have read and understood the attached project information sheet, dated 10-03-2020, or it has been read to me.

I have been given the opportunity to ask questions about the project and my questions have been answered to my satisfaction.

I agree to take part in the project. I understand that taking part in the project will include sharing materials (outputs of Three Amigo sessions) with the researcher and his research team.

I consent that my participation is voluntary and I understand that I can withdraw from the study at any time whilst it is ongoing without having to give any reasons for why I no longer want to take part.

#### **Use of the information I provide for this project**

I understand that information I provide will be used for the thesis report, publications and potential other research outputs.

I understand my personal data such as my name will not be revealed to people outside the project.

I understand that in any report on the results of this research my identity will remain confidential.

I agree that my (anonymised) words may be quoted in research outputs.

\*Insert name here\*

Name of participant

Signature

\*Insert date here\*

Date

Jasper Berends

Researcher [printed]

Signature

Date

For information please contact:  
Fabiano Dalpiaz ([f.dalpiaz@uu.nl](mailto:f.dalpiaz@uu.nl))



# Chapter C | Controlled Experiment

## Appendix C.1 User Story Outputs – Example Mapping

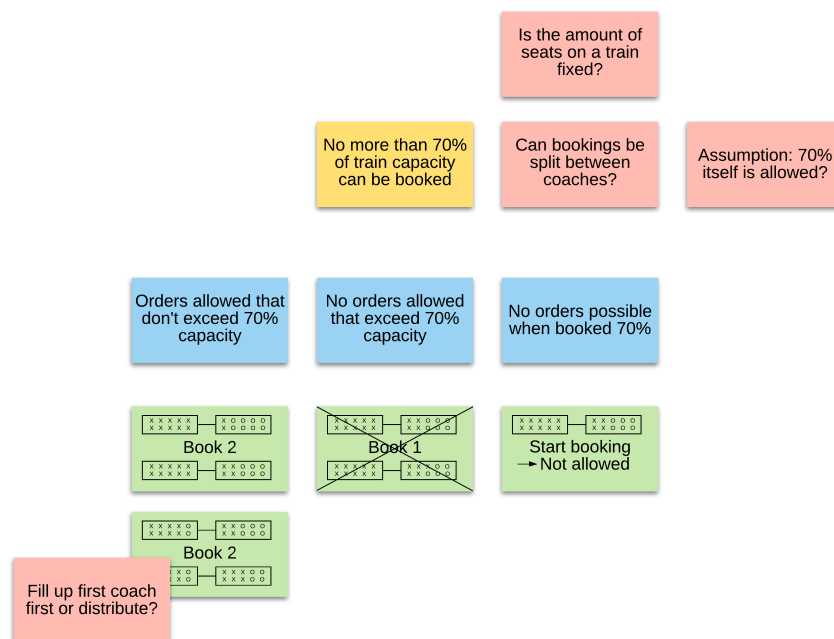


Figure. C.1: Example Mapping – US1

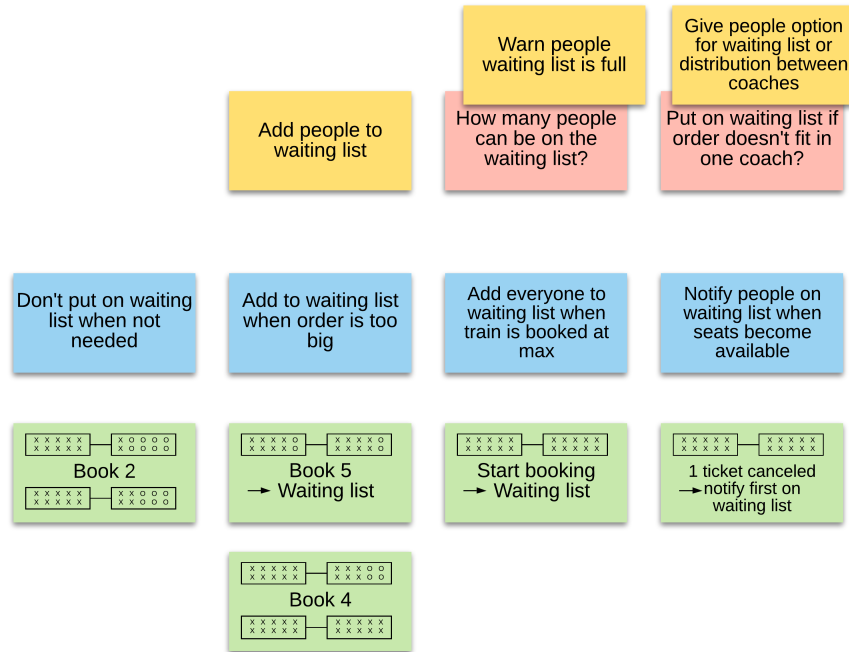


Figure. C.2: Example Mapping – US2

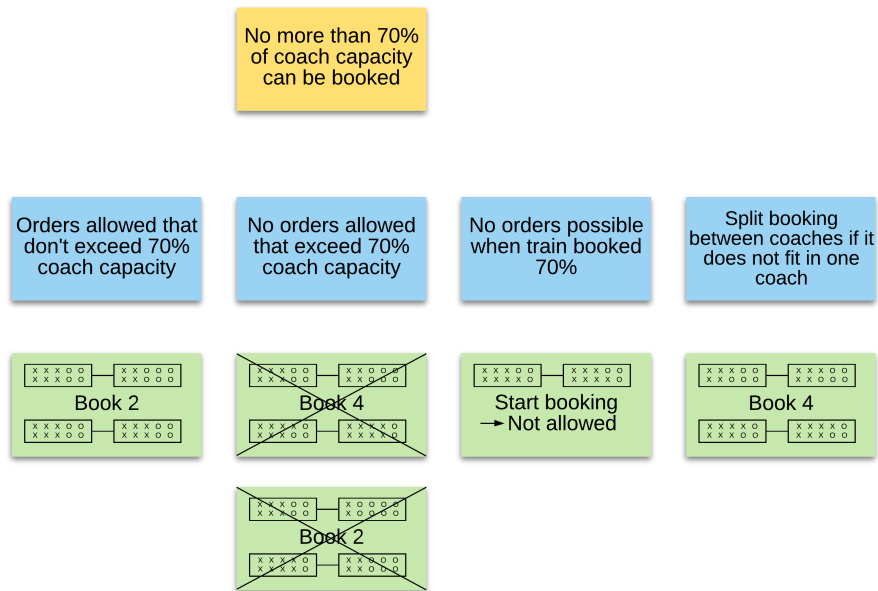


Figure. C.3: Example Mapping – US3

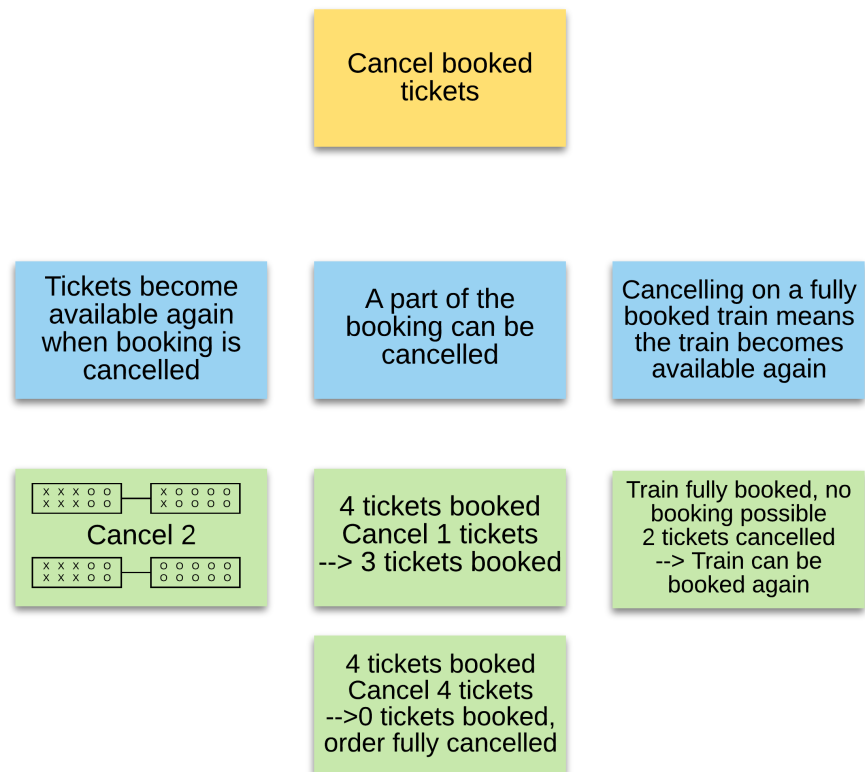


Figure. C.4: Example Mapping – US4

## Appendix C.2 User Story Outputs – Feature Mapping

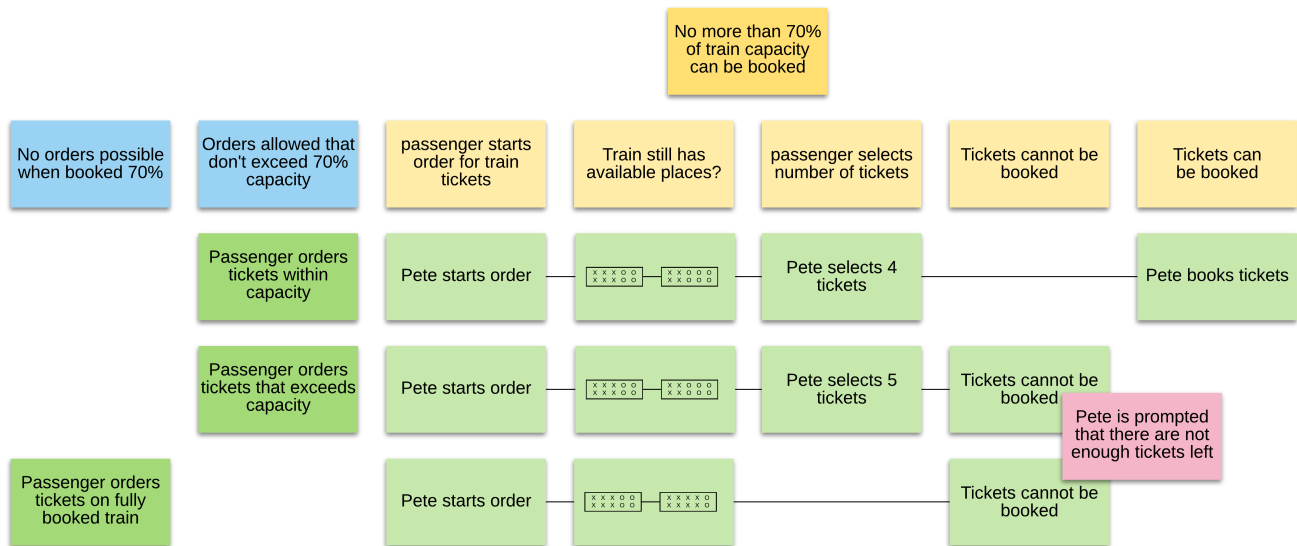


Figure. C.5: Feature Mapping – US1

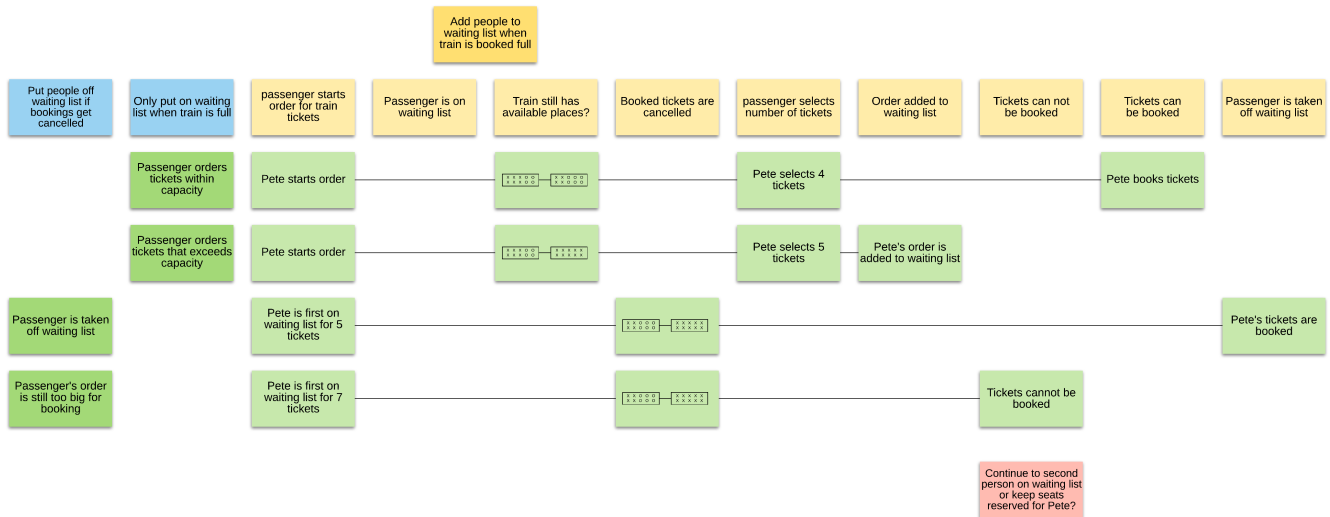


Figure. C.6: Feature Mapping – US2

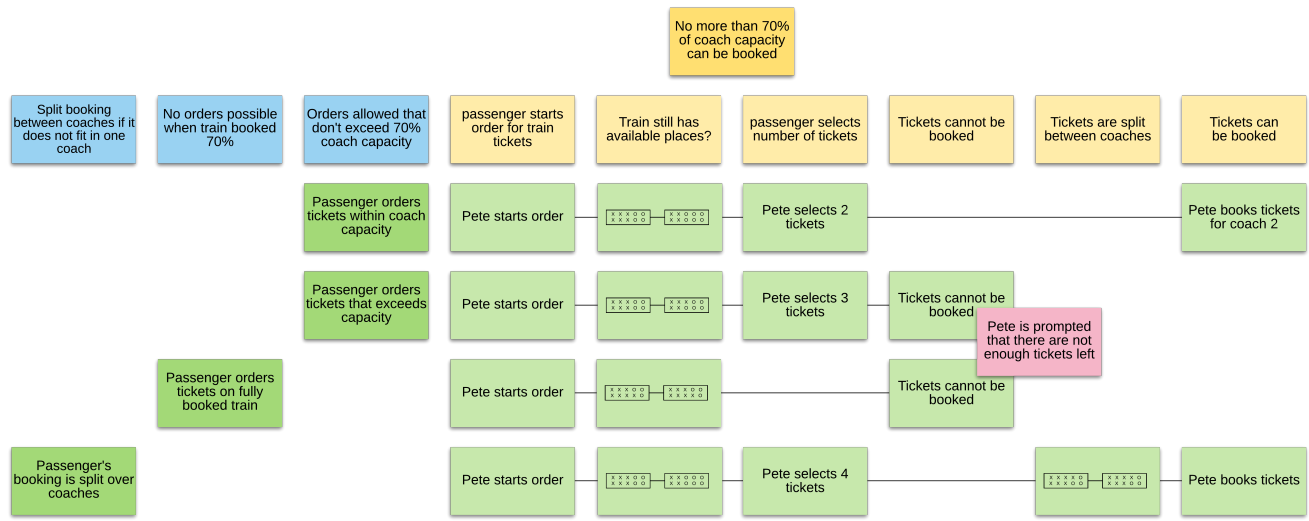


Figure. C.7: Feature Mapping – US3

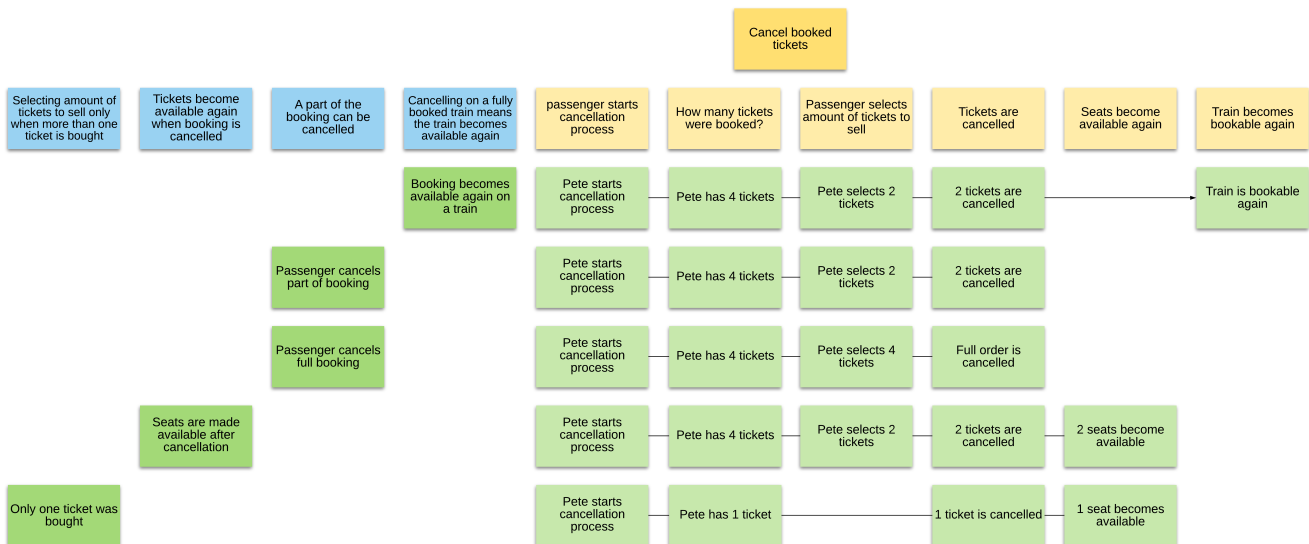


Figure. C.8: Feature Mapping – US4

## Appendix C.3 Lecture Slides

# User Story Refinement Using Three Amigo Sessions

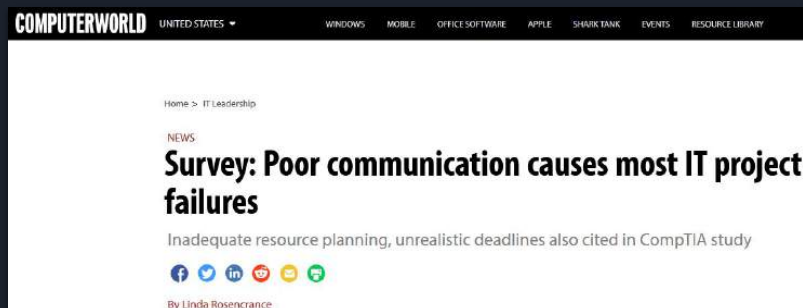


Requirements Engineering 2019/2020  
Jasper Berends and Fabiano Dalpiaz  
24-02-2020

## Contents

- **Introduction**
- Behaviour-Driven Development
- Three Amigo Sessions
- Example Mapping
- Feature Mapping
- Today's Experiment

## Introduction



### Poor Communication Is Still The Primary Contributor To Project Failure



**Stephan Zoder** Contributor

Manufacturing

Expert in data-driven physical-to-digital business transformation.



## Introduction

- Miscommunications occur a lot in software development and can be very costly
- There is a gap between jargon used by domain experts and software developers
- Presented solution → **Behaviour-Driven Development (BDD)**



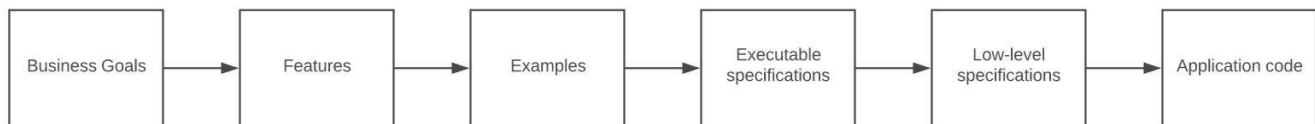
## Contents

- Introduction
- **Behaviour-Driven Development**
- Three Amigo Sessions
- Example Mapping
- Feature Mapping
- Today's Experiment



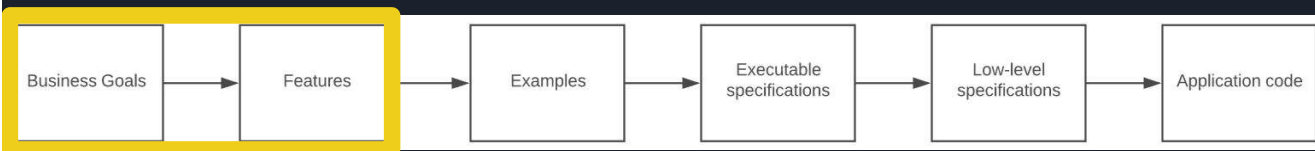
## Behaviour-Driven Development (BDD)

- Focus on **user behaviour**
- Writing **tests before** implementing **code**
- Writing specifications in **ubiquitous language**



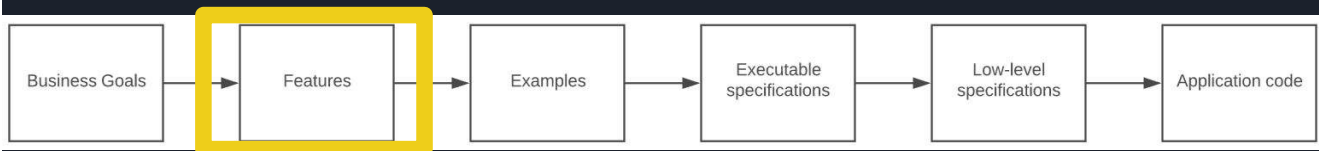
## BDD - Example

- *Business goal:* a bookstore wishes to create an online presence  
→ Create an **online webstore**
- *Feature:* Shopping Cart



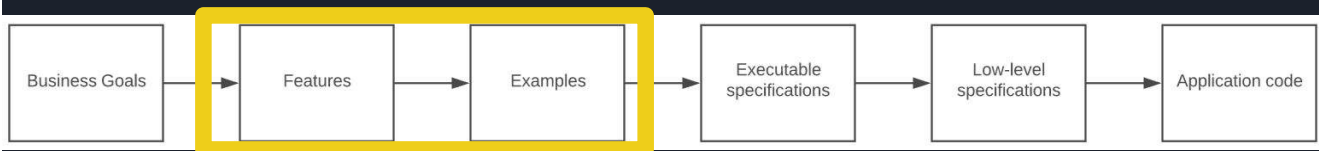
## BDD - Example

*As a* visitor of the online book store  
*I want to* save books in a shopping cart  
*So that* I can buy multiple books at once.



## BDD - Example

- Examples are written in the *Gherkin language*
- Focus: writing specific examples rather than general requirements



## BDD - Example

- 3 main Gherkin keywords:
  - Given
  - When
  - Then



## BDD - Example

- *Scenario: Adding a book to shopping cart*
  - Given** my shopping cart is empty
  - When** I add a book to my shopping cart
  - Then** my shopping cart should contain one book



## BDD - Example

- Executable specifications are automatically generated by tools
- Cucumber | SpecFlow



## BDD - Example

```
[Binding]
public class ShoppingCartSteps
{
    [Given(@"my shopping cart is empty")]
    public void GivenMyShoppingCartIsEmpty()
    {
        ScenarioContext.Current.Pending();
    }

    [When(@"I add a book to my shopping cart")]
    public void WhenIAddABookToMyShoppingCart()
    {
        ScenarioContext.Current.Pending();
    }

    [Then(@"my shopping cart should contain one book")]
    public void ThenMyShoppingCartShouldContainOneBook()
    {
        ScenarioContext.Current.Pending();
    }
}
```



## BDD - Example

- Tests are made functional
- Tests will fail at first as the application code is not yet written



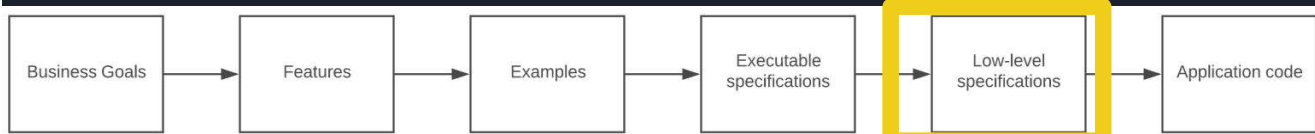
## BDD - Example

```
[Binding]
public class ShoppingCartSteps
{
    ShoppingCart Cart;

    [Given(@"my shopping cart is empty")]
    public void GivenMyShoppingCartIsEmpty()
    {
        Cart = new ShoppingCart();
        Cart.EmptyCart();
    }

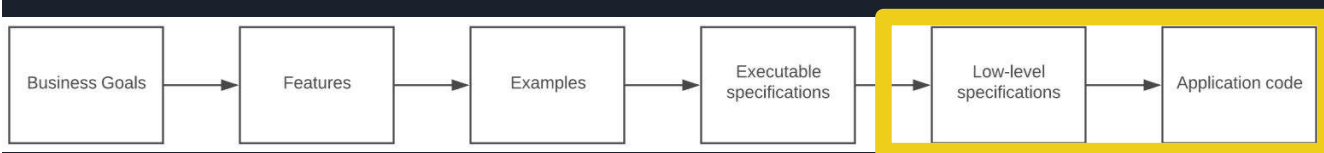
    [When(@"I add a book to my shopping cart")]
    public void WhenIAddABookToMyShoppingCart()
    {
        Book BDD = new Book("BDD IN ACTION");
        Cart.AddBook(BDD);
    }

    [Then(@"my shopping cart should contain one book")]
    public void ThenMyShoppingCartShouldContainOneBook()
    {
        Cart.BookCount.Should().Equal(1);
    }
}
```



## BDD - Example

- Application code is written, functionalities get implemented
- **Tests succeed → Done**



## Contents

- Introduction
- Behaviour-Driven Development
- **Three Amigo Sessions**
- Example Mapping
- Feature Mapping
- Today's Experiment

## Three Amigo Sessions

- Short refinement (<30 minutes) sessions with people from **different disciplines**
- Three Amigo sessions result in “a **clearer description** of an increment of work often in the form of **examples**, leading to a **shared understanding** for the team”



## Three Amigo Sessions

- 2 techniques are presented:
  - **Example Mapping**
  - **Feature Mapping**
- Session objective: shared understanding and clear description of **how a user story should be implemented**
- Examples do not follow a specific format
  - Text, drawings, icons, ...

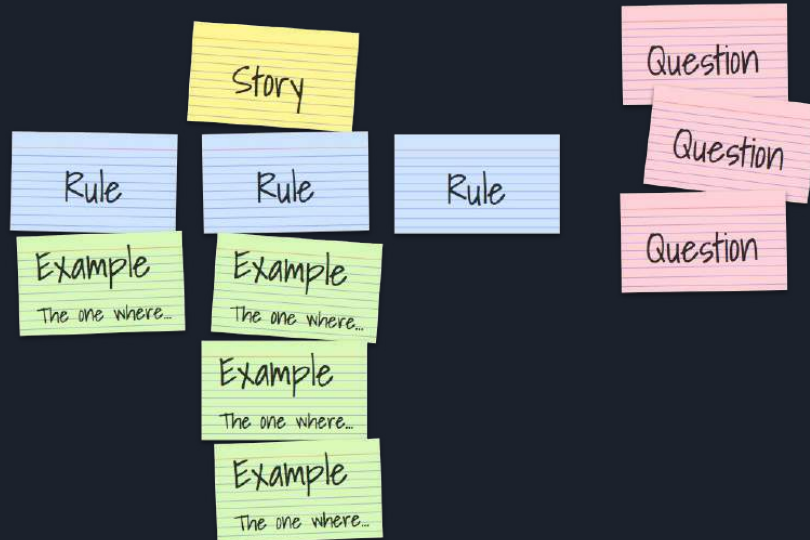


## Contents

- Introduction
- Behaviour-Driven Development
- Three Amigo Sessions
- **Example Mapping**
- Feature Mapping
- Today's Experiment



## Example Mapping







## Example Mapping

- Four different kinds of cards
  - Story
  - Rule
  - Examples
  - Questions
- **Rules make up the acceptance criteria of a user story, examples illustrate the rules**



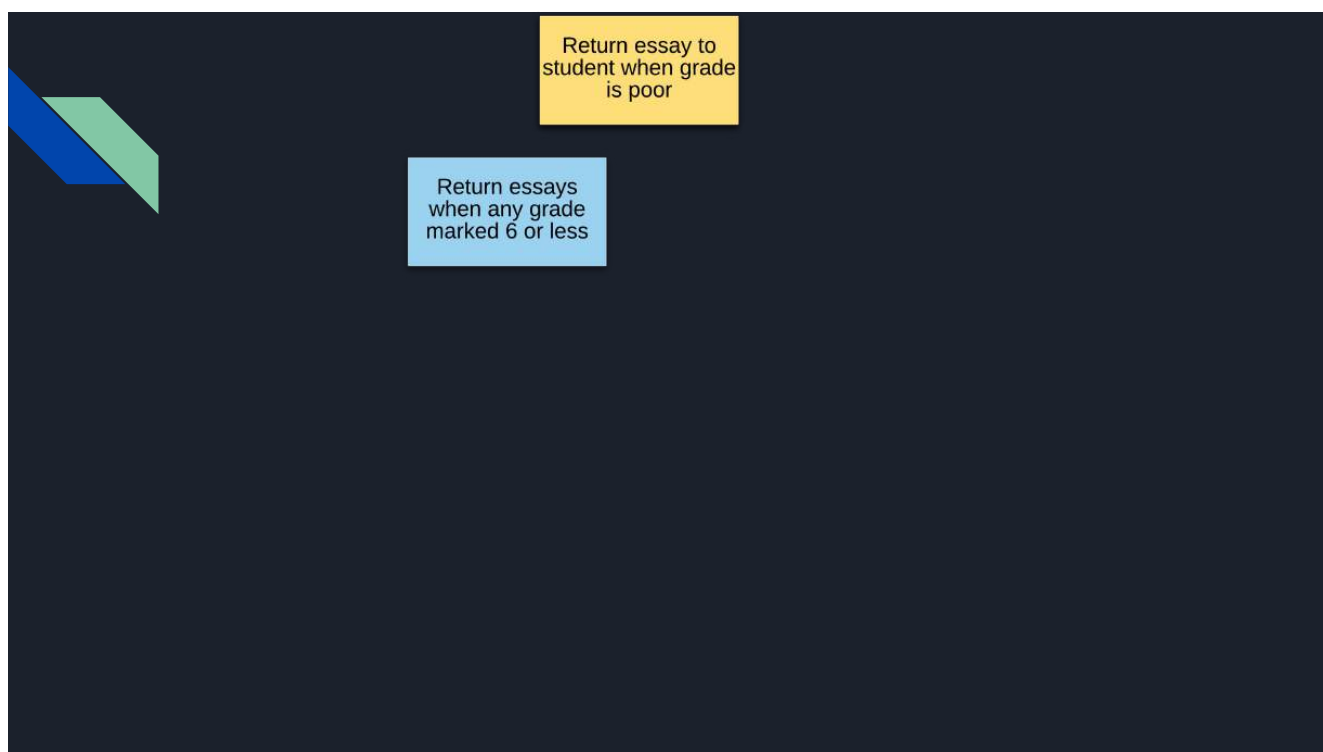
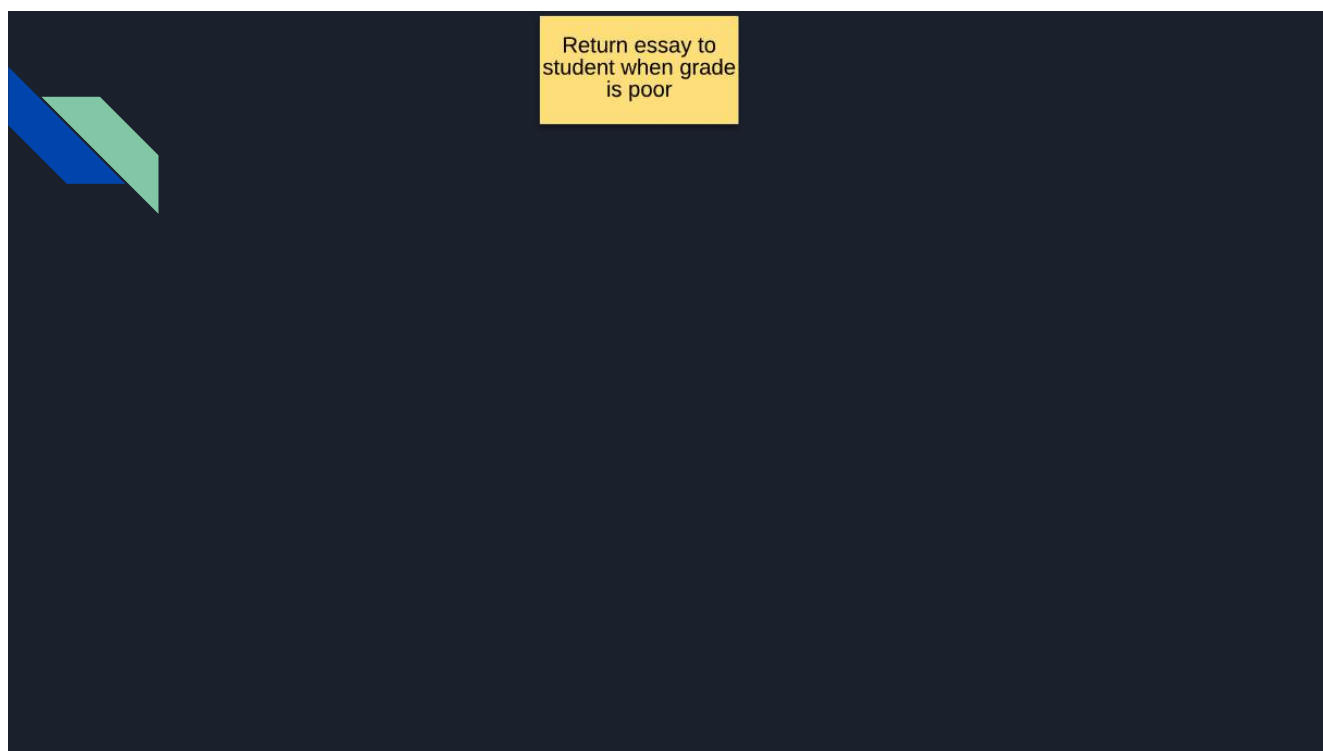
## Example Mapping

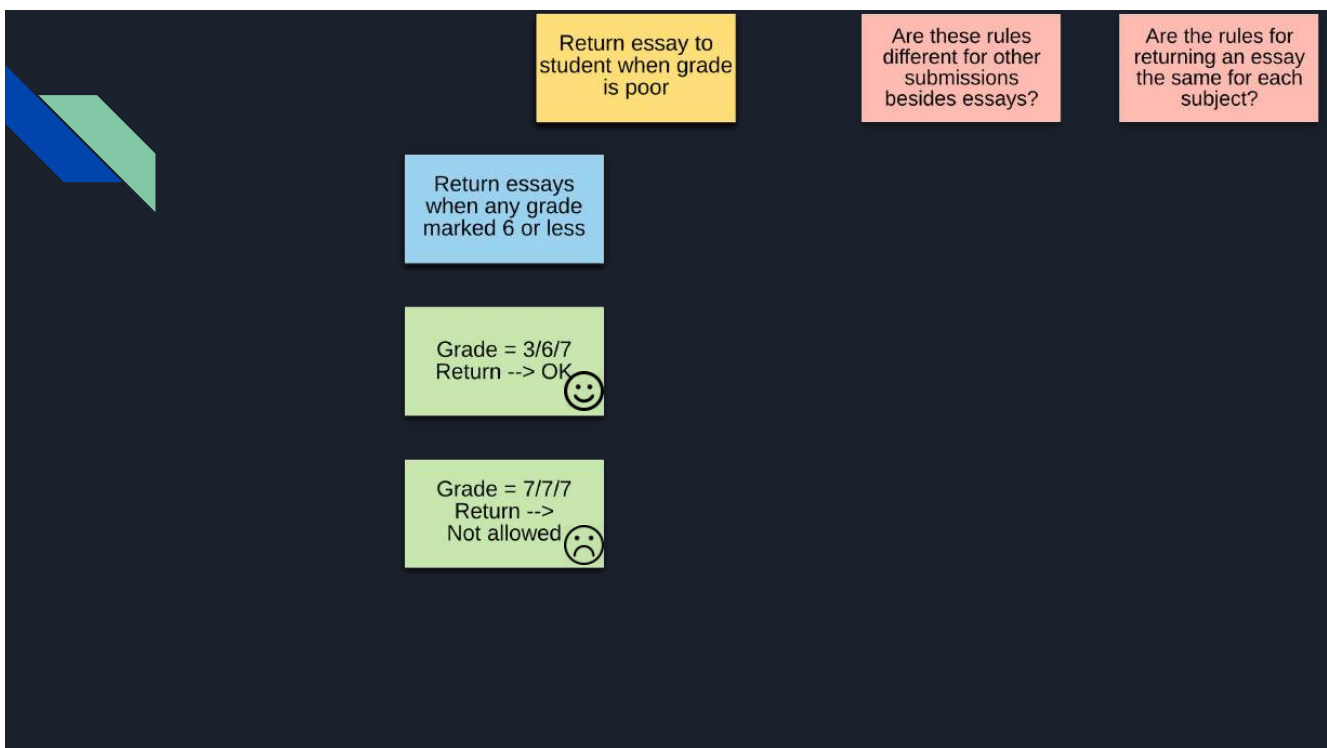
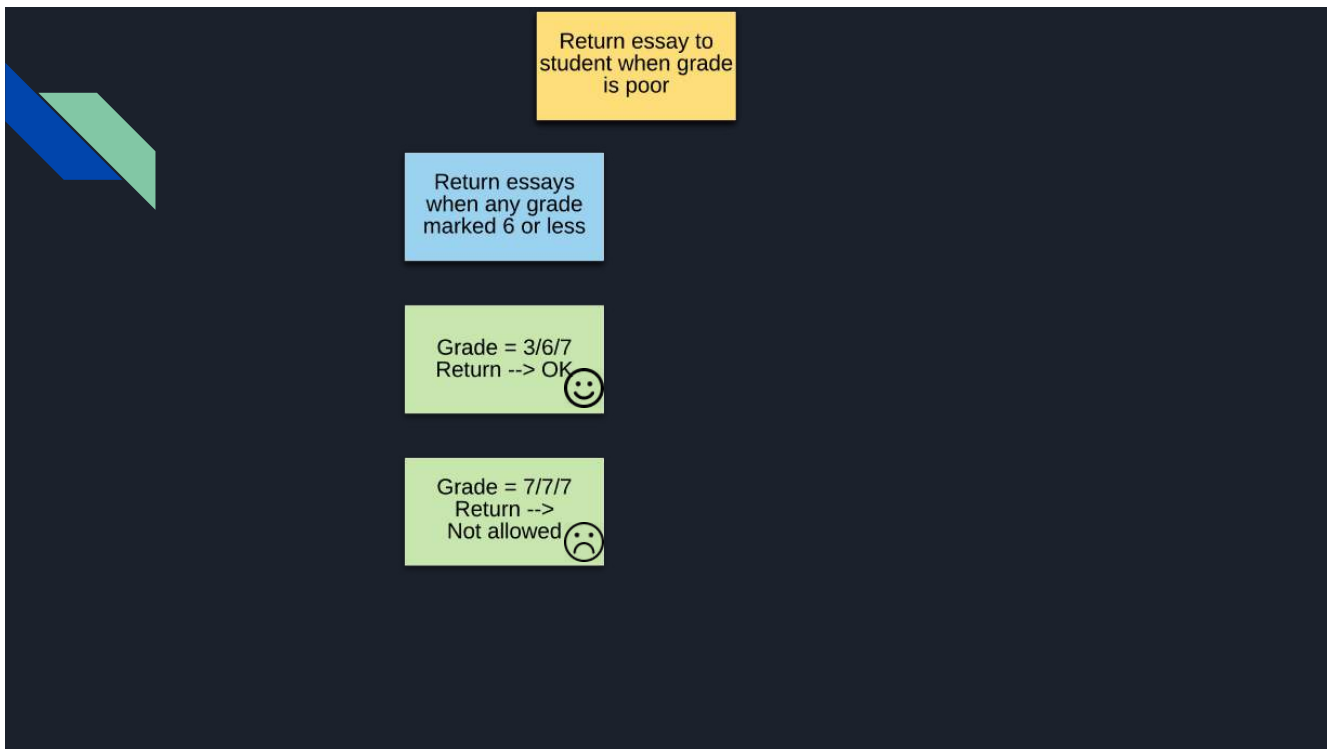
- Example → Grading system

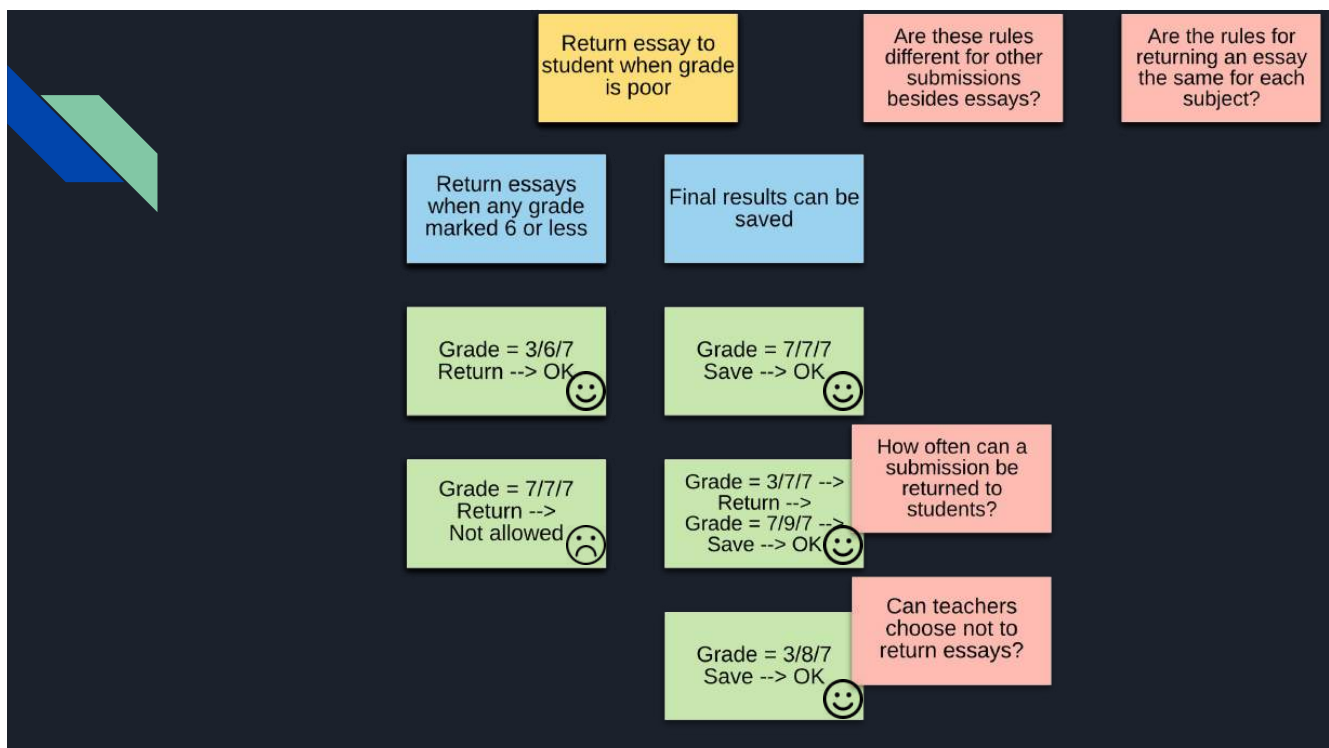
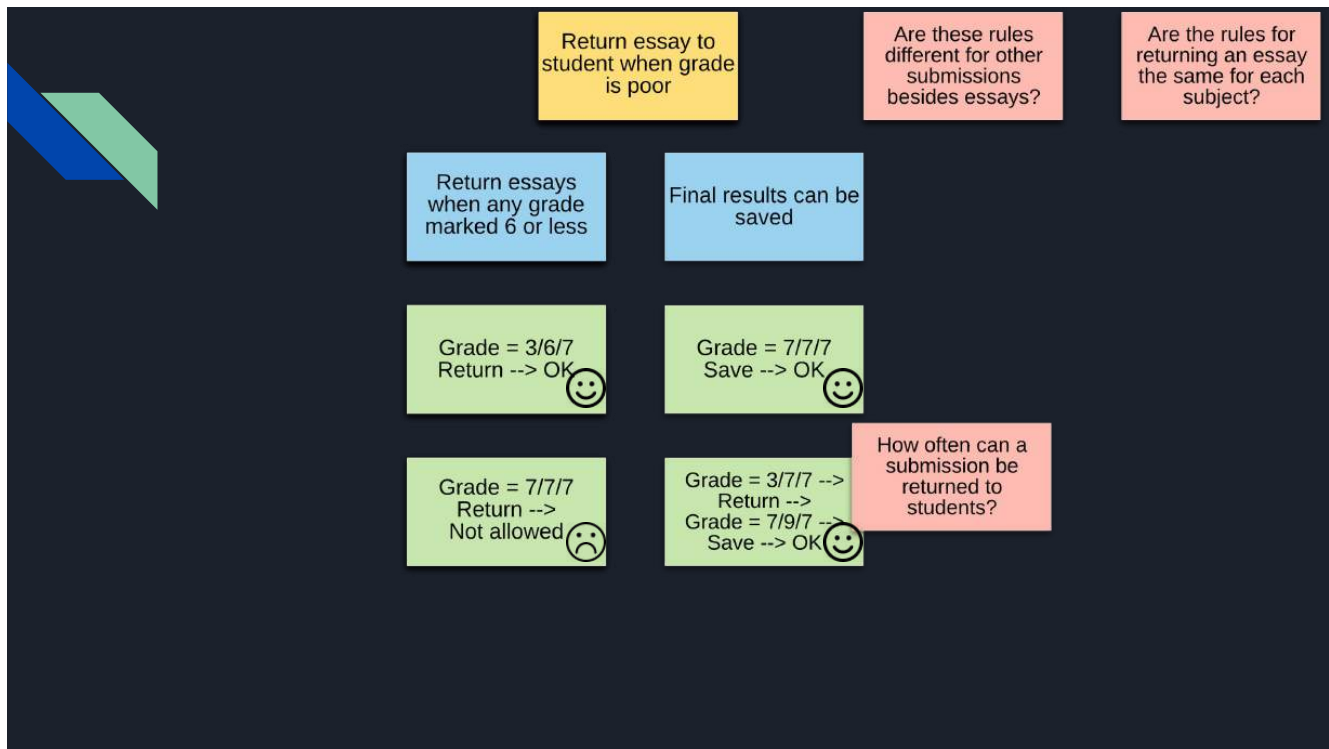
*As a* teacher marking student essays

*I want to* be able to return an essay to a student for corrections when the marks are poor

*In order to* allow students to learn from their mistakes









## Example Mapping

- **Free format**
  - Already know some rules? Start with those
  - Already know examples? Write them down and come up with a rule later!
- Questions give insight in what still needs to be answered



## Example Mapping - Definitions

- **Story**: represents a user story for an increment of work
- **Rule**: Acceptance Criterion of user story
- **Example**: illustrates specific functionality
  - Examples illustrate rules, and rules explain (or give context to) the examples
- **Question**: “known unknowns” of the user story
  - Assumptions should also be written down!



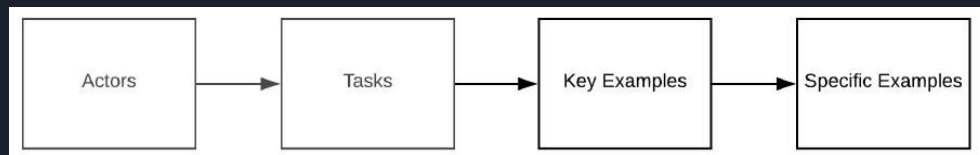
## Contents

- Introduction
- Behaviour-Driven Development
- Three Amigo Sessions
- Example Mapping
- **Feature Mapping**
- Today's Experiment



## Feature Mapping

- More structured than Example Mapping



## Feature Mapping

- Example → Grading system

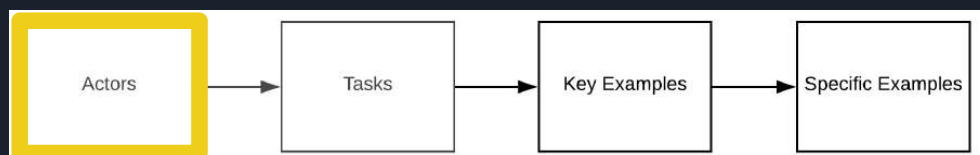
*As a* teacher marking student essays

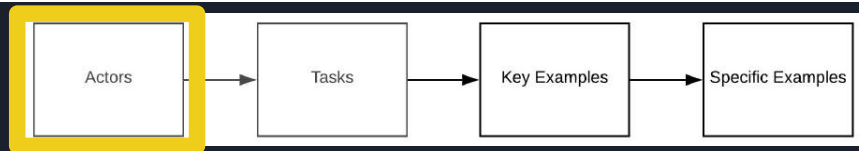
*I want to* be able to return an essay to a student for corrections when the marks are poor

*In order to* allow students to learn from their mistakes

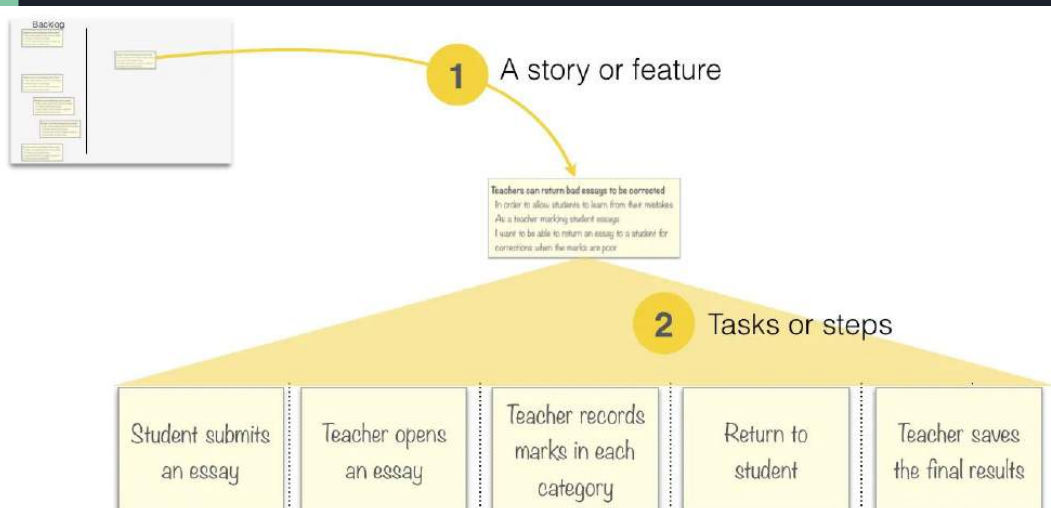
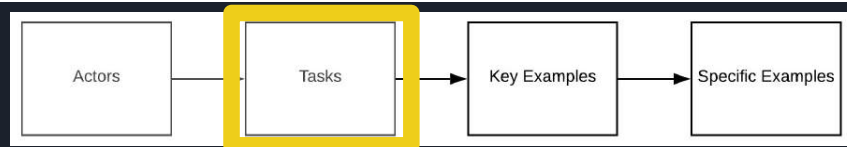
## Feature Mapping

- **Main actor** is often mentioned in User Story template
- **Other actors** and their involvement of the user story are specified as well

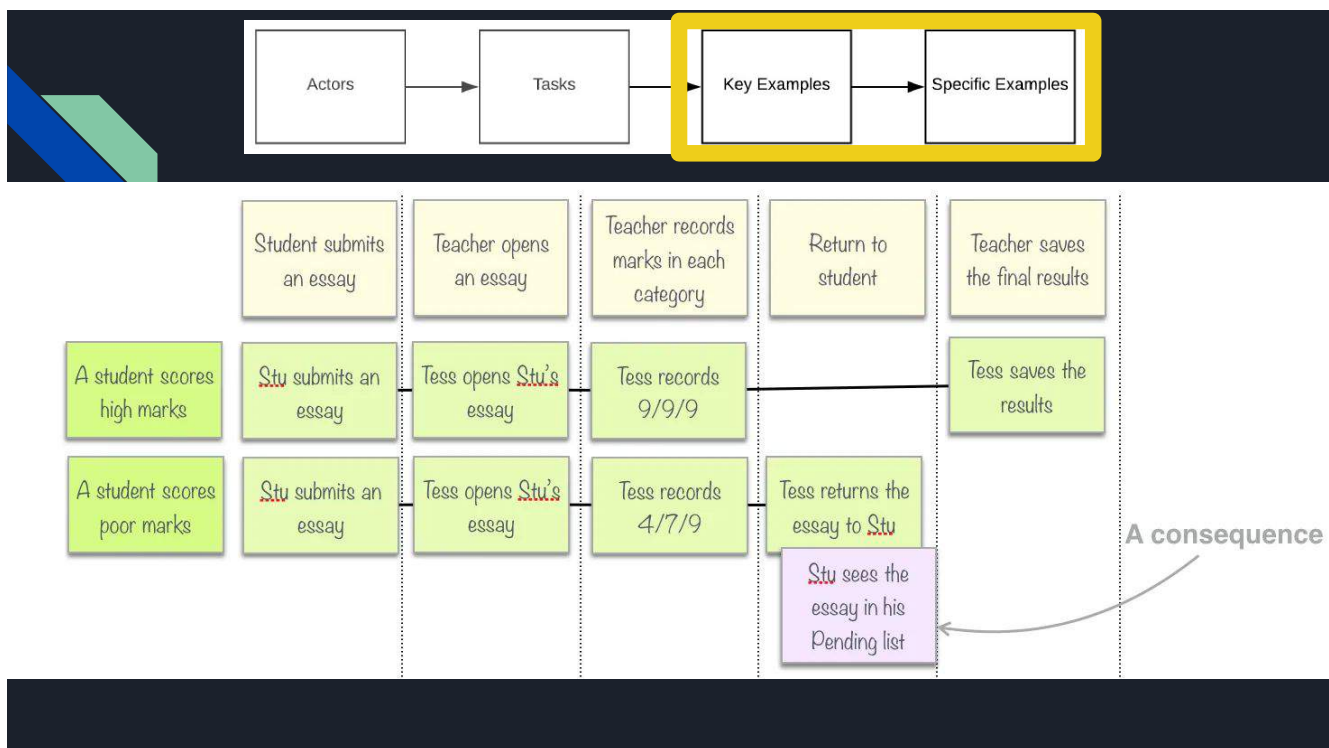
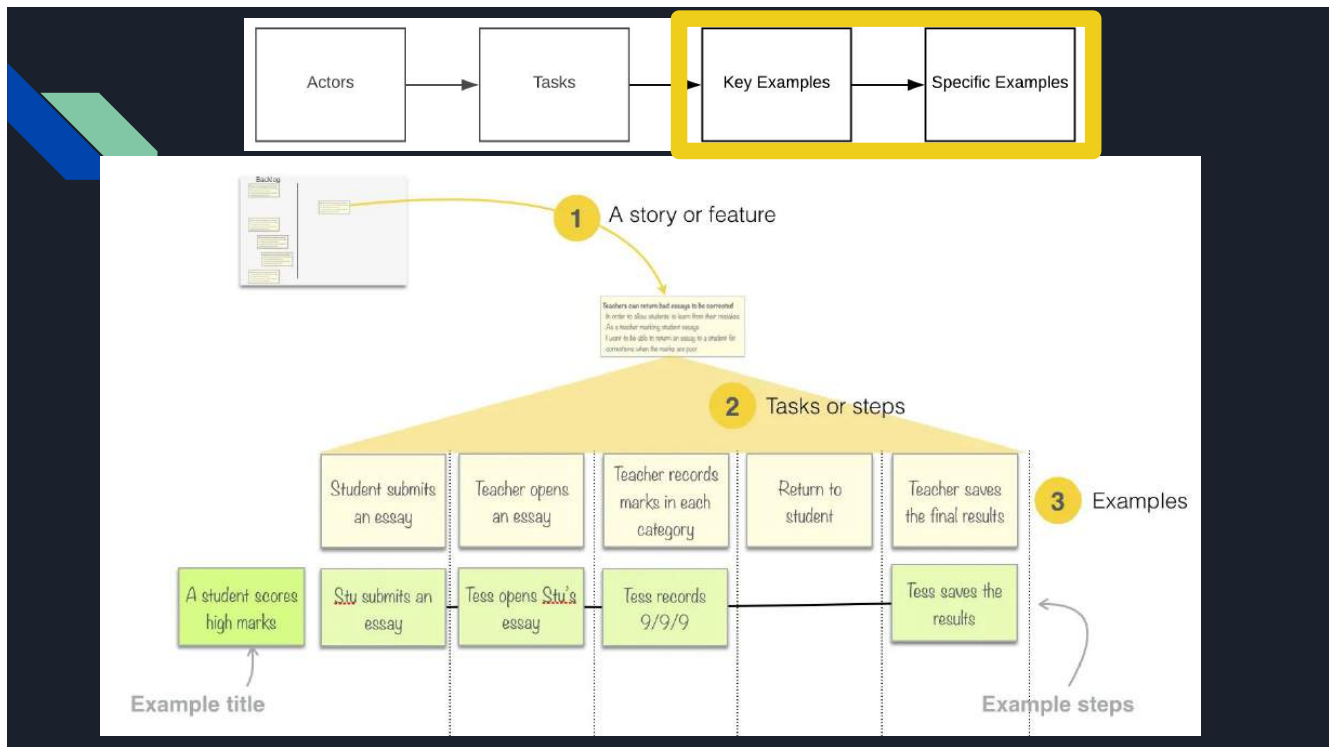




- The student who submits the essay (let's call this actor Stu)
- The teacher, who marks the essay (Tess)

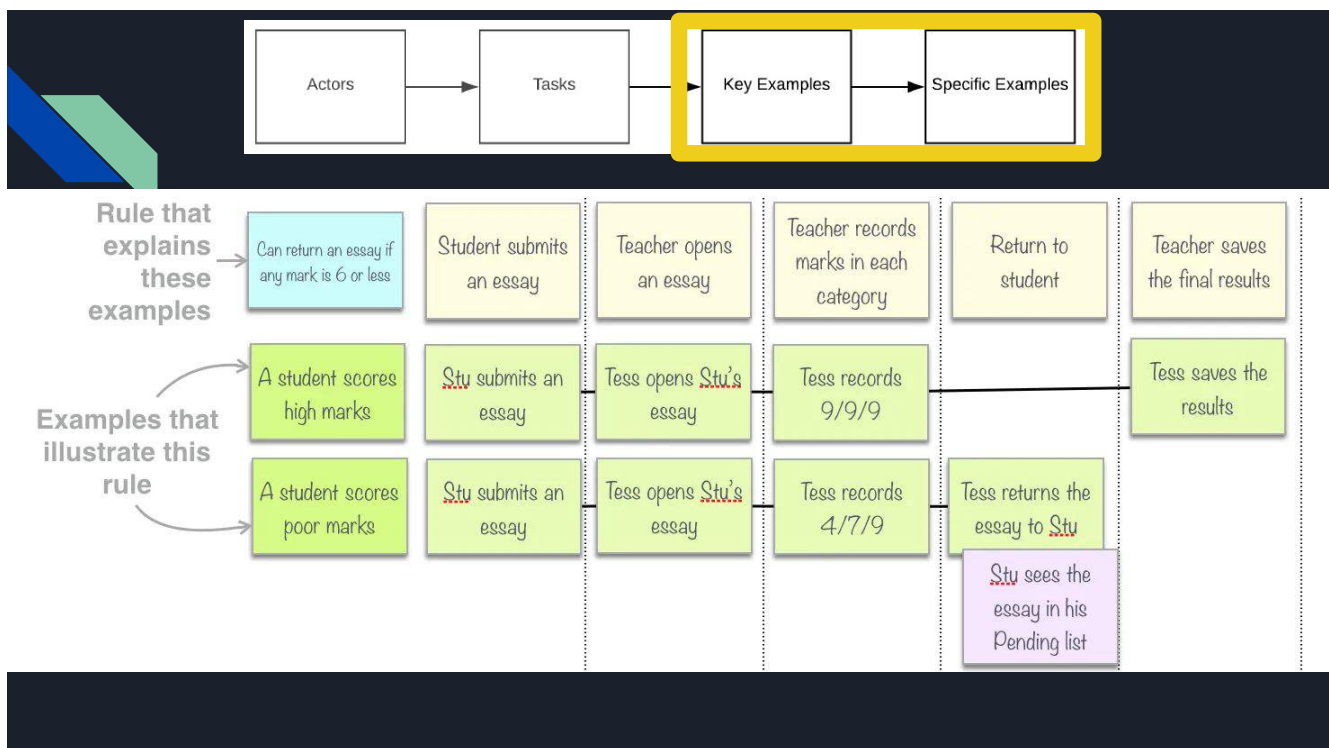


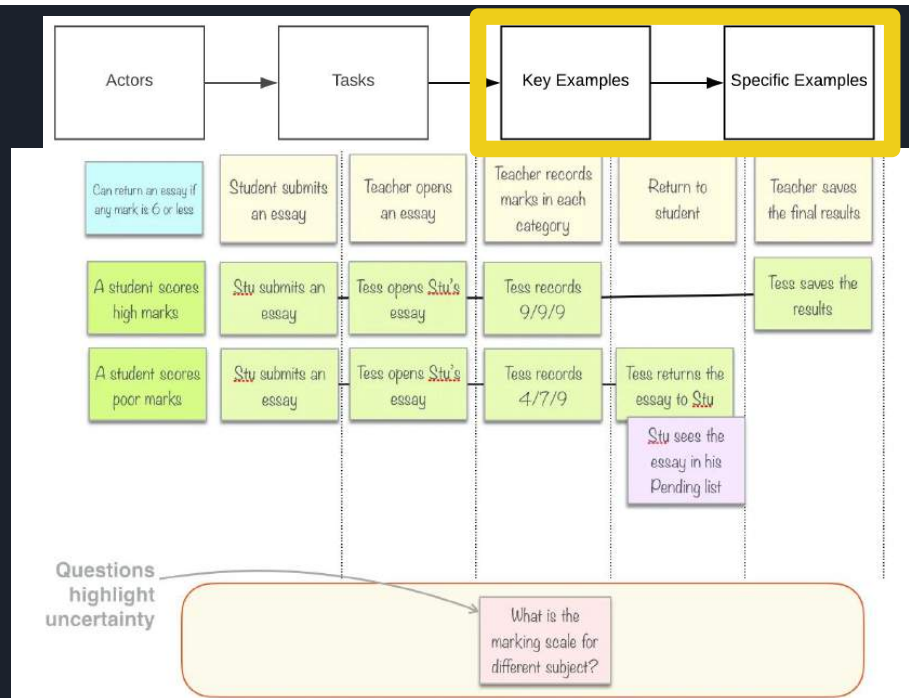




## Feature Mapping - Definitions

- **Story:** represents a user story for an increment of work
- **Rule:** Acceptance Criterion of user story
- **Example:** illustrates specific functionality
  - Examples illustrate rules, and rules explain (or give context to) the examples
- **Question:** “known unknowns” of the user story
  - Assumptions should also be written down!
- **Consequence:** Explicit result of an example

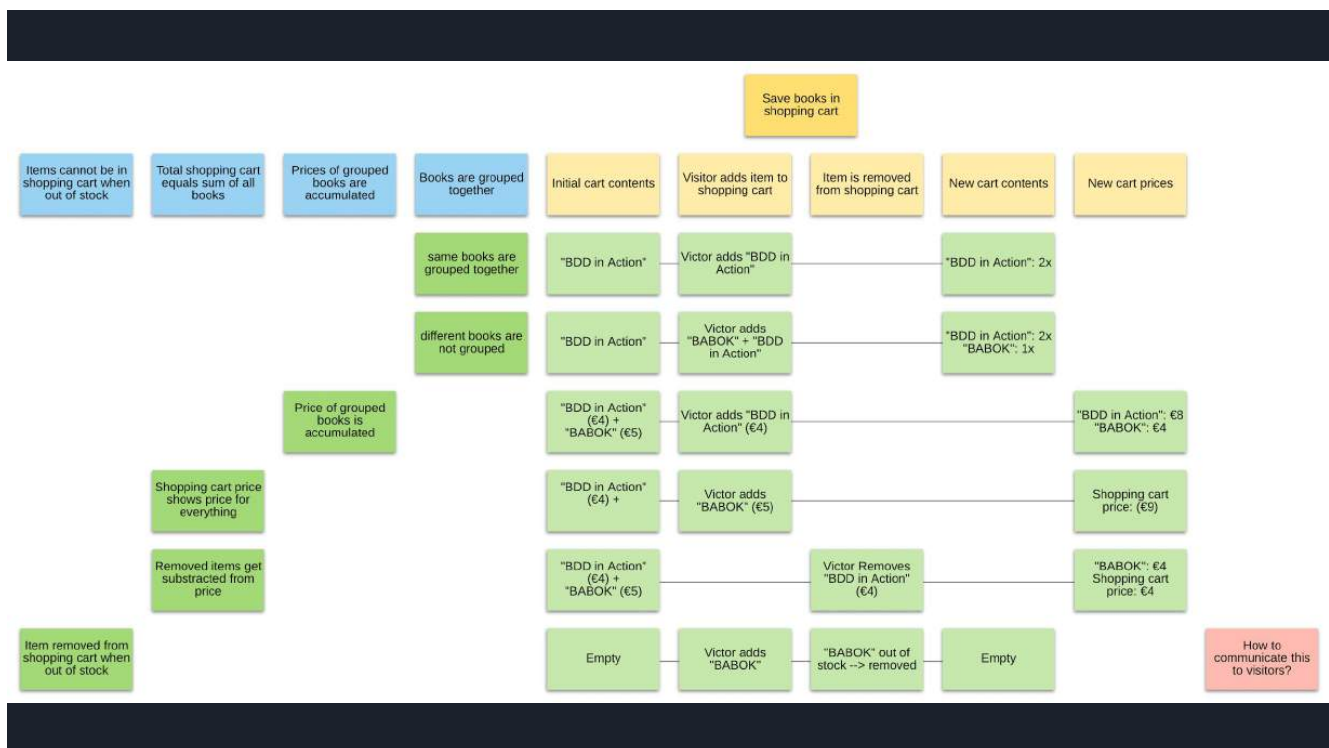
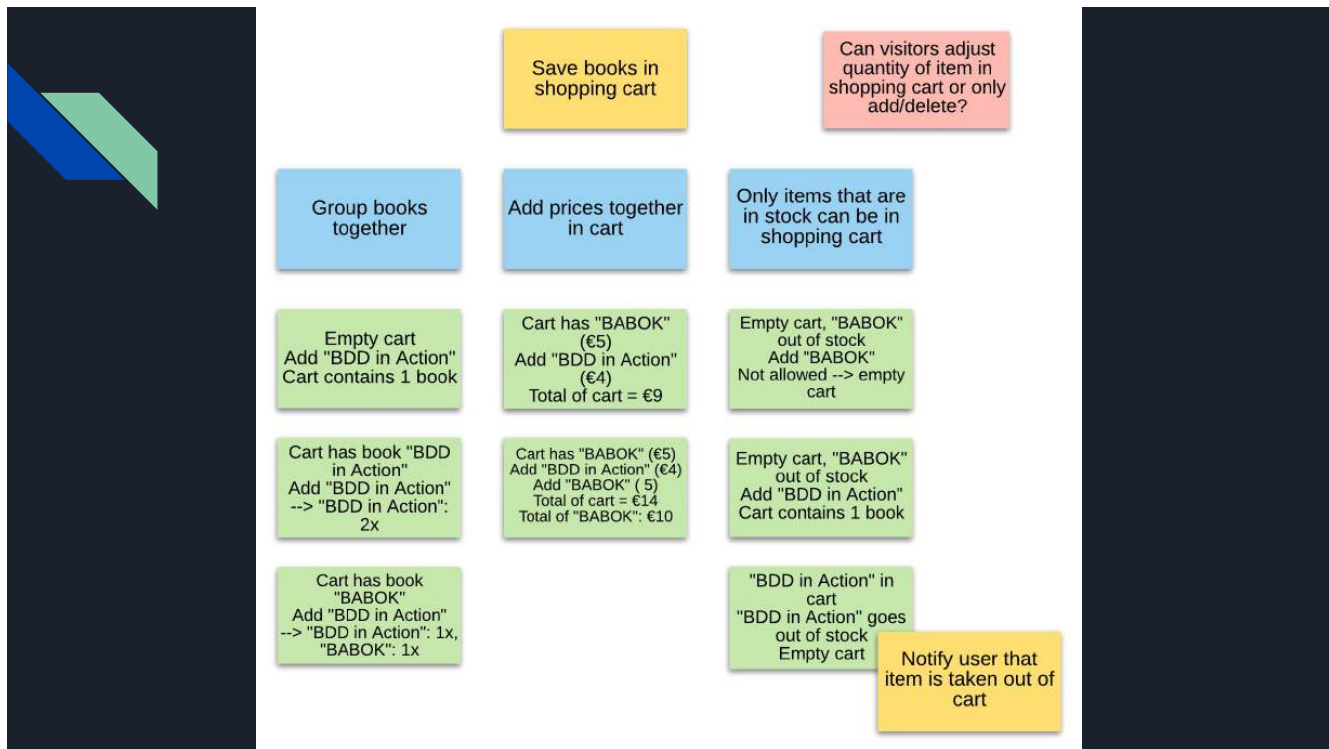




## Technique Comparison

- Example from earlier:

*As a* visitor of the online book store  
*I want to* save books in a shopping cart  
*So that* I can buy multiple books at once.





## Technique Comparison

- With both techniques, **each example can be translated to one Gherkin scenario**



## Contents

- Introduction
- Behaviour-Driven Development
- Three Amigo Sessions
- Example Mapping
- Feature Mapping
- **Today's Experiment**



## PremRide

- A new train service provider is introduced on the Dutch train tracks reservable seats
- For PremRide coaches, a ticket booking system is being developed



## PremRide, context and problems

### Context

- Passengers **can already** book or cancel tickets
- Multiple tickets can be booked in one reservation

### Problems

- Trains are now **often fully booked** during peak hours
- Passengers want to be **notified** when tickets are cancelled



## PremRide

*As a* train conductor

*I want to* make sure no more than 70% of all available seats in a coach can be booked

*So that* people without a prior reservation also have a chance to get on the train.



## PremRide

*As a* passenger

*I want* to sign up for tickets on a waiting list if a train is fully booked

*So that* I can still get a ticket if previous bookings get cancelled



## PremRide

- The following materials are provided:
  - Instructions
  - Case description
  - User Story descriptions
  - Technique overviews
  - Questionnaires
  - Markers
  - Post-its



## PremRide - Schedule

- 14:00 - 14:10: Fill in demographics questionnaire + Informed Consent, read case and user story description (BBG 2.14)
- 14:15 - 14:45: Perform Three Amigo Session 1 🕒
- 14:45 - 15:00: Upload photo of output, fill in questionnaire, take small break
- 15:00 - 15:30: Perform Three Amigo Session 2 🕒
- 15:30 - 15:45: Upload photo of output, fill in questionnaire, take small break
- 15:45: Meet back in BBG 2.14





## PremRide

- Rooms:
  - BBG 2.31
  - BBG 3.31
  - BBG 4.31
  - BBG 4.56
  - BBG 5.56
  - BBG 2.14



## PremRide

- Any questions?
- Good luck! We will meet again at 15:45

## Appendix C.4 Student Handout Package

### Instructions – Group 1

#### Order of sessions

Three Amigo Session 1: Example Mapping - US1

Three Amigo Session 2: Feature Mapping – US2

After filling in the questionnaire and reading the case and user story description, you may **stay in BBG 2.14**

#### Schedule

14:00 - 14:10: Fill in demographics questionnaire,  
read case and user story description (BBG 2.14)

14:15 - 14:45: Perform Three Amigo Session 1 🕒

14:45 - 15:00: Upload photo of output, fill in questionnaire, take small break

15:00 - 15:30: Perform Three Amigo Session 2 🕒

15:30 - 15:45: Upload photo of output, fill in questionnaire, take small break

15:45: Meet back in BBG 2.14

#### Materials

If you wish to review the full slides, you can find them on Blackboard. Some slides of Example Mapping and Feature Mapping are included in this handout.

If you miss any materials or run out of anything, more can be requested at **BBG 2.14**

#### Questions?

If you have any questions during the session and no coordinator is present in your room, see if you can wait for one come by, as we are regularly walking around between the different rooms. If not, a coordinator will always be present in BBG 2.14. This only concerns questions that you may have regarding the process of the techniques. If you have questions about the user stories, those should be written down on a question card.

#### Uploading materials

For each Three Amigo session, we would like for you to make a photo of your output once you are done with the session. You can upload your photos in a shared presentation by visiting <https://tinyurl.com/TAREUU>. Please use the template that is included in the presentation and put both of your pictures in consecutive slides. Make sure the everything is readable on the photos!

#### Finished?

Once you are finished with the experiment, please clean up all materials in your room. Return to BBG 2.14 at 15:45 for the final debriefing.

## Case Description – PremRide

In order to increase competition in the Dutch railroad sector, a new service provider is introduced: **PremRide**. PremRide differentiates itself by offering **reservable seats**, much like is often the case in Germany. They do allow people to also buy tickets at the station, mainly for the elderly who do not like using technological solutions. The system that is being developed focuses purely on the **online reservations**, not the on-site ticket sales.

Current situation: PremRide already has the ability to book and cancel tickets. A passenger can order multiple tickets at once in order to ensure an entire group can get aboard the train. In previous user story refinement sessions, examples for illustrating train capacity referred to a train with 2 coaches with 10 seats each.

Problems/needs: PremRide is still very limited to the basic functionalities. Train conductors have noted that the trains during peak hours are always fully booked, and that people who did not reserve a seat are often complaining (to the conductors themselves) that they have to wait too long before they can hop on a train. On the other hand, it also sometimes happens that passengers cancel their booked ticket. It would be preferable to be able to notify people who were told that the train was booked that those seats have become available. Therefore, the following two user stories are created, along with some preliminary information:

### US1:

*As a train conductor*

*I want to make sure no more than 70% of all available seats in a coach can be booked  
So that people without a prior reservation also have a chance to get on the train.*

Larger orders that do not fit in a single coach should be split up and divided across coaches. If all coaches are booked full for 70%, then we should not even let passengers start the booking process at all.

### US2:

*As a passenger*

*I want to sign up for tickets on a waiting list if a train is fully booked  
So that I can still get a ticket if previous bookings get cancelled.*

We want to make sure that passengers who cannot book a ticket because a train is fully booked, will have an ability to get on board the train in case tickets get canceled. We should have a waiting list for this, and if their booking could fit in the train after someone else cancels a booking, then we should also be able to take them off the waiting list.

## Performance of Three Amigo Session Techniques

### *Information sheet (24-02-2020)*

Thank you for considering to participate in this study. This information sheet outlines the purpose of the study and provides a description of your involvement and rights as a participant, if you agree to take part.

#### **Purpose of the research**

The purpose of this research is to investigate the performance of Three Amigo session techniques for user story refinement.

#### **Participation**

You are free to decide whether you want to participate or not. Your participation is entirely voluntary, meaning that you will not directly benefit from your participation. However, the results may be shared with you afterwards, if you would be interested in this. The researchers involved in this project do not foresee that there are any risks associated with your participation. If you do decide to participate, you are asked to sign a consent form which you will sign and return prior to the experiment.

#### **Withdrawal procedure**

You can withdraw from the study at any point whilst it is ongoing, without providing any reason for this. Withdrawing from the study will have no effect on you. If you withdraw from the study, we will not retain the information you have given thus far, unless you give us permission to retain the information.

#### **Usage of the data**

The collected information will be analysed and used for research studies by the researcher and may be used in publications and other research outputs.

Personal data will be processed on the legal basis of consent and is gathered for obtaining informed consent and for contacting you. Any personal information you communicate will be treated confidentially. All (personal) data will be treated as confidential as possible. Only the researchers involved in the project will have access to the personal data. Your data will be anonymised in research outputs, meaning that your name will not be used and any details which may reveal your identity will be disguised in any report or publication resulting from the study. You have the right to request access to and rectification or erasure of personal data while it is in storage. All personally identifying information collected will be destroyed once it is no longer needed for the project. Data will be kept in locked files that only the researchers involved in this study can open.

#### **Contact details**

If you have any questions or complaints regarding this study please contact dr. Fabiano Dalpiaz: [f.dalpiaz@uu.nl](mailto:f.dalpiaz@uu.nl).

## Performance of Three Amigo Session Techniques

### *Informed consent form*

Research Investigator: Fabiano Dalpiaz  
Institution name: Utrecht University

### **Please tick the appropriate boxes**

**Yes No**

#### **Taking Part**

I have read and understood the attached project information sheet, dated 24-02-2020, or it has been read to me. ☐ Yes ☐ No

I have been given the opportunity to ask questions about the project and my questions have been answered to my satisfaction. ☐ Yes ☐ No

I agree to take part in the project. I understand that taking part in the project will include sharing materials (outputs of Three Amigo sessions) with the researcher and his research team. ☐ Yes ☐ No

I consent that my participation is voluntary and I understand that I can withdraw from the study at any time whilst it is ongoing without having to give any reasons for why I no longer want to take part. ☐ Yes ☐ No

#### **Use of the information I provide for this project**

I understand that information I provide will be used for the thesis report, publications and potential other research outputs. ☐ Yes ☐ No

I understand my personal data such as my name will not be revealed to people outside the project. ☐ Yes ☐ No

I understand that in any report on the results of this research my identity will remain confidential. ☐ Yes ☐ No

I agree that my (anonymised) words may be quoted in research outputs. ☐ Yes ☐ No

\_\_\_\_\_  
Name of participant

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date

Jasper Berends



24-02-2020

\_\_\_\_\_  
Researcher [printed]

\_\_\_\_\_  
Signature

\_\_\_\_\_  
Date

For information please contact:  
Fabiano Dalpiaz ([f.dalpiaz@uu.nl](mailto:f.dalpiaz@uu.nl))

**Questionnaire Thee Amigo Session**    **Technique:** \_\_\_\_\_  
**User Story (US1 or US2):** \_\_\_\_\_    **Group:** \_\_\_\_\_

Question		Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1	I found the technique complex and difficult to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2	We have a good understanding of all rules of this user story that were discussed during the session	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3	We have a good understanding of all examples of this user story that were discussed during the session	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4	I believe that this technique allows me to express acceptance criteria with little effort.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5	We worked together in a well-coordinated fashion.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6	We could not reach agreement on certain rules.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7	Acceptance criteria represented using this technique would be easy for participants to understand.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8	Overall, I found the technique difficult to learn.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
9	This technique would make it easy for participants to verify whether acceptance criteria are correct.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10	I found the technique easy to learn.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
11	We understand what questions still need to be answered before we can proceed with implementing this user story	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
12	We had very few misunderstandings about what to do.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
13	We needed to backtrack and start over a lot.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Flip over for remaining questions →

Question		Strongly disagree	Disagree	Neutral	Agree	Strongly agree
14	Overall, I found the technique to be useful.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
15	We did not have a good understanding of the examples of this user story by the end of the session.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
16	Using this technique would make it more difficult to maintain the acceptance criteria.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
17	I found it difficult to apply the technique in the tasks of the experiment.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
18	I would use this technique in the future if I need to express acceptance criteria unambiguously.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
19	We accomplished the task smoothly and efficiently.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
20	I found the rules of the technique clear and easy to understand.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
21	We have agreement on the overall output of this session	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
22	Overall, I think this technique does not provide an effective solution to the problem of representing acceptance criteria.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
23	Using this technique would make it easy to communicate acceptance criteria with other stakeholders of a project.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
24	I am not confident that I am now competent to apply this technique in practice.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
25	Overall, I think this method is an improvement over other user story refinement techniques.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
26	There was much confusion about how we would accomplish the task.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
27	I would rather use a different way of expressing acceptance criteria if I ever need to define them in the future	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Questionnaire Thee Amigo Session**    **Technique:** \_\_\_\_\_  
**Session (US1 or US2):** \_\_\_\_\_    **Group:** \_\_\_\_\_

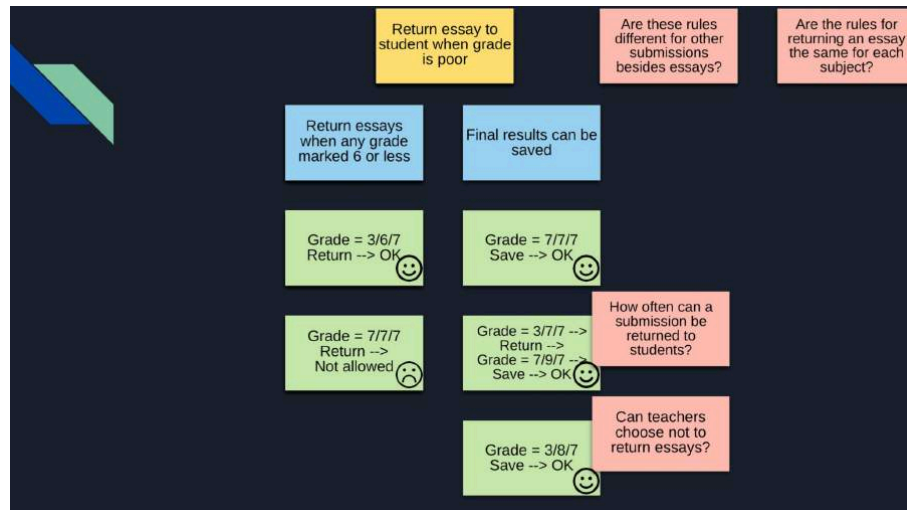
Question		Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1	I found the technique complex and difficult to use.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2	We have a good understanding of all rules of this user story that were discussed during the session	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3	We have a good understanding of all examples of this user story that were discussed during the session	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
4	I believe that this technique allows me to express acceptance criteria with little effort.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
5	We worked together in a well-coordinated fashion.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
6	We could not reach agreement on certain rules.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
7	Acceptance criteria represented using this technique would be easy for participants to understand.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
8	Overall, I found the technique difficult to learn.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
9	This technique would make it easy for participants to verify whether acceptance criteria are correct.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
10	I found the technique easy to learn.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
11	We understand what questions still need to be answered before we can proceed with implementing this user story	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
12	We had very few misunderstandings about what to do.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
13	We needed to backtrack and start over a lot.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Flip over for remaining questions →



Question		Strongly disagree	Disagree	Neutral	Agree	Strongly agree
14	Overall, I found the technique to be useful.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
15	We did not have a good understanding of the examples of this user story by the end of the session.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
16	Using this technique would make it more difficult to maintain the acceptance criteria.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
17	I found it difficult to apply the technique in the tasks of the experiment.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
18	I would use this technique in the future if I need to express acceptance criteria unambiguously.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
19	We accomplished the task smoothly and efficiently.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
20	I found the rules of the technique clear and easy to understand.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
21	We have agreement on the overall output of this session	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
22	Overall, I think this technique does not provide an effective solution to the problem of representing acceptance criteria.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
23	Using this technique would make it easy to communicate acceptance criteria with other stakeholders of a project.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
24	I am not confident that I am now competent to apply this technique in practice.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
25	Overall, I think this method is an improvement over other user story refinement techniques.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
26	There was much confusion about how we would accomplish the task.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
27	I would rather use a different way of expressing acceptance criteria if I ever need to define them in the future	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

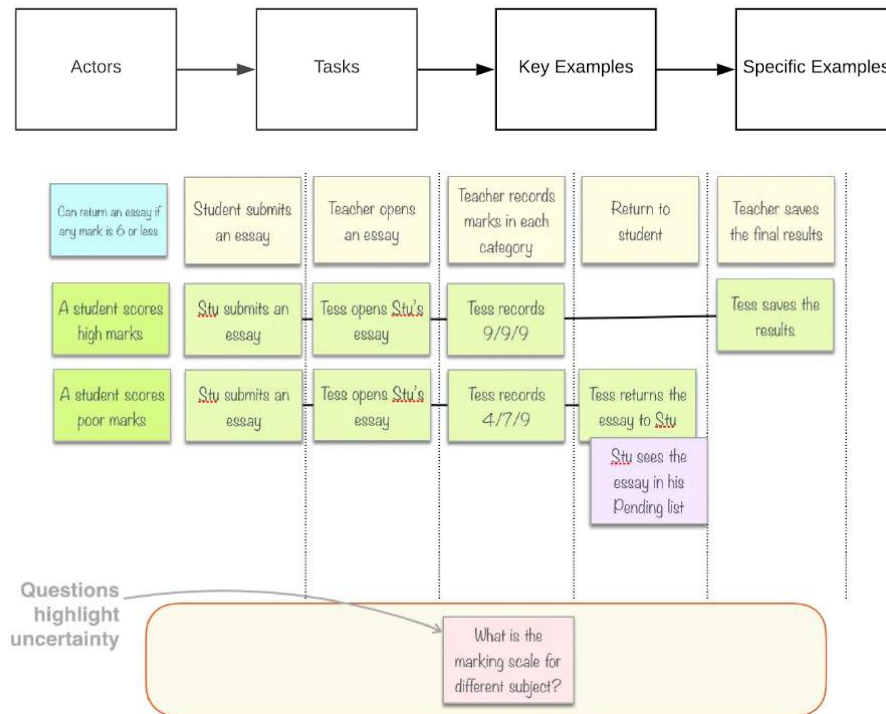
## Technique overview – Example Mapping



## Example Mapping - Definitions

- **Story:** represents a user story for an increment of work
- **Rule:** Acceptance Criterion of user story
- **Example:** illustrates specific functionality
  - Examples illustrate rules, and rules explain (or give context to) the examples
- **Question:** “known unknowns” of the user story
  - Assumptions should also be written down!

### Technique overview – Feature Mapping



### Feature Mapping - Definitions

- **Story:** represents a user story for an increment of work
- **Rule:** Acceptance Criterion of user story
- **Example:** illustrates specific functionality
  - Examples illustrate rules, and rules explain (or give context to) the examples
- **Question:** “known unknowns” of the user story
  - Assumptions should also be written down!
- **Consequence:** Explicit result of an example

## Demographics questionnaire

Age: \_\_\_\_\_

Previously obtained Bachelor's/Master's programs (if multiple, please name all):

\_\_\_\_\_

Please state your experience with the following concepts:

	No experience	Little experience (Some encounters over the past few years)	Some experience (Some small projects over the past few years)	Much experience (Some bigger projects over the past few years or many smaller projects)
User story refinement	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
working in an Agile software development environment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Gherkin	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Example Mapping	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Feature Mapping	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

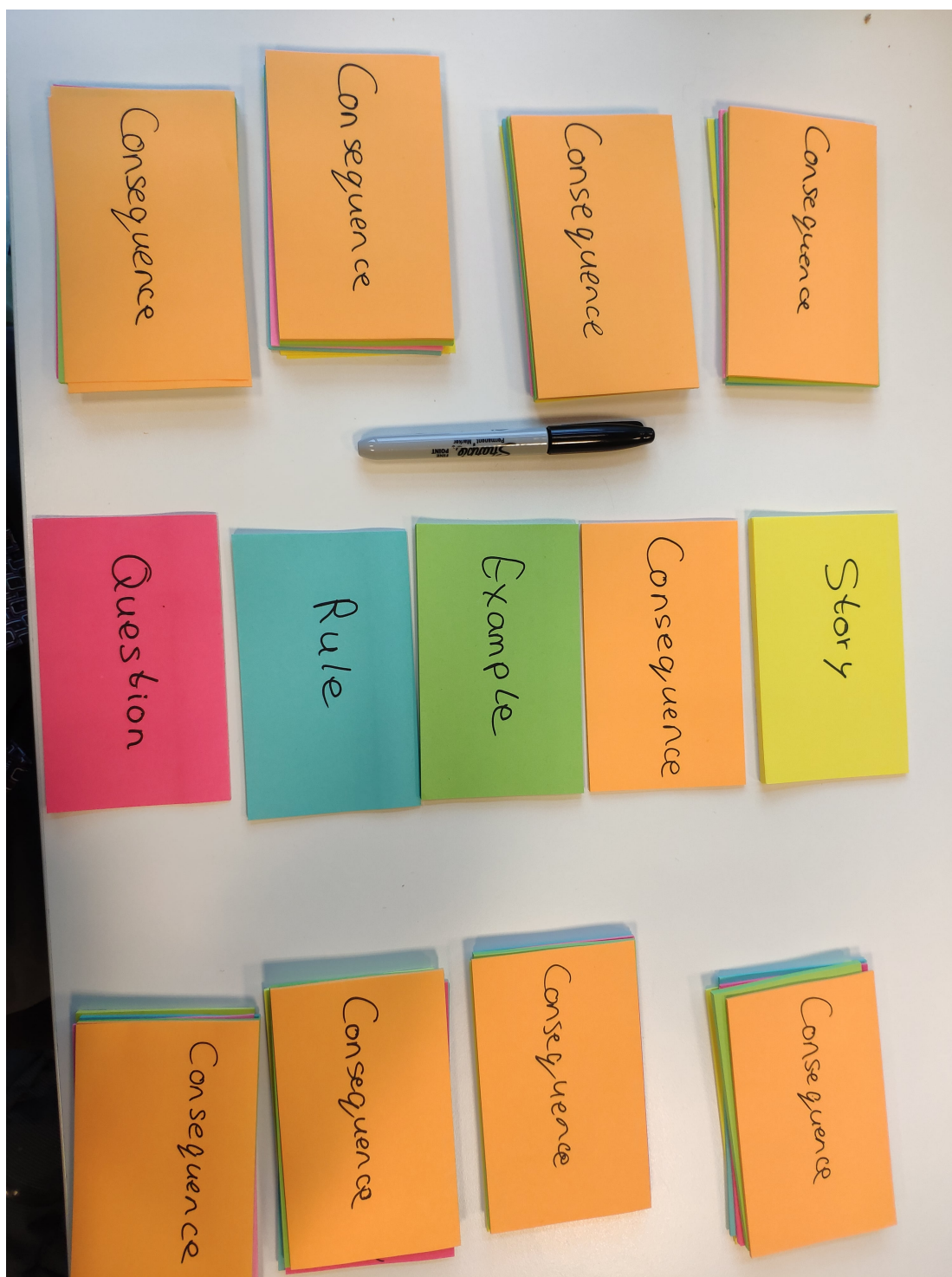


Figure. C.9: Handout - Post-its and markers

## Appendix C.5 Questionnaire Results

### C.5.1 Results per Technique - Combined

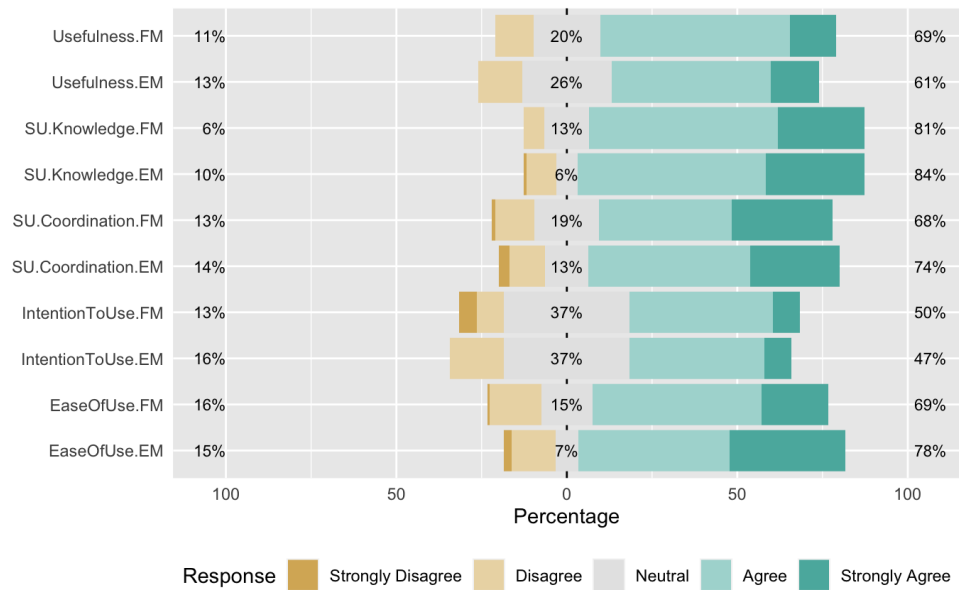


Figure. C.10: Responses per Aspect

## C.5.2 Results per Aspect

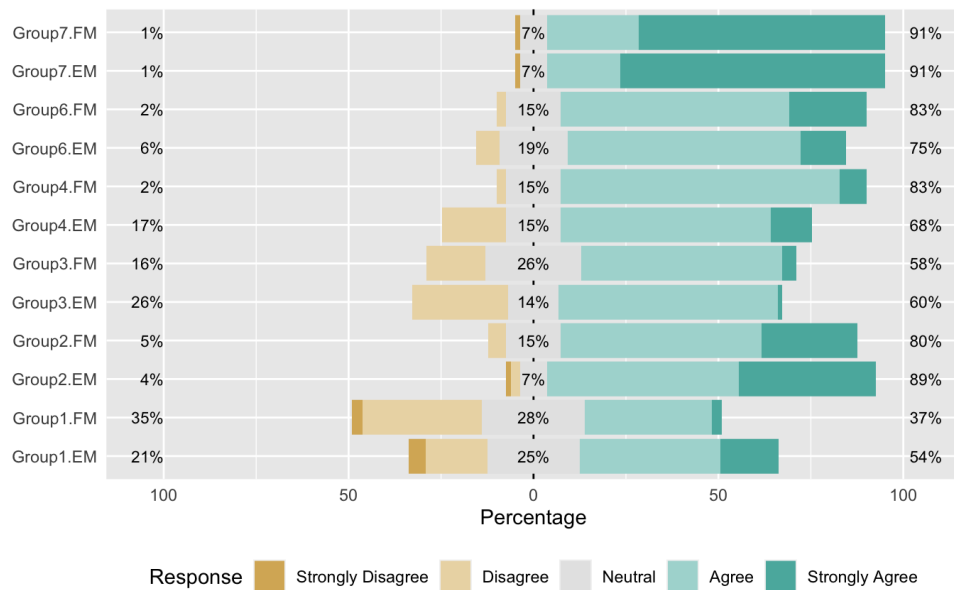


Figure. C.11: Group Results - All Aspects Combined

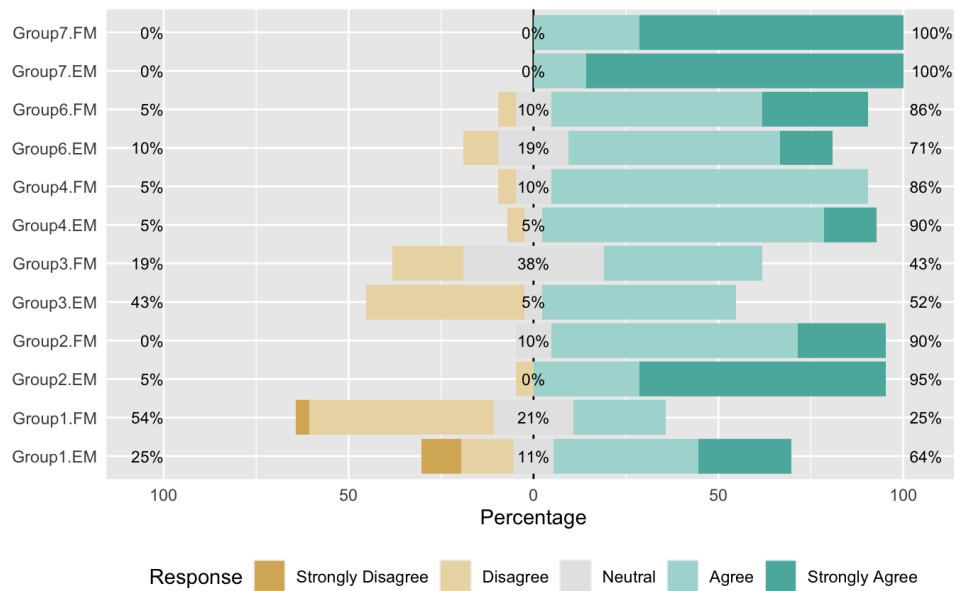


Figure. C.12: Group Results - Ease of Use

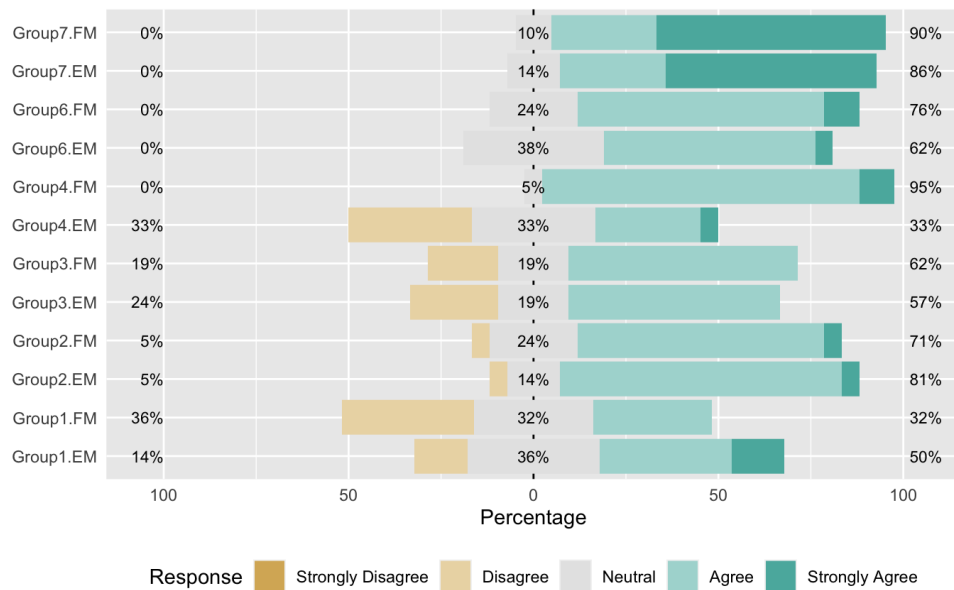


Figure. C.13: Group Results - Usefulness

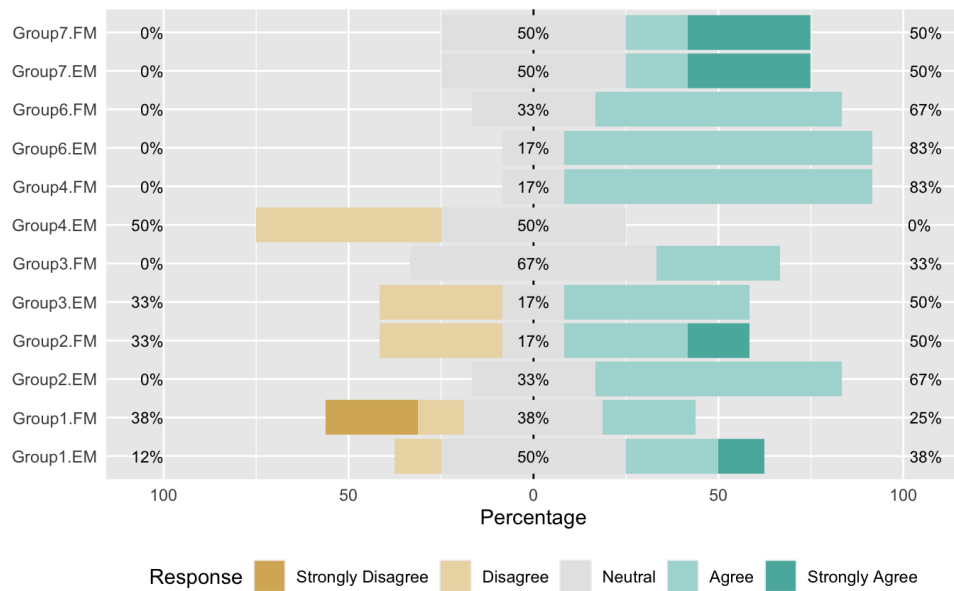


Figure. C.14: Group Results - Intention to Use



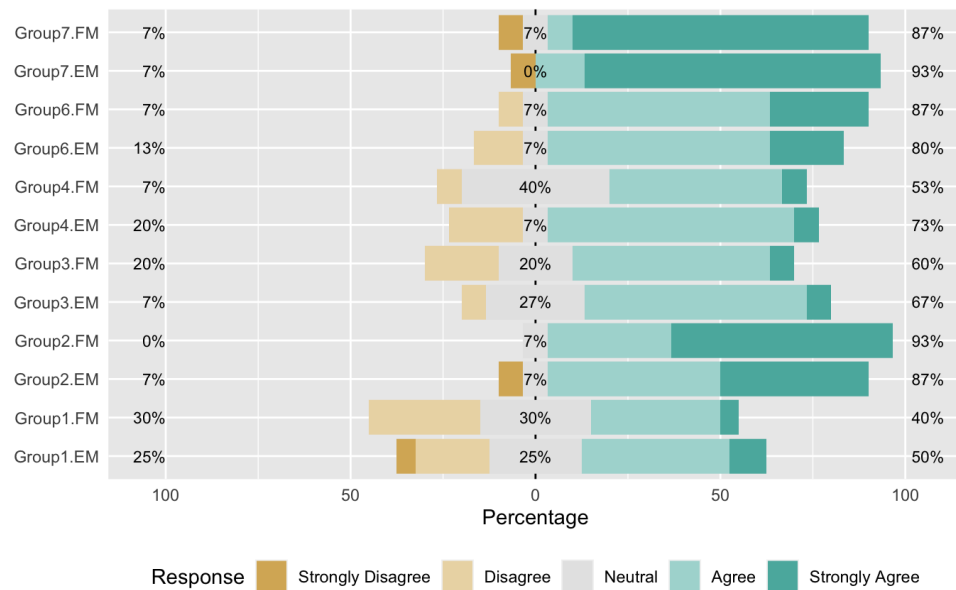


Figure. C.15: Group Results - SU Coordination

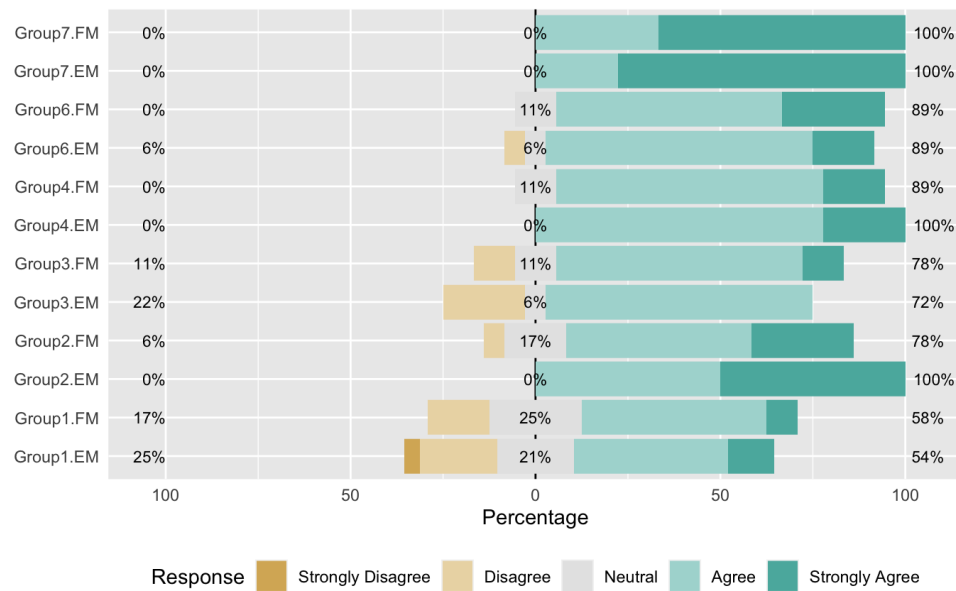


Figure. C.16: Group Results - SU Knowledge

### C.5.3 Results per Group

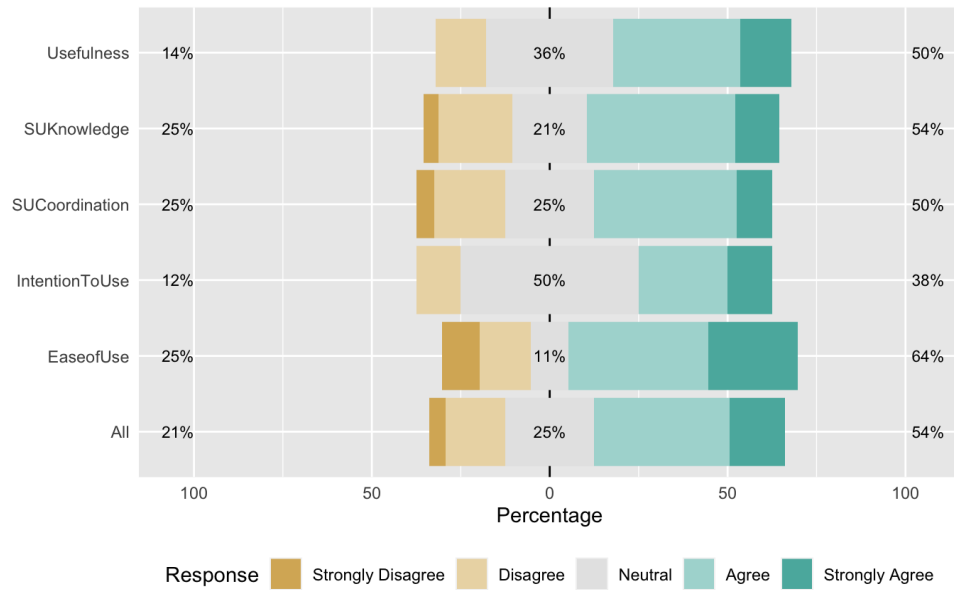


Figure. C.17: Group 1 - Example Mapping

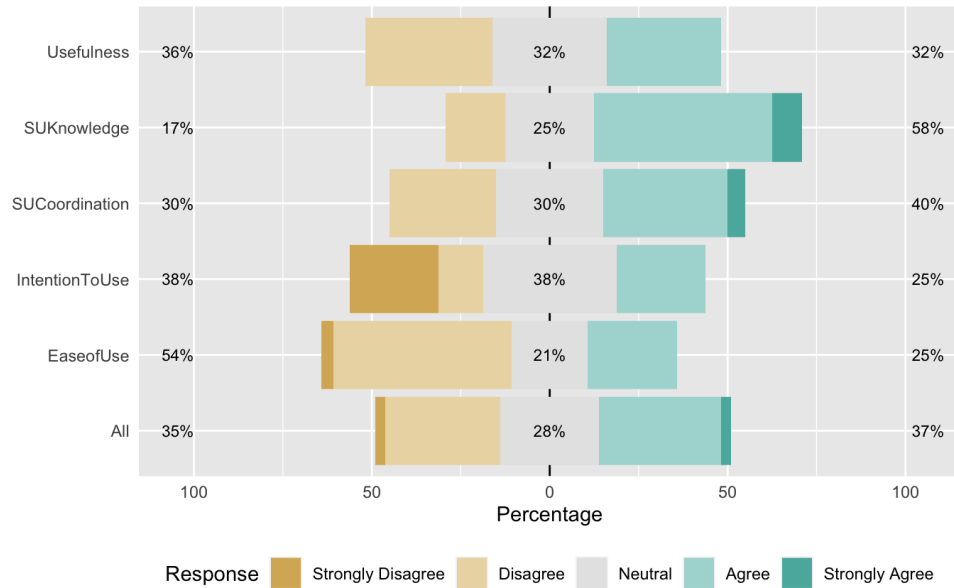


Figure. C.18: Group 1 - Feature Mapping

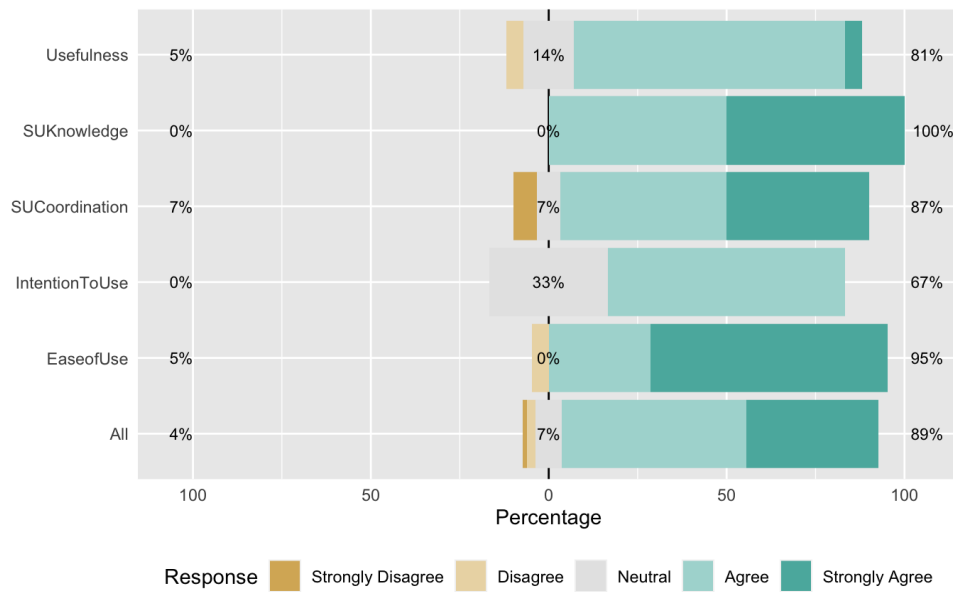


Figure. C.19: Group 2 - Example Mapping

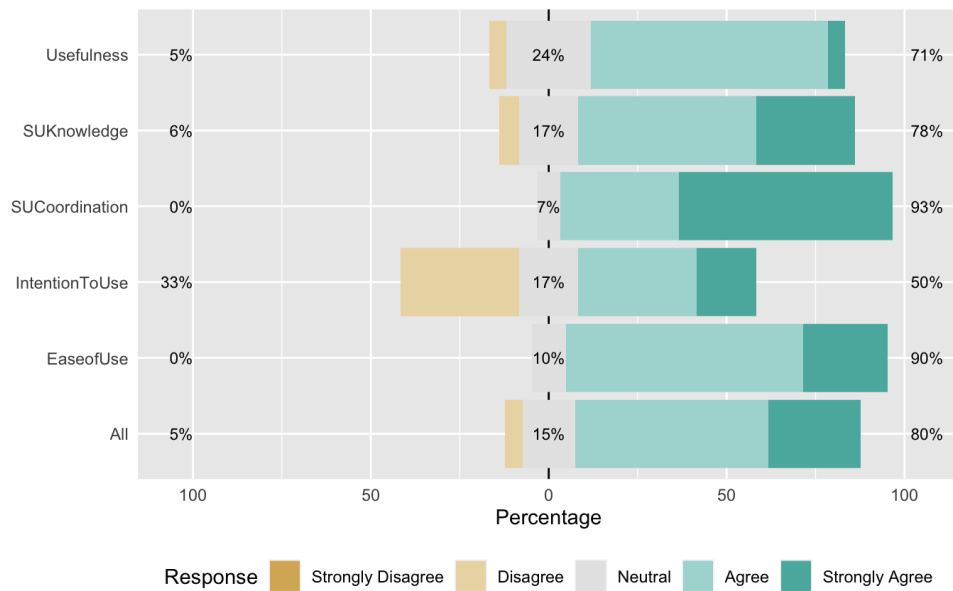


Figure. C.20: Group 2 - Feature Mapping

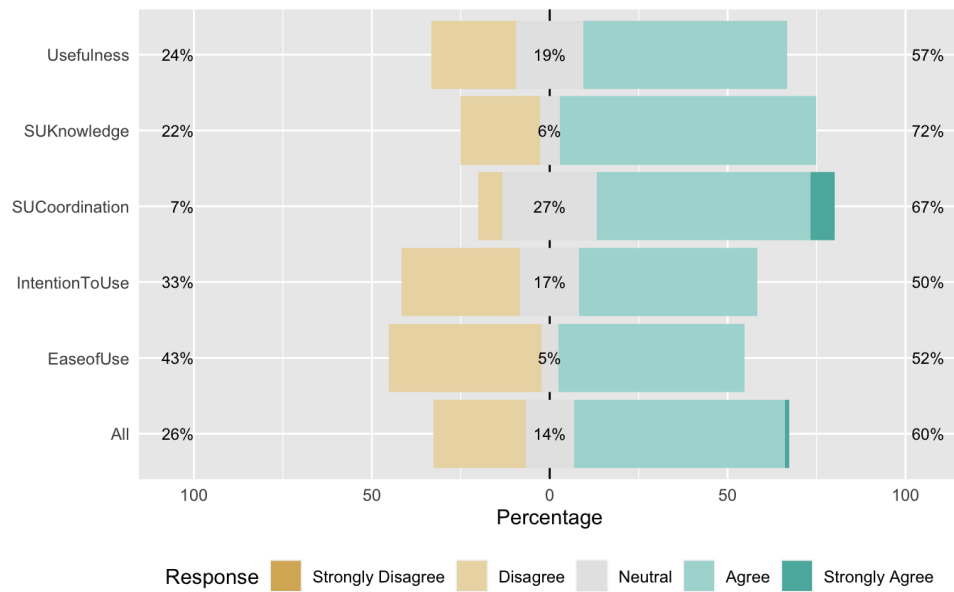


Figure. C.21: Group 3 - Example Mapping

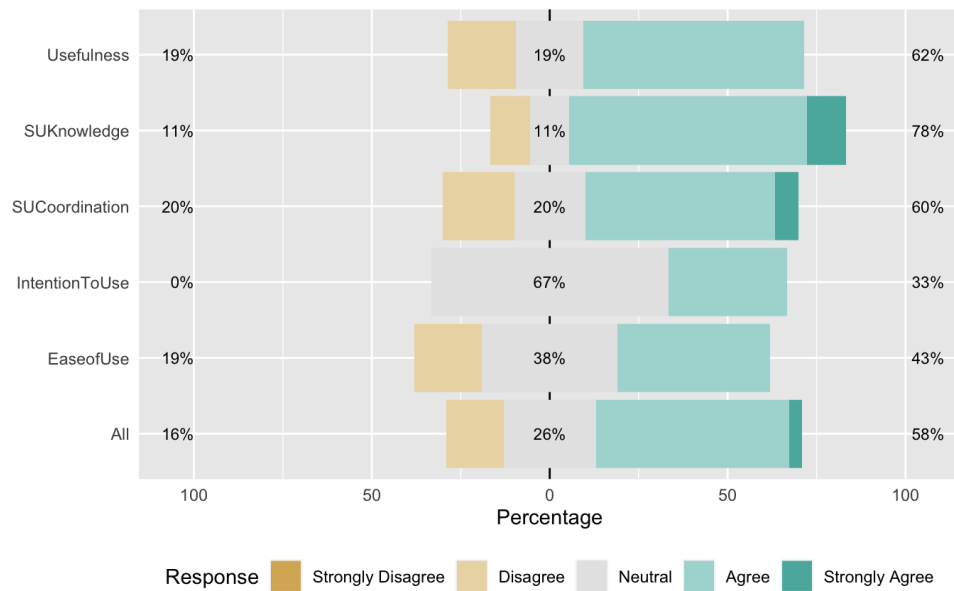


Figure. C.22: Group 3 - Feature Mapping

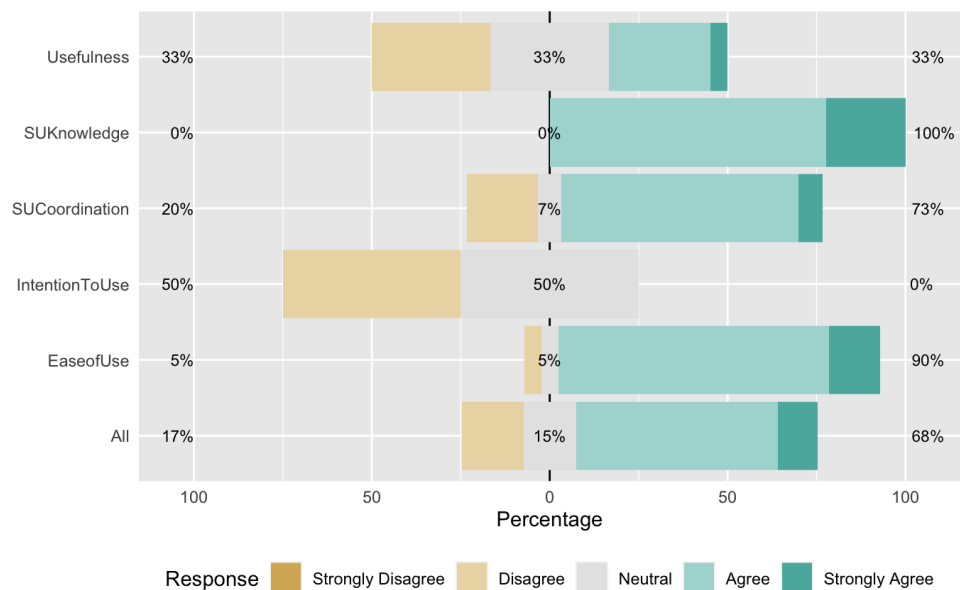


Figure. C.23: Group 4 - Example Mapping

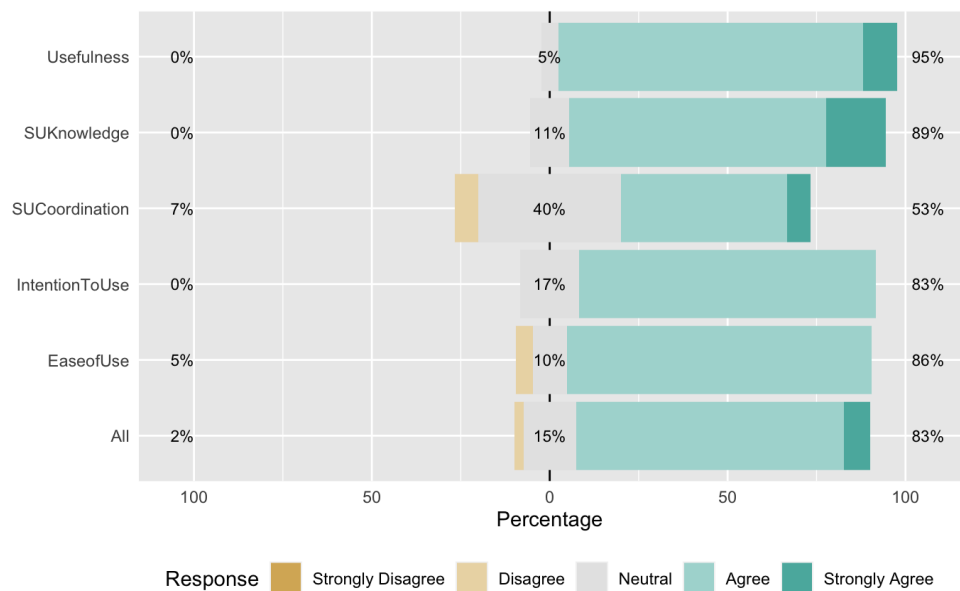


Figure. C.24: Group 4 - Feature Mapping

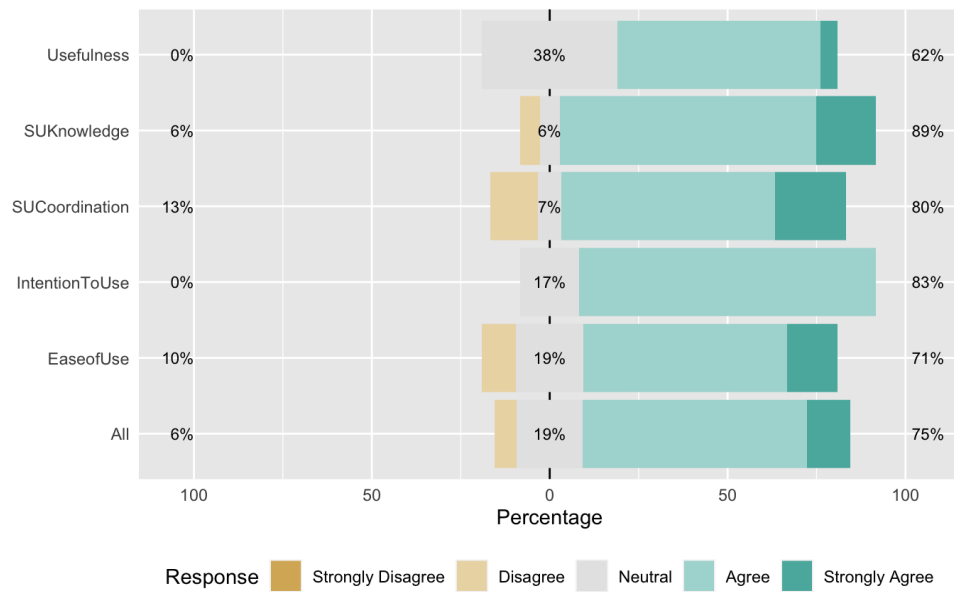


Figure. C.25: Group 6 - Example Mapping

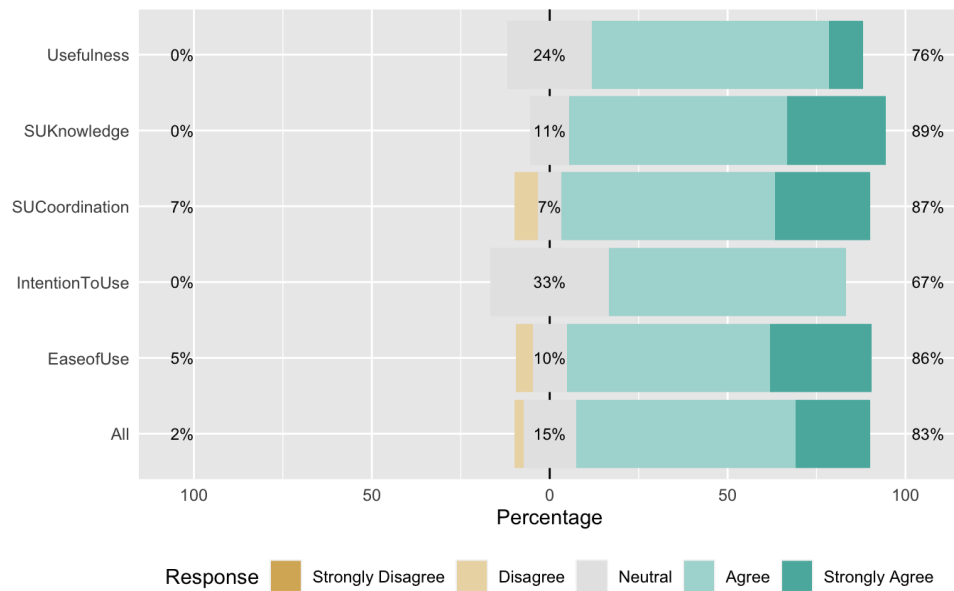


Figure. C.26: Group 6 - Feature Mapping

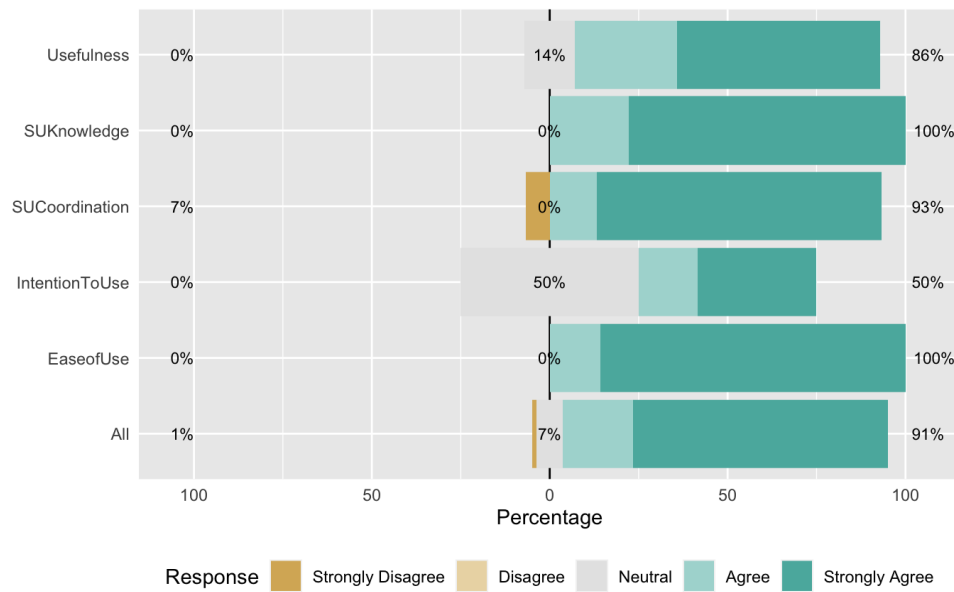


Figure. C.27: Group 7 - Example Mapping

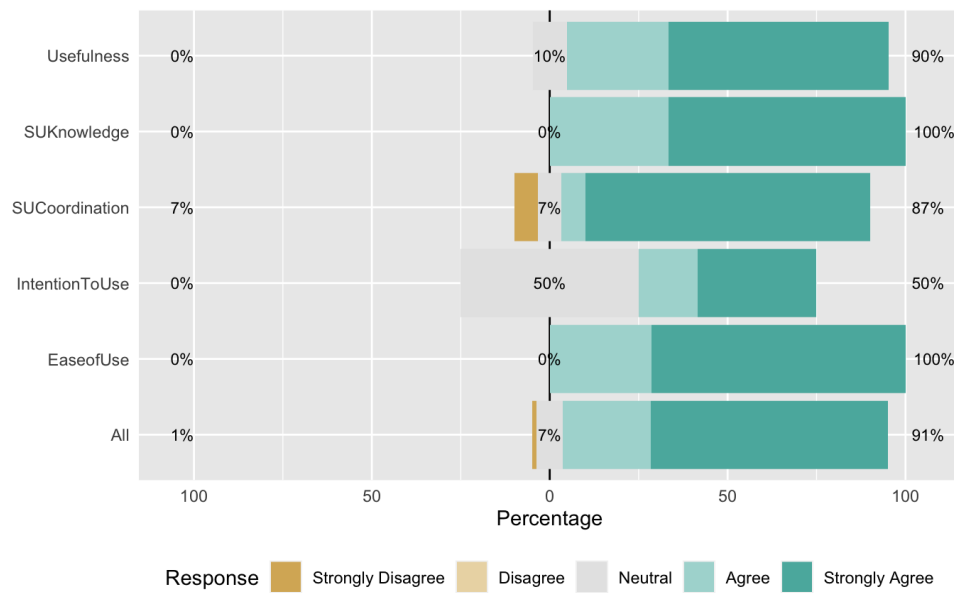


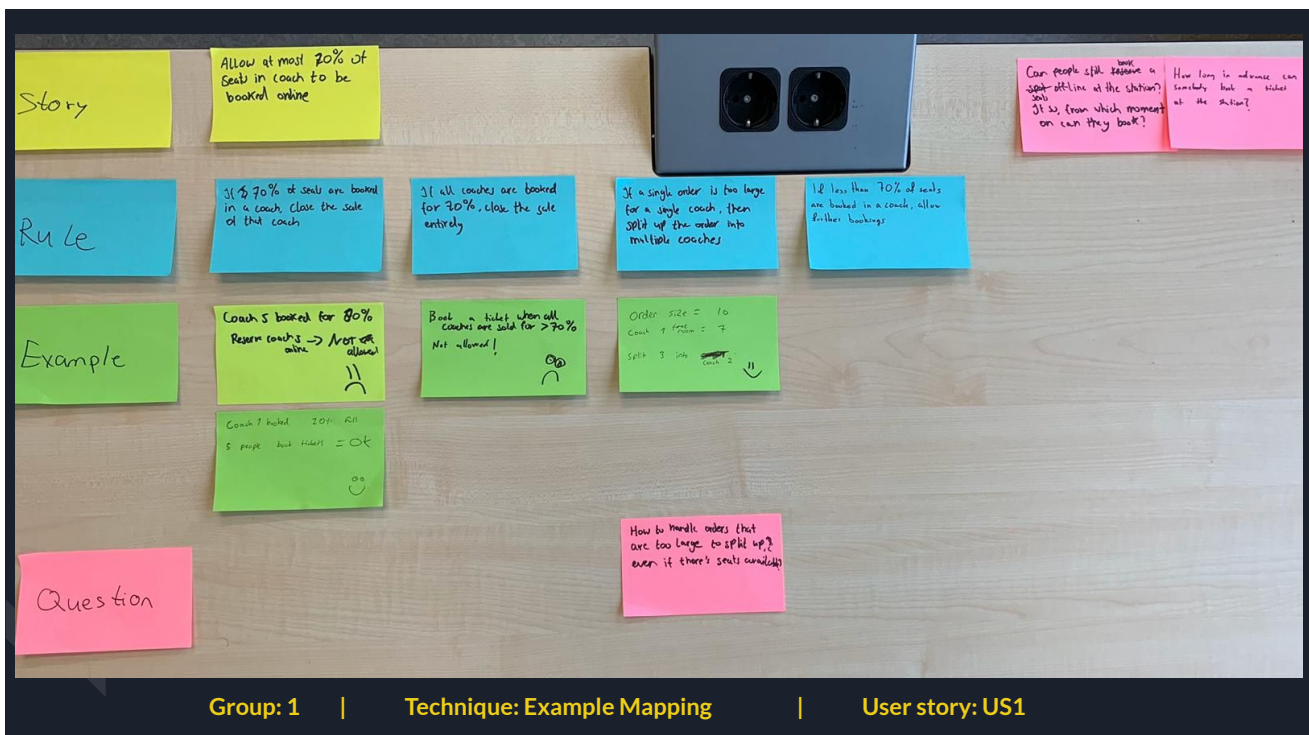
Figure. C.28: Group 7 - Feature Mapping

## Appendix C.6 TA Session Outputs



Three Amigo  
Session Outputs

Requirements Engineering 2019/2020  
Jasper Berends and Fabiano Dalpiaz  
24-02-2020



**Story**

Allow at most 70% of seats in coach to be booked online

**Rule**

If 70% of seats are booked in a coach, close the sale of that coach

If all coaches are booked for 70%, close the sale entirely

If a single order is too large for a single coach, then split up the order into multiple coaches

If less than 70% of seats are booked in a coach, allow further bookings

**Example**

Coach 5 booked for 80%  
Return coach's status → Not allowed

Coach 7 booked 20% Rn  
5 people book tickets = OK

Book a ticket when all coaches are sold for >70%  
Not allowed!

Order size = 10  
Coach 1 + Coach 2 = 7  
Split 3 into Coach 2

Can people split between a seat-off-line at the station? No  
If so, from which moment on can they book?

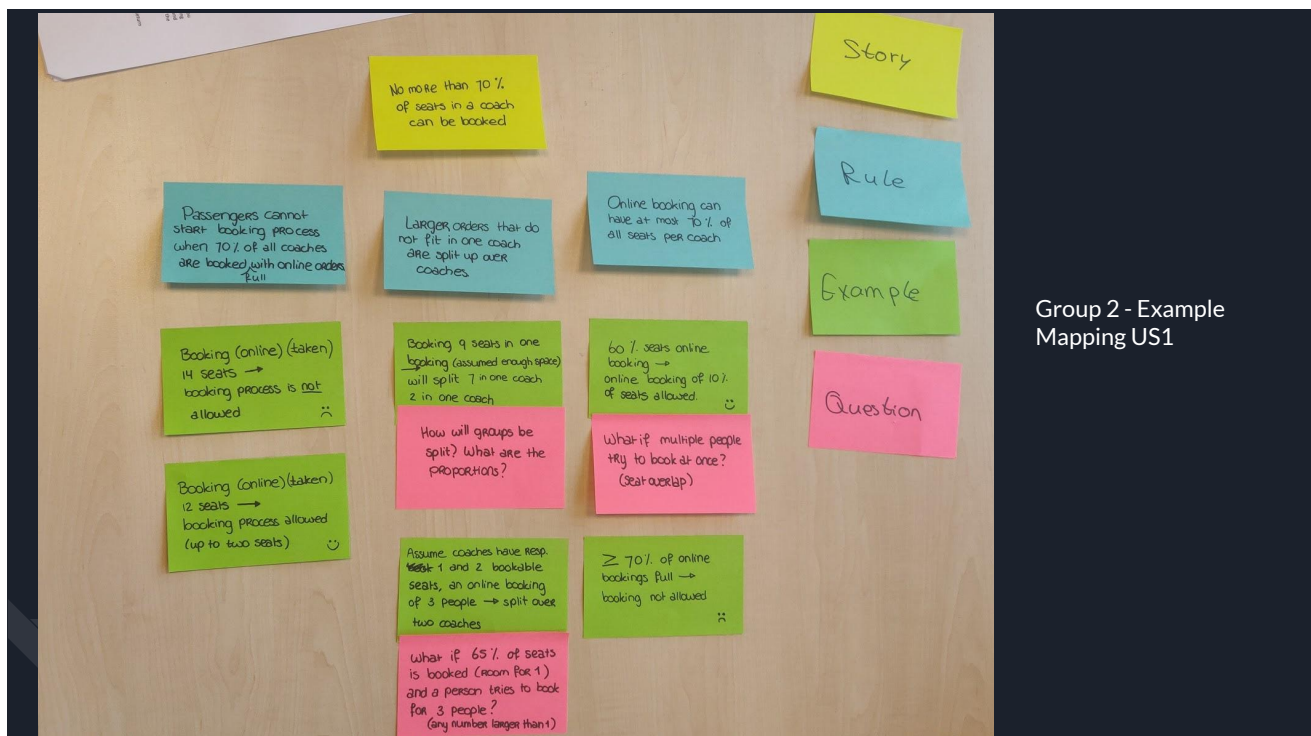
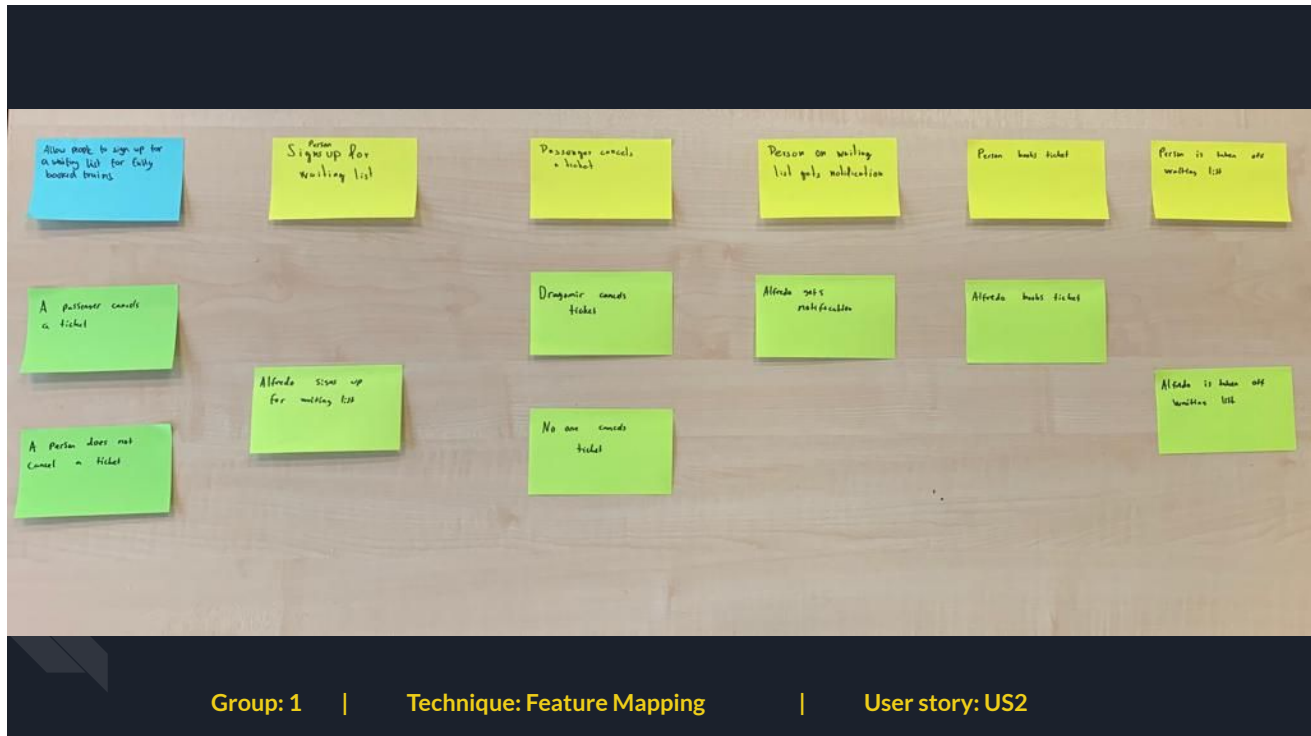
How long to advance can somebody book a ticket at the station?

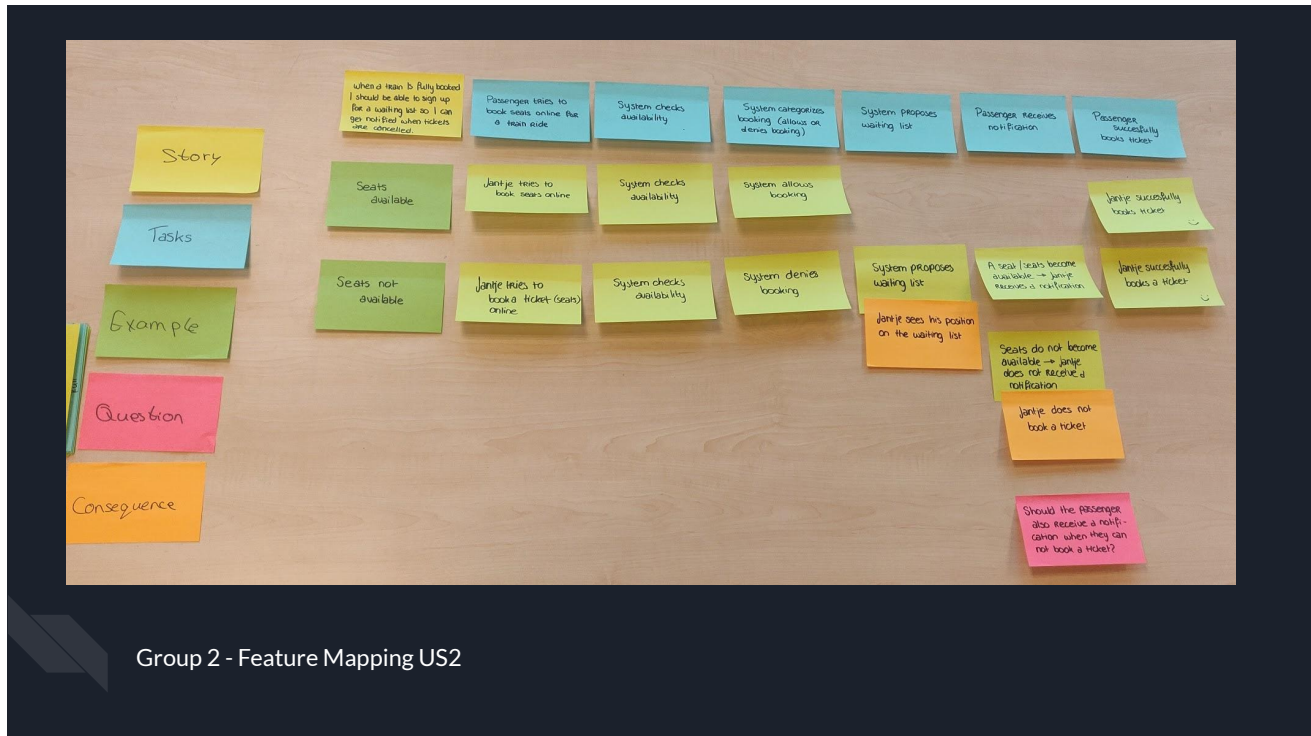
**Question**

How to handle orders that are too large to split up? even if there's seats available

Group: 1 | Technique: Example Mapping | User story: US1



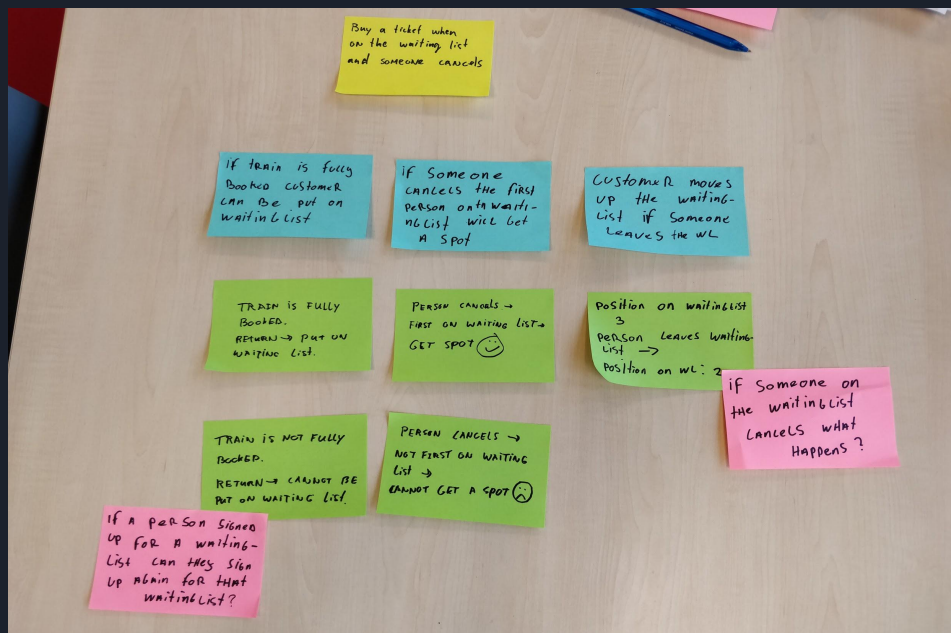




Group 2 - Feature Mapping US2



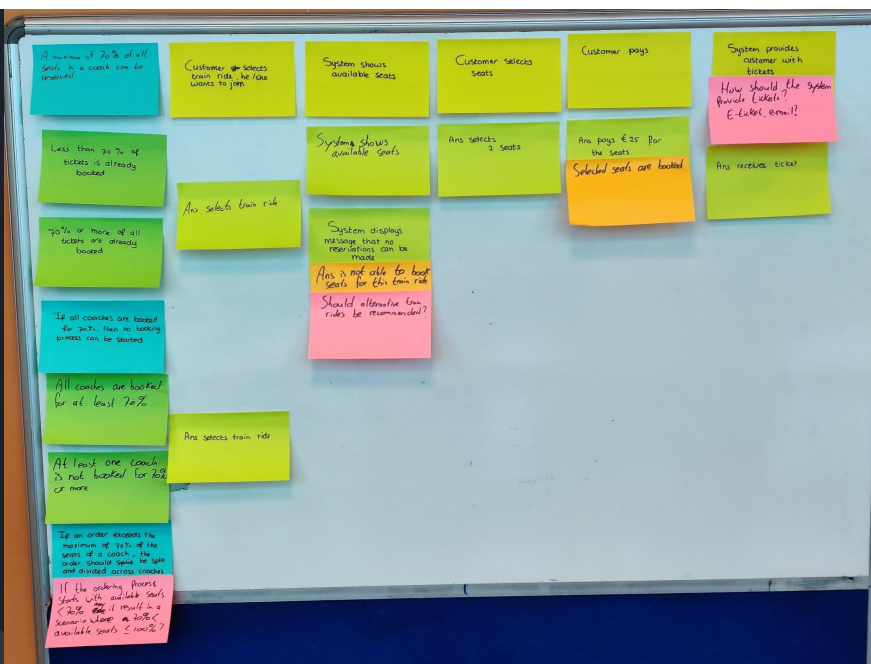
Group: 3 | Technique: Feature Mapping | User story: US1



Group: 3

Technique: Example Mapping

User story: US2

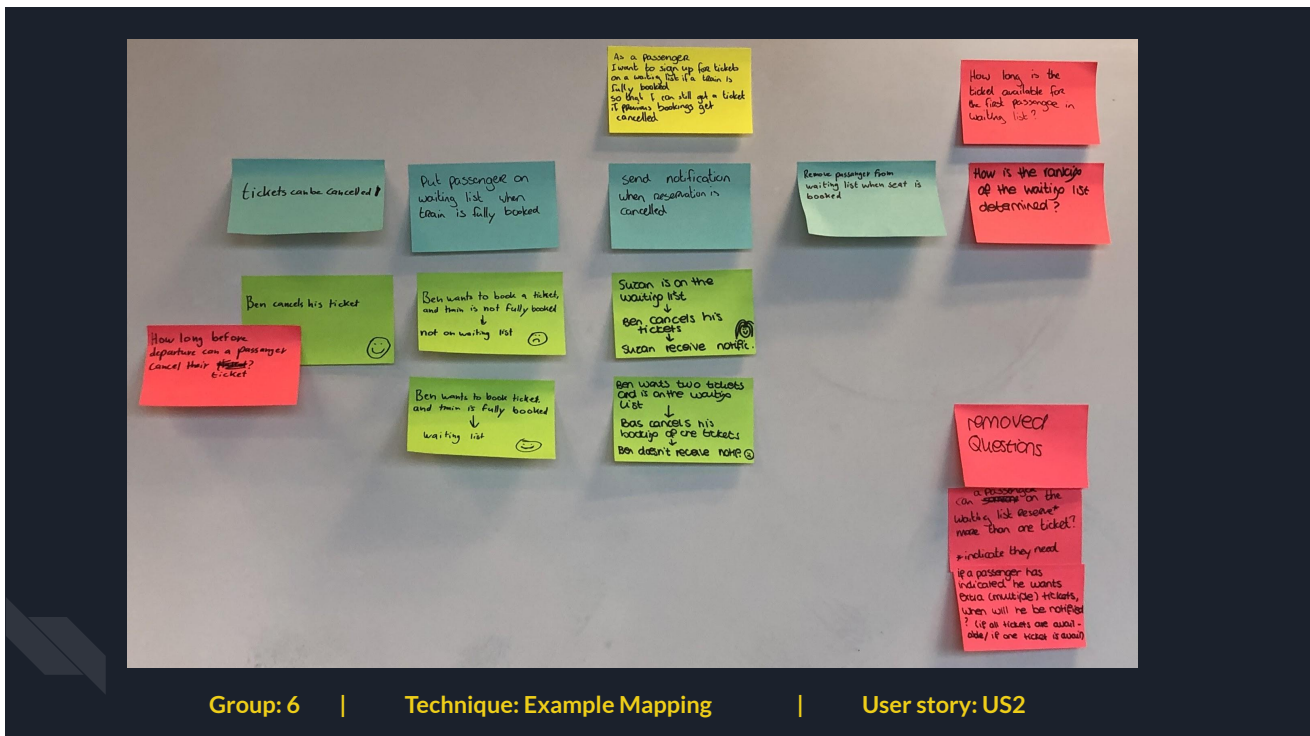
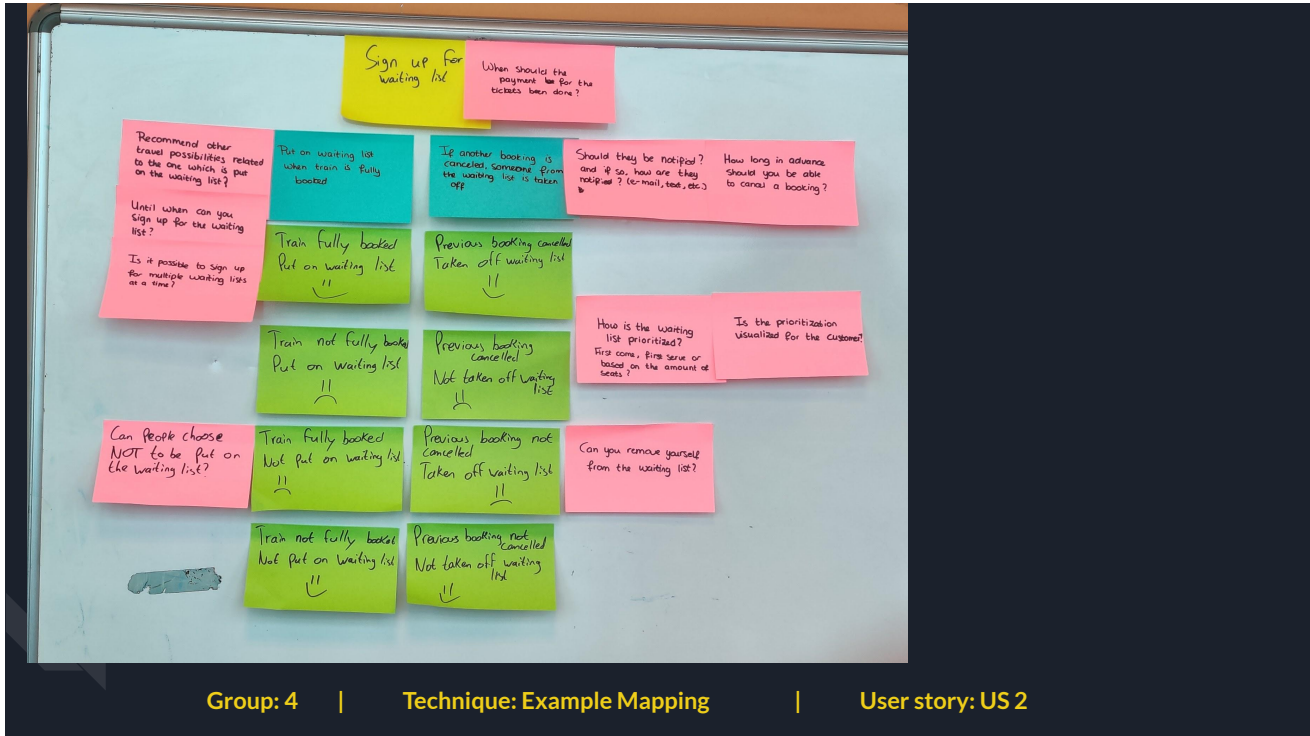


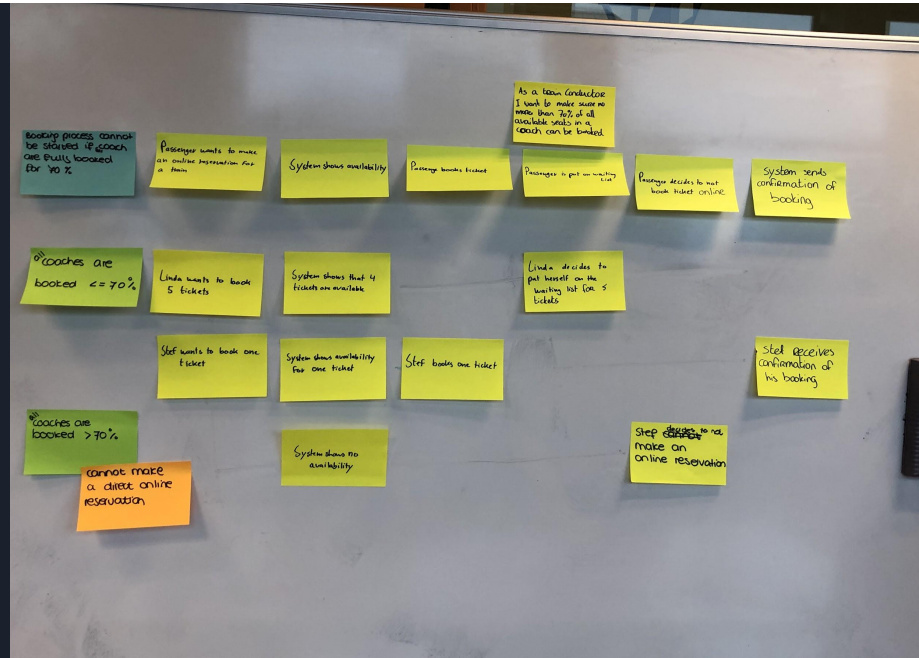
Group: 4

Technique: Feature Mapping

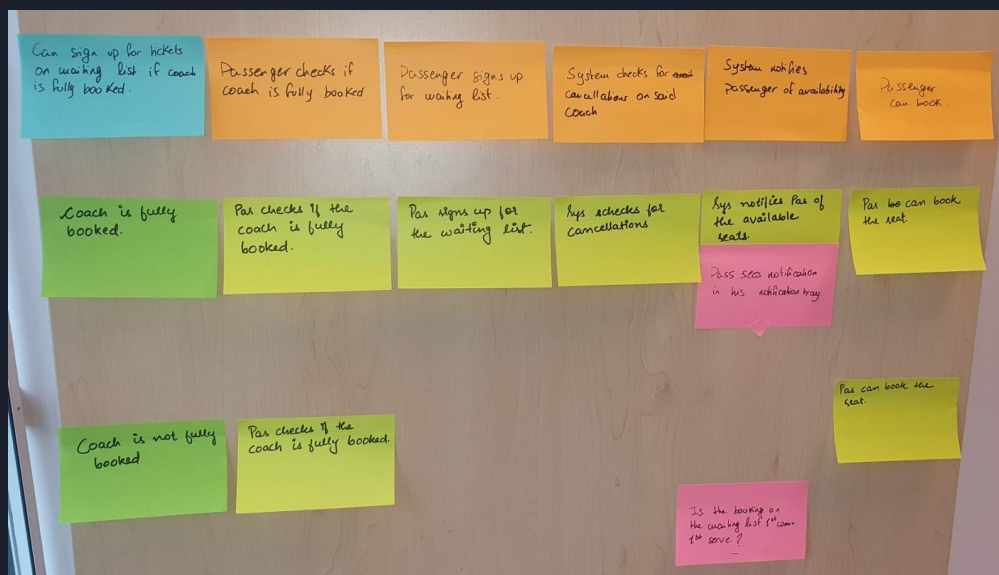
User story: US 1



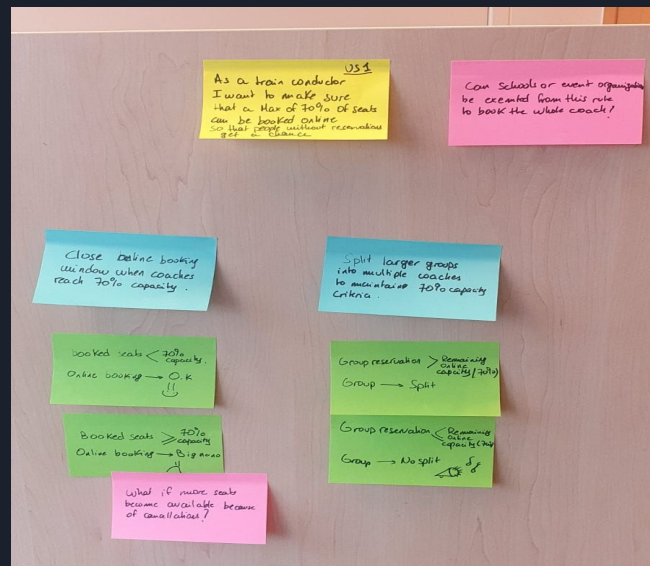




Group: 6 | Technique: Feature Mapping | User story: US1



Group 7 | US2 | Feature Mapping



Group 7 | US1 | Example Mapping

## Appendix C.7 Output Analysis

### C.7.1 US1 Aggregation

Rules | *Examples*

- **No more than 70% of seats on a coach can be booked**
  - *Positive scenario*
  - *Negative scenario*
- **Close sales when entire train is 70% booked**
  - *Positive scenario*
  - *Negative scenario*
- **Split up large orders**
  - *Example where order is split*
  - *Example where order is not split*

Questions (*italic questions are not crucial for this user story*)

- What if people try to order more than total number of train seats available? / What if orders are too large to split up?
- What if multiple people try to book at once?
- How are groups split up?
- *Can people book offline at the station?*
- *How long in advance can someone book a ticket at the station?*
- *How are tickets provided to passengers?*
- *Can school/organisations be exempted from this rule?*

### C.7.2 US2 Aggregation

#### Rules | *Examples*

- Add to waiting list when order does not fit train
  - *Scenario where someone is immediately put on waiting list*
  - *Scenario where order booked*
- Add everyone to waiting list when train is fully booked
  - *Scenario where put on waiting list*
- Notify/book people on waiting list when seats become available
  - *Scenario where passenger is notified*
  - *Scenario where order is still too big*
- People on waiting list move up positions when others get out
  - *Scenario where someone is moved up*
- Remove a passenger from the waiting list when a ticket is booked

Questions (*italic questions are not crucial for this user story*)

- How many people can be on the waiting list?
- Should order be put on waiting list if it does not fit one coach?
- Should the second person on the waiting list be notified if order of first person on waiting list is too big or should seats still be reserved for first person?
- How are people notified when seats are available?
- How is the waiting list prioritised?
- Can people choose not to be put on the waiting list?
- Can you remove yourself from the waiting list?
- *How long is a ticket available on someone from the waiting list?*
- *Until when can you sign up for the waiting list?*
- *Recommend other travel possibilities?*
- *Can you sign up for multiple waiting lists at once*
- *How long in advance can tickets be cancelled?*



# Chapter D | Case Study - Fizzor

## Appendix D.1 Correlation Matrices

Correlation	Perceived Ease of Use	Perceived Usefulness	Intention to Use	SU - Coordination	SU - Shared Knowledge
Perceived Ease of Use		0.51	-0.04	0.89	0.14
Perceived Usefulness	p = 0.1989		0.08	0.52	0.05
Intention to Use	p = 0.9173	p = 0.8596		-0.38	0.80
SU - Coordination	p = 0.0029	p = 0.1857	p = 0.3464		0.04
SU - Shared Knowledge	p = 0.7328	p = 0.9079	p = 0.0171	p = 0.9213	

Table D.1: Pearson's Correlation between aspects - Long Questionnaire

Correlation	Perceived Ease of Use	Perceived Usefulness	SU - Coordination	SU - Shared Knowledge
Perceived Ease of Use		0.57	0.36	0.39
Perceived Usefulness	p = 0.0259		0.10	0.46
SU - Coordination	p = 0.1818	p = 0.7140		0.37
SU - Shared Knowledge	p = 0.1485	p = 0.0812	p = 0.1808	

Table D.2: Pearson's Correlation between aspects - Session Questionnaire

## Appendix D.2 Participant Results

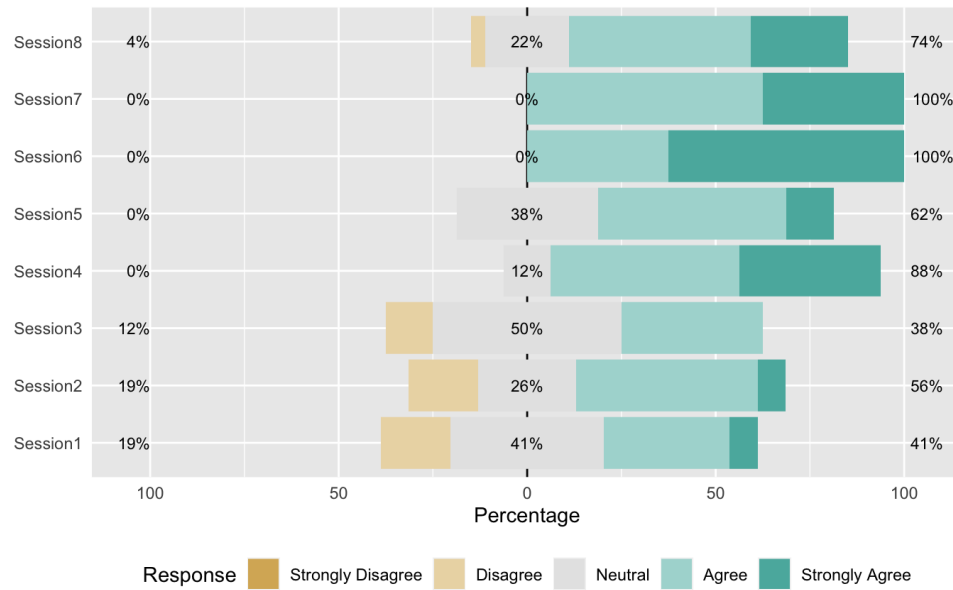


Figure. D.1: Results Person 1

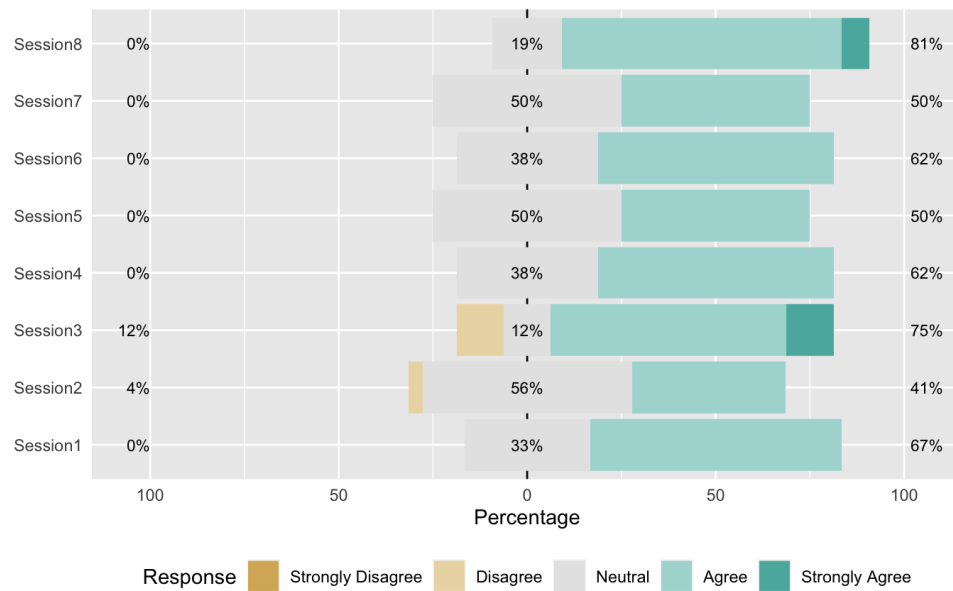


Figure. D.2: Results Person 2

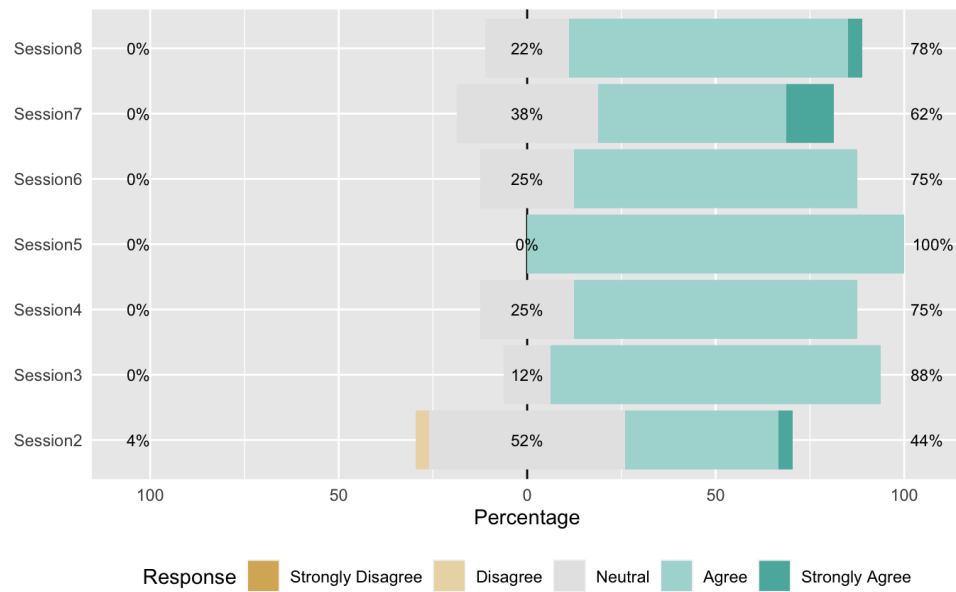


Figure. D.3: Results Person 3

## Appendix D.3 Session Results

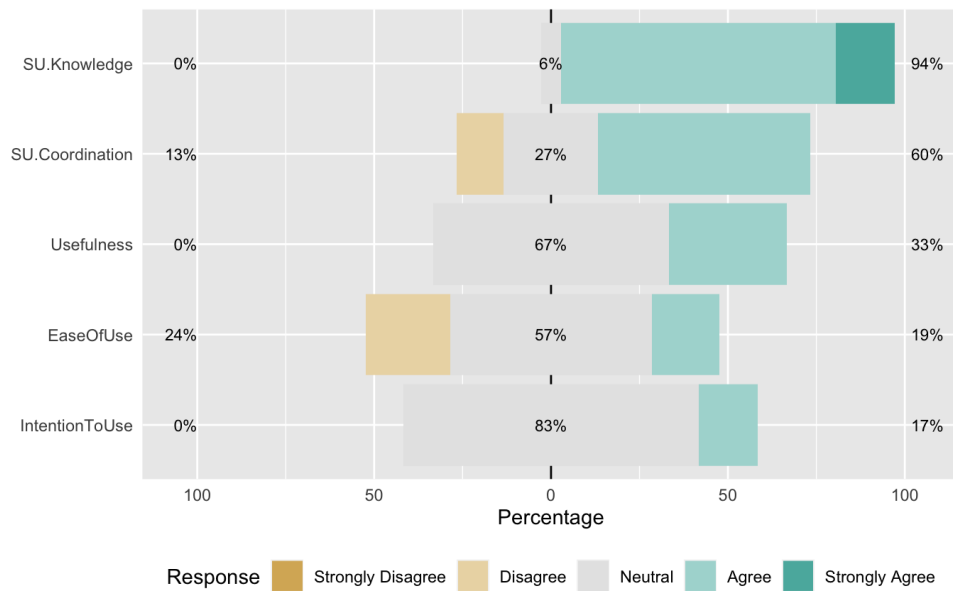
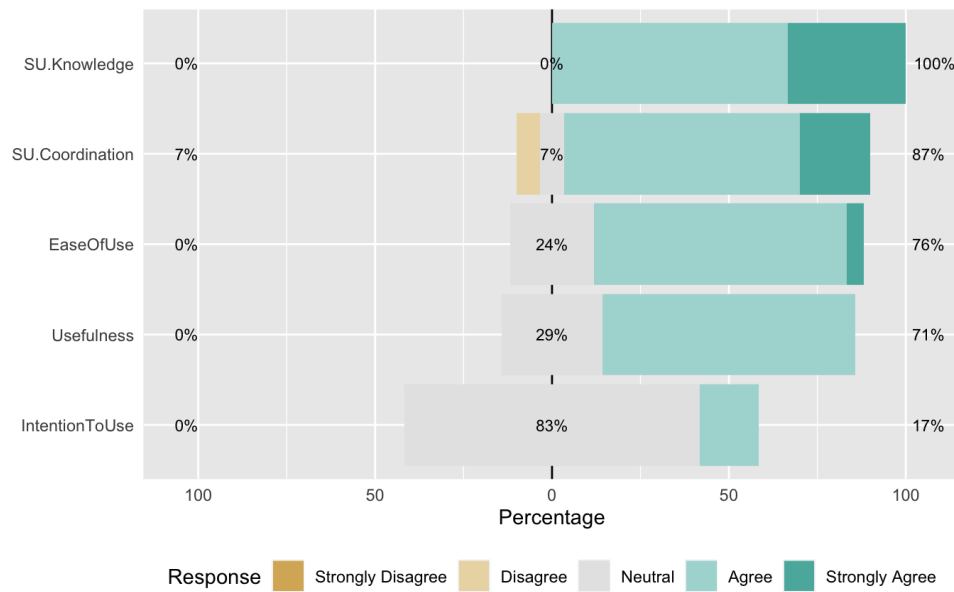


Figure. D.4: Results Session 2

**Figure. D.5:** Results Session 8

# Chapter E | Case Study - Pension Manager Example Mapping

## Appendix E.1 Correlation Matrices

Correlation	Perceived Ease of Use	Perceived Usefulness	Intention to Use	SU - Coordination	SU - Shared Knowledge
Perceived Ease of Use		0.66	0.68	0.53	0.56
Perceived Usefulness	p = 0.0759		0.68	0.61	0.69
Intention to Use	p = 0.0650	p = 0.0641		0.09	0.09
SU - Coordination	p = 0.1742	p = 0.1050	p = 0.8316		0.93
SU - Shared Knowledge	p = 0.1521	p = 0.0584	p = 0.8336	p = 0.0009	

Table E.1: Pearson's Correlation between aspects - Long Questionnaire

Correlation	Perceived Ease of Use	Perceived Usefulness	SU - Coordination	SU - Shared Knowledge
Perceived Ease of Use		0.67	0.32	0.68
Perceived Usefulness	p = 0.0667		0.79	0.80
SU - Coordination	p = 0.4400	p = 0.0195		0.59
SU - Shared Knowledge	p = 0.0636	p = 0.0181	p = 0.1257	

Table E.2: Pearson's Correlation between aspects - Session Questionnaire