Refining User Stories via Example Mapping: An Empirical Investigation

Jasper Berends
Info Support
Veenendaal, the Netherlands
Jasper.Berends@Infosupport.com

Fabiano Dalpiaz
Utrecht University
Utrecht, the Netherlands
F.Dalpiaz@uu.nl

***Abstract*—New techniques for managing, specifying, and analyzing requirements in software engineering projects are frequently presented by consultants and agile trainers. However, the effectiveness of these techniques is not evaluated in a rigorous manner, leaving practitioners with the question "Will it work in our company?" In this paper, we investigate the performance of a user story refinement technique named Example Mapping (EM). This is a time-boxed workshop in which people from different disciplines work collaboratively in order to refine, or clarify, a user story with the use of examples. The creation of such examples is intended not only to obtain a more precise specification, but also and mostly to achieve shared understanding on the user story to develop among the team members. We investigate the performance of EM via two longitudinal case studies. To enable a rigorous validation of EM, we first define the Refinement Evaluation Tool (RET), a survey-based measurement instrument that extends the Method Evaluation Model with questions that cover the shared understanding dimension. The results from our case studies show that EM contributes to the shared understanding within a team; certain conditions are necessary: the user stories should not be too small-sized. We also investigated the learning effect for EM; our data indicates that two sessions are generally necessary for the team members to use the technique effectively.**

***Index Terms*—Example Mapping, Shared Understanding, Refinement, User Stories**

I. Introduction

Many of the recent innovations in software engineering have been proposed and made popular by industry professionals and trainers. The manifesto for agile software development [1] is a prime example, alongside general techniques such as Scrum, Kanban and DevOps, as well as more specific approaches such as test-driven development [2], behavior-driven development such as the Gherkin approach [3], and user stories [4].

These industrial innovations are made necessary by the increasing need for speed in the software development landscape [5] and by the fierce competition. However, they lack a rigorous empirical validation that a practitioner may consider when deciding whether to switch to a new technique, which one to choose, and what are the contextual factors that make a technique suitable for the situation at hand.

We study one specific innovation that pertains to requirements engineering (RE): Example Mapping (EM) [6]. This is a short workshop in which people from different disciplines gather in order to refine a software increment (a requirement), through the definition of examples of how the increment should function. EM adopts the principles of the Three Amigos: an increment should be studied from multiple perspectives, i.e., business, development, and testing. An EM session, therefore, may involve a product owner, a lead developer, and a quality assurance member who need to agree on the criteria for implementing and validating a user story.

Through two longitudinal case studies in two organizations, we investigate the performance of EM as a technique to refine a user story. While doing so, we focus on one crucial feature of EM: its ability to raise shared understanding (SU) among the team members involved with that user story. At the end of an EM session, the involved team members should converge to a high degree on what that user story entails.

Note that shared understanding has been acknowledged as an important factor in RE [7], [8], especially when software development adopts agile methodologies. Informal and frequent communication has been considered a crucial part of Agile Requirements Engineering (RE) [9]. The agile manifesto [1] makes this clear through its principles "individuals and interactions over processes and tools" and "working software over comprehensive documentation."

To provide a thorough answer to the *how does it work?* question, we first construct the Refinement Evaluation Tool (RET): a performance measurement instrument that allows evaluating refinement techniques. The RET builds on and extends the well-known Method Evaluation Model [10] by adding questions that allow measuring the shared understanding within a team.

In our case studies, we use the RET as a probing tool. In addition to measuring the performance in a quantitative fashion, we focus on two contextual factors: (i) the characteristics of a user story that makes it suitable for use within an EM session; and (ii) the learning curve for a team to master the use of EM within a session effectively.

In this paper, we make two contributions to RE research and practice:

- We present two longitudinal studies on the effectiveness of EM sessions that zooms in on factors that affect EM's suitability and its learning curve;
- We propose the RET tool as a practical, yet theoretically founded, instrument for studying the effectiveness of requirements refinement techniques.

The rest of the paper is structured as follows. In Section II, we present an overview of EM. In Section III, we discuss

the notion of SU and its importance in RE. We introduce the RET tool in Section IV. We outline the research method in Section V. Section VI reports on the case studies, while Section VII presents a cross-case analysis. Finally, we conclude and present future directions in Section VIII.

II. Example Mapping

In this research, we evaluate Example Mapping (EM). This is a technique for organizing TA sessions that was introduced by Matt Wynne in 2015 [6]. This is an example of an industrial innovation which, besides inspiring agile trainers, has triggered the creation of variants, such as Feature Mapping [11].

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. Organizing 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" [12].

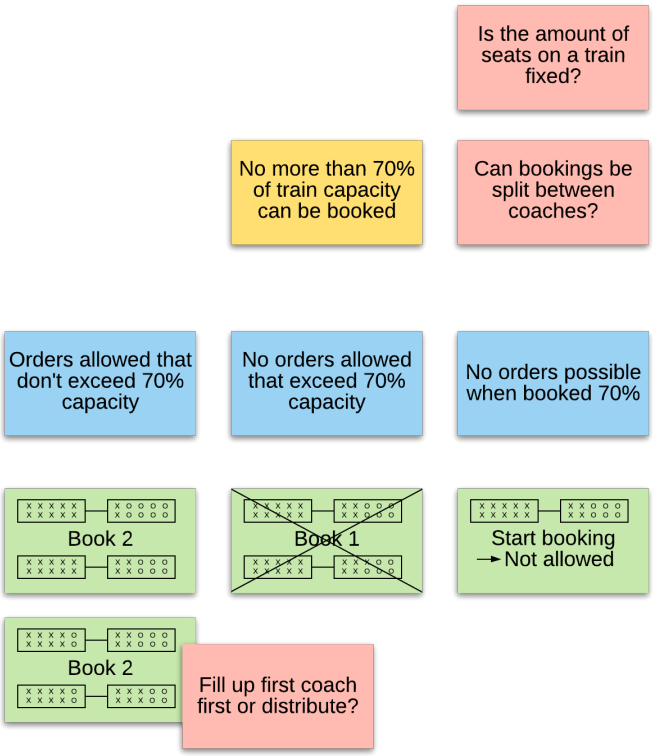


Fig. 1. Example Mapping output for a train reservation system.

Fig. 1 shows an example of EM output concerning a train reservation scenario. Four different types of cards can be used:

- *Story*: a yellow card[1] that represents a user story, often in an abbreviated form. In the figure, the story states that in the system to-be, no more than 70% of the train capacity can be booked (so as to allow people to buy tickets at the train station);
- *Rule*: a blue card that refines the condition stated by the story while summarizing a bunch of examples. For example, the user story in the figure is refined to three rules that state that (i) orders should be allowed when occupancy is below 70% capacity; (ii) no orders should be possible when this would increase the occupancy above 70%; and (iii) no orders should be possible when occupancy is already above 70%;
- *Example*: a green card that presents a concrete situation that illustrates the rule. In the figure, the example on the left shows that booking is possible when it would lead to occupancy of exactly 70%; the one in the middle shows that a booking where 29 out of 40 spots is forbidden, while the example on the right shows—with a mix of text and picture—that when 14/20 spots are already booked, new bookings should not be allowed;
- *Question*: a red card that represents questions that are revealed through the user story refinement session. These may need to be discussed in further meetings. For example, in Figure 1, three questions arise: Is there a fixed amount of seats on a train? Can a booking be split across coaches? Should the first coach be filled up first, or can bookings be distributed across coaches?

The process of an EM session is rather free format. The starting point is selecting the user story, from the backlog, that will be refined during the session. Then, if a business representative such as a stakeholder, product owner or business analyst is present, they may first introduce the story with some initial information on what it is about. After picking out the user story to refine and possibly giving some initial information, the EM session can start. There is no prescribed order for the writing of the cards: while one team may start from examples and then generalize them into rules, another may take a more top-down approach.

III. Shared Understanding

Shared understanding (SU) is a term that has become popular in agile software development, and previous research has shown that it increases team performance and software quality [13]–[15]. Still, there are numerous variants of this term, with different definitions. As TA session techniques consider shared understanding of a user story to be a key benefit, it is important to define what exactly is shared understanding, how it is built, and what are enablers and inhibitors.

According to Glinz and Fricker [7], multiple situations of SU exist: (i) true implicit shared understanding, (ii) true explicit shared understanding, (iii) false implicit shared understanding, and (iv) false explicit shared understanding. These situations define both the way information is shared as well as its correctness. Implicit shared understanding regards non-specified knowledge, whereas explicit shared understanding mostly regards written down concepts.

[1]The card coloring is recommended, not prescribed. Also, there is no fixed structure for the contents of the cards.

A. Research embedding of shared understanding

Cooke *et al.* [16] have investigated shared cognition and team cognition in cognitive sciences; they argue that the focus should be on processes and interactions at a team level instead of on an individual level. They explain the difference between team cognition and shared cognition: the former is about a group as a whole, whereas the latter refers to individual cognition. They note several issues regarding the definition of shared cognition in relation to team cognition. Therefore, they propose Interactive Team Cognition (ITC) as an alternative theory to shared or team cognition. In 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 [17], [18]. In this research, we investigate an interactive technique. Therefore, considering the interaction itself as part of SU will be essential to our research.

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 make this assumption: team cognition can be both more or less than that of the sum of the individuals' cognition.

Another important implication about 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 expected to increase team cognition.

Wildman *et al.* [17] conducted a literature review on team cognition across multiple disciplines. They determine five 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 as the similarities of mental models of members of a team and the accuracy of those mental models. The definition of transactive memory systems is two-fold: it regards both the knowledge of individuals in a group, as well as the processes used to "encode, store, and retrieve that knowledge" [19]. The second part of this definition corresponds with the focus on interaction of ITC.

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. However, situation awareness overlaps with team mental models [17].

Strategic consensus is most studied in literature concerning top management teams and is defined as a "team's shared understanding regarding the high-level strategic goals of the team or organization" [17]. Albeit a different focus than the other research domains, shared understanding is generally considered as the degree of agreement or sharedness between individuals [20], making its concepts 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 considers 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.

In their research, Wildman and colleagues have come up with context-dependent recommendations on how to measure team cognition [17]. In order to select the appropriate technique to analyze team cognition, the first question to answer is if team cognition is conceptualized 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. Observation and self-reported perceptions of team cognition are the two prevailing options.

B. Shared understanding in this research

For the purpose of this research, where a TA session technique is 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, SU should include the interaction aspect that is described in theories such as ITC.

However, it cannot focus solely on the interaction. Although creating shared understanding may be the primary goal of a TA session, in the end, the user story should 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 focuses on the dispersion of knowledge, rather than on knowledge that all team members possess, while TA sessions put emphasis on sharing information with team members in order to all get the same understanding.

Lewis has created a list of questions used to measure transactive memory systems [21]. He distinguishes three different dimensions within transactive memory systems: (i) knowledge specialization, (ii) credibility, and (iii) coordination. Knowledge specialization refers to the dispersion of knowledge. This is a general view within 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.

As such, we define SU 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.

IV. The Refinement Evaluation Tool

We describe the *Refinement Evaluation Tool (RET)*, a measurement instrument for measuring the effectiveness of a user story refinement session.

Relatedness ratings [17] are a way of measuring the actual knowledge in a team. They do test individual knowledge, where domain concepts are compared to one another by participants in terms of relatedness [22]. However, a technique like this will not work well for this research. User stories are only small increments of a larger product: there are not many concepts that are unique to a single user story (for example, for an information system, there may be dozens of user stories to manipulate the *user profile* concept). This makes it difficult to test the knowledge on a specific user story based on a TA session about that user story. The recommendations by Wildman and colleagues to measure SU as self-reported perceptions of knowledge and process fit best with this research.

In order to measure user perception within information system design methods or techniques, the Method Evaluation Model (MEM) has been proposed by Moody [10]. The MEM evaluates a method or a technique by means of perceived ease of use, perceived usefulness, and intention to use, based on a questionnaire. Since shared understanding is an essential aspect of EM, this aspect should also be incorporated in the measurement of performance for a refinement technique. A link between SU and effectiveness has been observed in various studies [23]–[25]. However, previous research seems inconclusive with regards to how SU influences efficiency [13], [26]. As such, we expand the MEM to incorporate this aspect. The proposed new model, which we call *Refinement Evaluation Tool (RET)*, is shown in Fig. 2. In the RET, 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 the perceived usefulness of the technique.

A. Full Questionnaire

The RET is operationalized via a questionnaire, listed in our online appendix[2]. The questions on perceived ease of use, perceived usefulness and intention to use are adapted from the Method Evaluation Model [10]. SU is built up as a combination of coordination and shared knowledge in the questionnaire:

[2]Our online appendix includes the questionnaire as well as the responses collected through the two case studies reported in this paper: https://doi.org/10.5281/zenodo.5068421

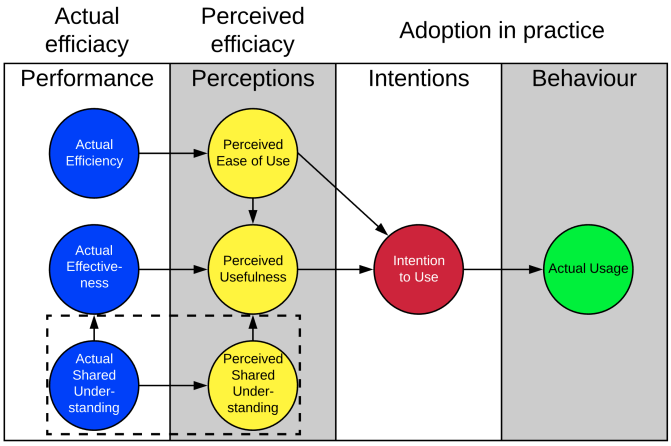


Fig. 2. The Refinement Evaluation Tool (RET), which extends the Method Evaluation Model. The extension to MEM is shown through the dashed box.

- The *coordination* questions evaluate the interaction between team members and are adapted from Lewis [21].
- 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 [27], Smidth et al. [14], and Hu et al. [28].

These 11 questions together measure the perceived SU of a participant. Questions are answered on a five-point scale, from strong disagreement to strong agreement.

As per the original MEM, the questions are put in randomized order. Moody took this measure to reduce the possible ceiling effect in which monotonous responses are given to questions regarding the same concept [10]. This was based on the earlier work of Hu *et al.* [29].

B. Session Questionnaire

The RET can be used for evaluating the performance of a refinement technique. Ideally, in order to measure the learning curve and how the effectiveness of the technique evolves over time, one should employ the questionnaire after every refinement session. However, this is undesirable: some participants may attend all refinement sessions, which may occur once or twice per week, and they may get annoyed with filling in the full questionnaire, thereby influencing their willingness to participate and the quality of their answers. Therefore, we also designed a shortened questionnaire, which can be found in our online appendix. The full questionnaire can be given to participants several times during the case study, at minimum the first and last session of one TA technique. In the other sessions, the shortened questionnaire is used.

The shortened questionnaire does not focus on a participant's perception of the TA technique—which is the case for the full questionnaire—, but rather on the particular session they just had. This way, the perception of both individual sessions and of the technique in general is evaluated.

This session questionnaire adapts the full questionnaire and has four out of five of the same categories. Only intention to use is left out, as this is more about the technique as it is about

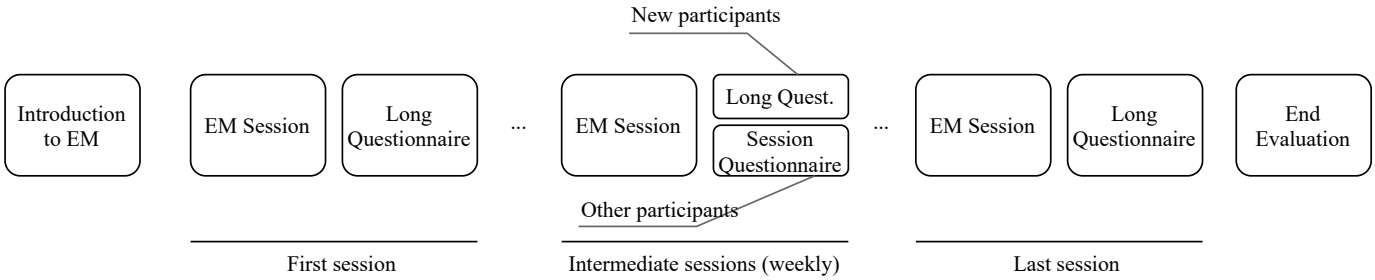


Fig. 3. Illustration of the case study research protocol.

a particular session. No reversed questions are asked either, so to keep the questionnaire short.

V. Case Study Research Method

This research adopts the case study research method [30]. Prior to conducting the case studies, we did perform the planning phase, which led to the protocol shown in Fig. 3 and described in the following:

1) *Introduction to EM*: the researcher provides a short tutorial (15-20 minutes) on the EM technique to the teams who are going to participate in the case study.
2) *First session*: the first EM session is conducted on one or more user stories that are selected by a team member such as the product owner. The researcher acts as moderator when the interaction is not smooth since this is the first time the participants use EM. At the end of the session, the participants take the long questionnaire as a baseline measurement.
3) *Intermediate sessions*: with a weekly frequency, the participants have one EM session for refining one or more user stories. The researcher still acts as moderator if necessary, although the degree of intervention decreases and (s)he becomes mostly an observer. The session questionnaire is used at the end of each session, except for new participants, who take the long questionnaire.
4) *Last session*: the last session is conducted with the researcher as an observer, and the long questionnaire is delivered so to compare the initial results with the final ones.
5) *End evaluation*: the researcher conducts a group interview with as many participants as possible, with the goal of retrieving qualitative feedback on the technique.

Note that the introduction to EM can be combined with the first session, and the end evaluation with the last session.

Observation guidelines. The researcher is expected to attend as many sessions as possible. This allows the researcher also to 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, which is why Wildman *et al.* recommend it in their research [17]. Therefore, when using the RET, we propose to take observations regarding:

- All five questions of SU coordination;
- The perceived involvement of participants;
- The order of writing cards;
- Any other remarkable events that occur in 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 only answer the coordination questions. Besides that, a question on "perceived involvement" is added. This question was added as the work of Van den Bossche and colleagues shows that active contribution is necessary to gain a proper SU [31]. Finally, it is interesting to assess whether there is a link 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 many examples first, or each time write one rule followed by examples. Besides these three observational items, any additional events should 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.

VI. Example Mapping Performance

Using the RET, we have conducted two longitudinal case studies in order to evaluate the performance of Example Mapping (EM). The first case study was performed at Fizor (https://fizor.io/), a low-code software development company. The second case study is conducted at a large pension management firm in the Netherlands, which has requested to remain anonymous.

In both case studies, the first author of this paper taught the techniques to the participants and also facilitated the first two sessions in order to help the teams get started.

8 EM sessions were held in the Fizor case study over 7 weeks, whereas 5 sessions were held over 4 weeks in the second case study. Besides using the two RET questionnaires and conducting the observations, we also conducted an end evaluation after the final session with both case studies in order to evaluate the overall study.

The spreadsheets that include the results that we used to create the charts are also available in our online appendix. Results were collected through Google Forms and analyzed in R 4.0.1 using the `likert` package.

A. Fizor Case Study

The project had an existing set of user stories assigned among several features. For these user stories, no detailed requirements were specified yet. However, each user story had an effort estimation expressed in hours. The user stories were generally small: some user stories would take one or two working days to implement, but the effort of many user stories was estimated between 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 grouped together user stories that regarded similar functionalities before the sessions took place. The first three sessions regarded one user story, and the remaining five sessions were held by combining 2–9 user stories per session.

1) Overall Results: The overall questionnaire results per aspect of all eight sessions are presented in Fig. 4. In this figure, we combined the results of the long questionnaire with the session questionnaire in order to get an overview of all eight sessions. An exception is intention to use, which is not evaluated in the session questionnaire, but only through the data collected in sessions 1, 2 (a new participant joined in that session), and 8.

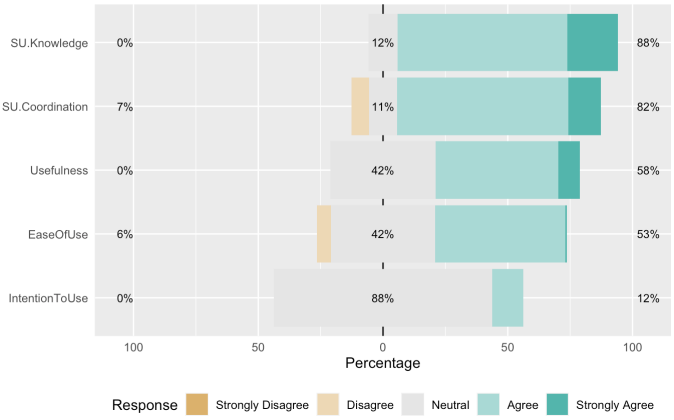


Fig. 4. Per-aspect results for the Fizor case study.

The overall results indicate 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. Interestingly, the results regarding usefulness and ease of use are mildly positive, with a large number of neutral answers. Intention to use is, unsurprisingly, the lowest-scoring aspect, for this is influenced by the other aspects, as shown in Fig. 2.

2) Learning Curve: In order to analyze the sessions in detail, and to assess the learning curve, we created Fig. 5 for the session ratings from participants, together with Fig. 6, which shows the observational ratings of the sessions.

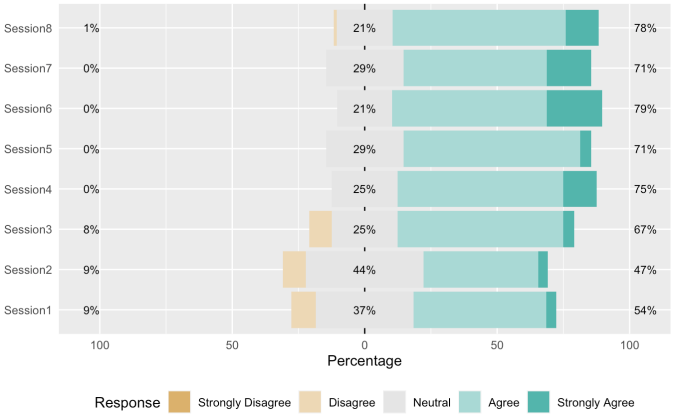


Fig. 5. Per-session results for the Fizor case (self-reported data): all factors but intention to use are included.

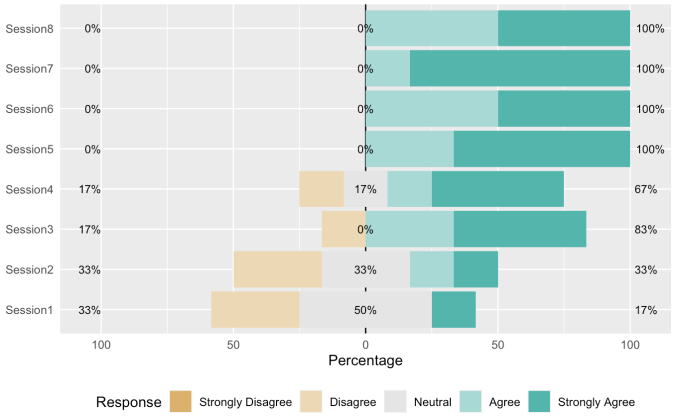
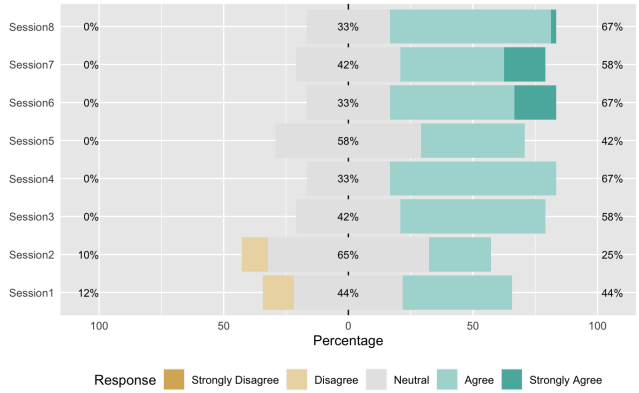


Fig. 6. Per-session results for the Fizor case (observation).

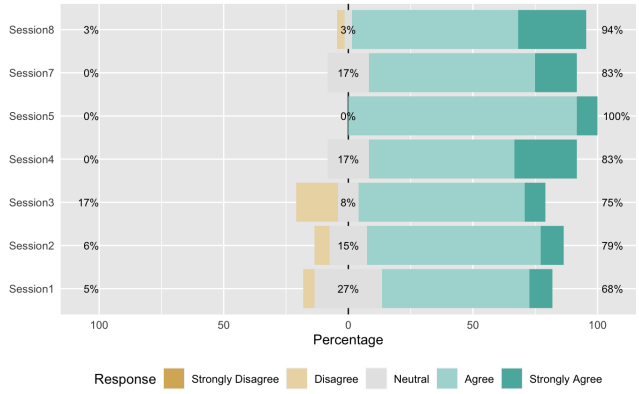
The session ratings are based on all aspects, except for intention to use when the session questionnaire was used, whereas the observation was purely about the process itself and the coordination between team members. Furthermore, Fig. 7 compares side-by-side the evolution per session, when separating the Method Evaluation Model results from the shared understanding questions.

Participation of the team members was not ideal in the first four sessions: of the three participants, one did most of the work. A second participant gave a few comments during one session, while the third participant only contributed when explicitly asked to do so. To improve participation of team members, we had a small intervention before session 5 with the active participant regarding the participation of the other two. We agreed that this participant 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: the other team members 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.

Besides this intervention, we see a learning effect of EM



(a) Method Evaluation Model



(b) Shared Understanding

Fig. 7. Per-session results for the Fizor case, separating method evaluation and shared understanding.

during this case study. Ratings have gradually increased since the first session. These improvements show how the team members have started to master the technique and to reap more benefits from it as well. The trend is particularly evident when evaluating our observations in Fig. 5. When splitting the results as in Fig. 7, we can see how both the MEM and SU aspects increase over time, with an evident appreciation of SU aspects throughout the sessions, and some more varied opinion on the method itself especially in the first two sessions.

The trend was also confirmed by interviews held in an end evaluation: participants had become more familiar and comfortable with the EM technique. During this end evaluation, the participants also mentioned that the technique is really useful for getting an organized overview of what a user story represents and giving them insights that they would otherwise have missed. They believe that the technique is mostly valuable for vague or complex user stories or when a user story has multiple possible implementations that can be visualized using EM. On the other hand, they also believe that user stories that are very small or straightforward are not worth the effort of refining with EM.

Overall, participants became 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, for which EM can still be useful but requires more time than 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. For example, we had sessions where we grouped 5 user stories that still worked out well, but we also had one session with nine grouped user stories and this was deemed too much, causing a lack of focus during the 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.

B. Pension Management Firm Case Study

The second case study was executed with a software development team at a large pension management firm. In total, five sessions were held over the course of this case study. The team that participated in the case study designs and develops APIs that allow their pension management software to interact with systems from third parties.

At first, we had our concerns if what they develop would not be too technically-focused, and therefore maybe not suitable for EM. However, we have found out 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 shared understanding on this after the first or second session.

The participating team has been working together on their products for a long time already. Many user stories were already created, of which several 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.

TABLE I
NUMBER OF ATTENDEES AND RESPONSES PER SESSION.

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

Unfortunately, we did not have a 100% response rate for the questionnaire. As members of this team often had meetings right after the EM session, they were sometimes unable 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. To mitigate this threat, we conducted an evaluation after the fourth session, which we triangulate with our observations. An overview of the number of responses and the number of attendees is presented in Table I. The difference between the number of attendees and responses shows the missing responses per session.

1) Overall Results: The results per aspect are shown in Fig. 8. In this figure, we can observe that SU knowledge is rated the 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.

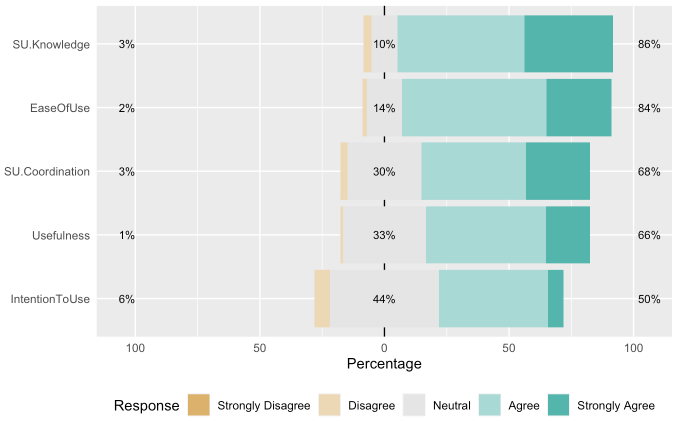


Fig. 8. Per-aspects results for the pension management firm case study.

2) Learning Curve: To analyze the sessions individually, we present the self-reported values per session in Fig. 9, our observations in Fig. 10, and we split MEM and SU in Fig. 11. By comparing the observation results with the total aspect results, we see that coordination was rated by the observer much higher 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 Fizor 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.

Overall, the results from this case study are positive. During

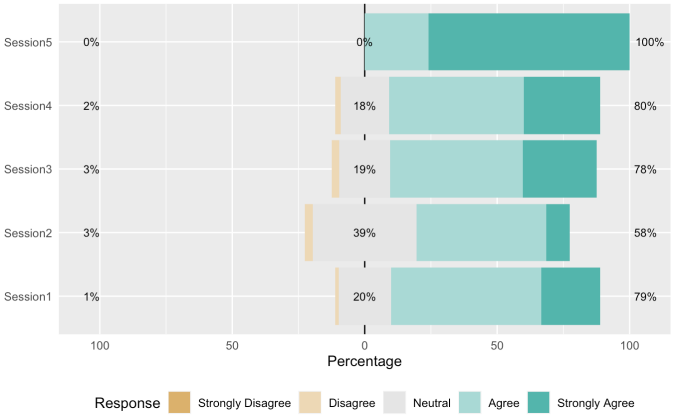


Fig. 9. Per-session results for the pension management firm case (self-reported data).

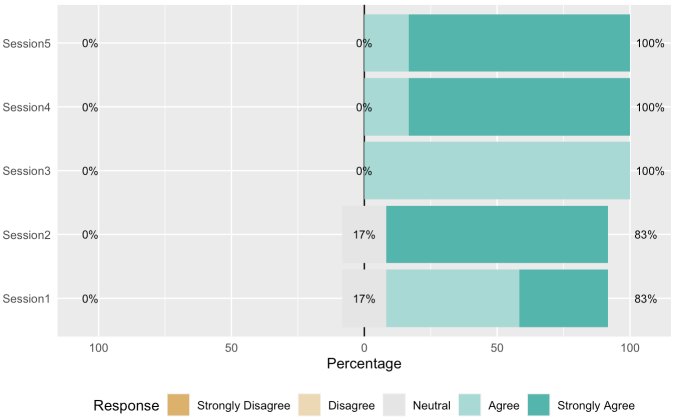
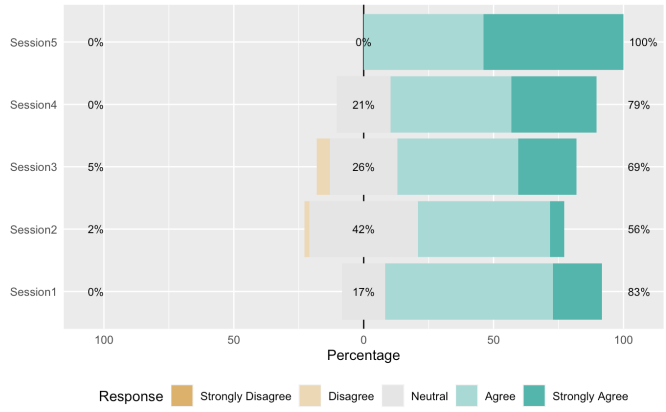


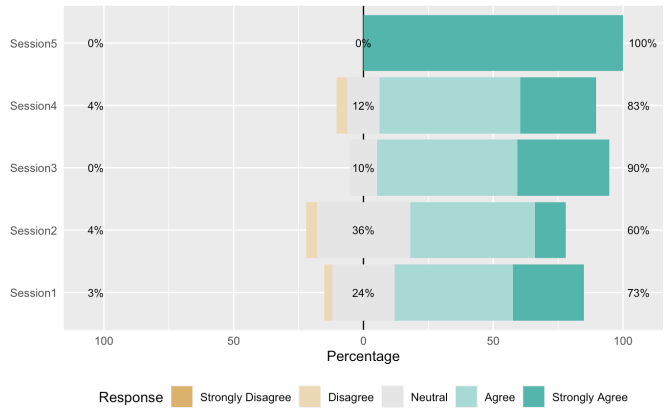
Fig. 10. Per-session results for the pension management firm case (observation).

the second session, EM was not a right fit for the user story, and the session turned mostly into brainstorming. However, even then, the Three Amigos principles of only having one person from each discipline, as opposed to the whole team, still helped to make it a beneficial session to the participants. The TA principles helped to make the meetings effective and efficient. In other sessions, where the entire team was present, the EM outputs were used as a guide for further refinement, and they provided a good overview of the functional requirements and acceptance criteria of the user story.

A learning effect was also observed during this case study. The self-reported data shown in Fig. 9 and Fig. 11 shows a generally positive opinion and an improving trend both on the method and on shared understanding, with the exception of the second session, as explained earlier. These findings were also mentioned during the end evaluation, where a participant mentioned that the sessions went better from the third session onward. The participants also added that the EM output was also valuable to other team members who were not present during the refinement session as it helped for them to quickly see what a user story is about. The Product Owner and Scrum Master of the team also agreed that they want to keep EM



(a) Method Evaluation Model



(b) Shared Understanding

Fig. 11. Per-session results for the pension management firm case, separating method evaluation and shared understanding.

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 technique in their ways of working. This indicates that the team is positive about EM and that it has given them added benefits compared to their previous way of working.

From this case study, we conclude that EM performs well under certain conditions, and that the TA aspects help even when EM is not suitable. EM works well in giving structure to a refinement session, and the resulting output gives a good overview of the requirements of a user story.

VII. Cross-case analysis

Based on the cross-analysis of our case studies, we draw implications for future research and for practitioners who may be interested in adopting EM in their development processes.

In both of our case studies (8 sessions for Fizor, and 5 sessions for pension management firm), we observed that the first couple sessions are necessary for the team to grasp the dynamics of an EM session. This can be seen, for example, in the self-reported values for shared understanding in Fig. 7(b) and in Fig. 11(b). After that, a learning effect is visible and the team starts working in an effective manner. The participants from the pension management firm case also confirmed, in the final interview, that they observed the interaction went smoothly from the third session on. This leads to our first implication on the learning effect.

Implication 1: Learning effect

It takes around two sessions, assisted by a moderator, for a team to learn to use EM sessions. Our cases reveal that, starting from the third session, a team should be able to conduct an EM session largely on their own.

Regarding the effectiveness of EM in general: for Fizor, EM performed well, especially on both the SU aspects of knowledge and coordination, see Fig. 4. Intention to use was rated neutrally, perhaps influenced by the small size of the user stories. The results for the second case are positive on most aspects, including SU, see Fig. 8. According to the participants, EM helps to give structure to a refinement session and to have a good overview afterwards of the requirements that a user story entails. Moreover, the outputs of the EM sessions were also valuable to other team members that were not present during the session. The team, in fact, continued with the use of EM after this case study and will even encourage other teams to adopt it. Thus, our second implication.

Implication 2: EM is beneficial for team SU

EM is beneficial for shared understanding within a team, both in terms of knowledge and coordination. The sticky notes produced in an EM session may be used to increase SU for those team members who did not attend the session.

In general, intention to use is rated lower than the other dimensions. This was expected: as visible in Fig. 2, and as explained in the MEM model, intention to use is affected both by perceived ease of use and by perceived usefulness.

Furthermore, the research reveal an important contextual factor that we could not foresee ahead of time, and that the grey literature behind EM [6] did not mention: the size of the selected user stories for use in an EM session. In the Fizor case, we identified the importance of choosing stories that are not too small. Applying EM to those would be a waste of time: conducting a 30 minutes refinement for a user story with effort estimated to 1 or 2 hours is clearly too expensive. The Fizor participants said that 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. We therefore draw a third implication.

Implication 3: low-effort USs

Contextual factors exist that determine the suitability of EM as a refinement technique. One key factor we identified is that too small user stories, i.e., with an estimated effort of 1-2 hours, are unsuitable. However, it is still possible to use EM by grouping related user stories in the same EM session.

One additional consideration regards the specific technique (EM) or the more general paradigm (Three Amigos sessions). While in this research we measured the effectiveness of EM, in the pension management firm case study, the participants highlighted how TA sessions may be a useful technique for refining user stories in a smaller, yet diverse team, which may lead to shared understanding also thanks to the small number of people who are involved in the discussion. It will be interesting, thus, to investigate to what extent the effectiveness depends on EM, or whether alternative TA approaches, such as feature mapping [11], would also deliver similar results. However, as stated earlier in this section, the participants in the pension management firm case study argued that the specific structure of EM improved their sessions.

We cannot draw any definite conclusion regarding the order in which the EM canvas (the sticky notes) are created. All the participants started with rules, and then added examples to illustrate those rules. While this behavior is quite consistent, it is plausible that this happened because of the training we gave or because of the limited sample size of our case studies.

VIII. Discussion

We present the main conclusions, discuss threats to validity, and sketch directions for future research.

A. Conclusions

This research has investigated the effectiveness of requirements refinement techniques, in particular example mapping, via longitudinal case studies conducted in two organizations. To do so, we have proposed the Refinement Evaluation Tool, a performance measurement instrument for refinement techniques with an emphasis on shared understanding.

The research has practical relevance, as we provide insights into the performance of EM and state several conditions when technique may be more or less suitable. In particular, EM is an industry-pushed technique, and we attempted to move first steps toward a rigorous assessment of its effectiveness, beyond the claims of the agile coaches who promote its use.

We have found TA sessions result in a good SU among the team members, which previous research has shown to improve overall team performance [7]. As such, this research can act as a guidance for teams that are considering different refinement techniques. They can either rely on the results for EM that are presented in this paper, or re-use the measurement tool RET for their own setting, prior to adopting the technique widely in their organizational context.

B. Threats to validity

A first limitation is the number of cases that were researched. As this was mainly qualitative research, results are context-specific, which makes generalization difficult. We tried to mitigate this threat by having two data sources for each case, namely the questionnaires and observations. Given the consistency in our findings across the two cases, we believe the findings do hold for teams that are in a similar context as the ones we have researched, but additional case studies in different contexts are necessary for further generalization.

A second limitation is that the longitudinal case studies did not allow us to investigate how TA sessions affect long-term aspects. Our initial plan was to have case studies that would last around three months, which would also allow us to investigate how EM outputs might affect the implementation of a software increment. However, due to the COVID-19 pandemic, our initial case studies had been canceled and new studies had to be found and set up. These new longitudinal case studies did not allow us to investigate the long-term effects, which is why that aspect was refrained from the research.

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 limitation of this research is that part of it 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 a TA session technique. This is a risk to the validity of the research. The severity is likely to be low, as teams already had one or two months to adjust to the new situation before the case studies actually started. The 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.

C. Outlook

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 generalizing 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 long-term effects can be researched. By looking at the entire software development cycle, insights could be found about TA session techniques that we were unable to find.

Additionally, research could be conducted to investigate the RET in more detail. The questionnaire is based on previous research and also partially validated in combination with the observations, but further 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.

Acknowledgment

We would like to thank Info Support and the employees from Fizor and from the pension management firm for enabling our in-vivo research during the COVID-19 pandemic. We thank the members of the RE-Lab at Utrecht University for the discussions and feedback around the research.

References

[1] K. Beck, M. Beedle, A. Van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries *et al.*, "The agile manifesto," 2001.

[2] K. Beck, *Test-driven development: by example*. Addison-Wesley Professional, 2003.

[3] J. F. Smart, *BDD in Action: Behavior-driven development for the whole software lifecycle*. Manning, 2015.

[4] M. Cohn, *User stories applied: For agile software development*. Addison-Wesley Professional, 2004.

[5] J. Bosch, "Speed, data, and ecosystems: The future of software engineering," *IEEE Software*, vol. 33, no. 1, pp. 82–88, 2015.

[6] M. Wynne, "Introducing example mapping," Dec 2015, last accessed April 6, 2021. [Online]. Available: https://cucumber.io/blog/bdd/example-mapping-introduction/

[7] M. Glinz and S. A. Fricker, "On shared understanding in software engineering: an essay," *Computer Science-Research and Development*, vol. 30, no. 3-4, pp. 363–376, 2015.

[8] C. Werner, Z. S. Li, N. Ernst, and D. Damian, "The lack of shared understanding of non-functional requirements in continuous software engineering: Accidental or essential?" in *Proc. of the IEEE International Requirements Engineering Conference (RE)*, 2020, pp. 90–101.

[9] E.-M. Schön, J. Thomaschewski, and M. J. Escalona, "Agile requirements engineering: A systematic literature review," *Computer Standards & Interfaces*, vol. 49, pp. 79–91, 2017.

[10] D. L. Moody, "The Method Evaluation Model: A theoretical model for validating information systems design methods," *Proc. of the European Conference on Information Systems (ECIS)*, p. 79, 2003.

[11] J. F. Smart, "Feature mapping – a simpler path from stories to executable acceptance criteria," Jan 2017, last accessed April 6, 2021. [Online]. Available: https://bit.ly/31Q6fwH

[12] A. Alliance, "Three amigos," Sep 2019, last accessed April 6, 2021. [Online]. Available: https://www.agilealliance.org/glossary/three-amigos/

[13] A. Açıkgöz, A. Günsel, N. Bayyurt, and C. Kuzey, "Team climate, team cognition, team intuition, and software quality: The moderating role of project complexity," *Group Decision and Negotiation*, vol. 23, no. 5, pp. 1145–1176, 2014.

[14] C. Schmidt, T. Kude, A. Heinzl, and S. Mithas, "How agile practices influence the performance of software development teams: The role of shared mental models and backup," *International Conference on Information Systems (ICIS)*, vol. 35, 2014.

[15] X. Yu and S. Petter, "Understanding agile software development practices using shared mental models theory," *Information and Software Technology*, vol. 56, no. 8, pp. 911–921, 2014.

[16] N. J. Cooke, J. C. Gorman, C. W. Myers, and J. L. Duran, "Interactive team cognition," *Cognitive Science*, vol. 37, no. 2, pp. 255–285, 2013.

[17] J. L. Wildman, E. Salas, and C. P. Scott, "Measuring cognition in teams: A cross-domain review," *Human Factors*, vol. 56, no. 5, pp. 911–941, 2014.

[18] E. Salas, S. M. Fiore, and M. P. Letsky, "Theoretical underpinning of interactive team cognition," in *Theories of team cognition: Cross-disciplinary perspectives*, vol. 49. Routledge, 2013, pp. 187–207.

[19] Y. Ren and L. Argote, "Transactive memory systems 1985–2010: An integrative framework of key dimensions, antecedents, and consequences," *Academy of Management Annals*, vol. 5, no. 1, pp. 189–229, 2011.

[20] F. W. Kellermanns, J. Walter, C. Lechner, and S. W. Floyd, "The lack of consensus about strategic consensus: Advancing theory and research," *Journal of Management*, vol. 31, no. 5, pp. 719–737, 2005.

[21] K. Lewis, "Measuring transactive memory systems in the field: scale development and validation." *Journal of Applied Psychology*, vol. 88, no. 4, p. 587, 2003.

[22] J. C. Gorman and N. J. Cooke, "Changes in team cognition after a retention interval: the benefits of mixing it up." *Journal of Experimental Psychology: Applied*, vol. 17, no. 4, p. 303, 2011.

[23] J. Cannon-Bowers, E. Salas, and S. Converse, "Cognitive psychology and team training: Shared mental models incomplex systems," in *Annual Conference of the Society for Industrial and Organizational Psychology (SIOP)*, 1990.

[24] R. Klimoski and S. Mohammed, "Team mental model: Construct or metaphor?" *Journal of Management*, vol. 20, no. 2, pp. 403–437, 1994.

[25] S. Mohammed, R. Klimoski, and J. R. Rentsch, "The measurement of team mental models: We have no shared schema," *Organizational Research Methods*, vol. 3, no. 2, pp. 123–165, 2000.

[26] K. A. Smith-Jentsch, J. E. Mathieu, and K. Kraiger, "Investigating linear and interactive effects of shared mental models on safety and efficiency in a field setting." *Journal of Applied Psychology*, vol. 90, no. 3, p. 523, 2005.

[27] B.-C. Lim and K. 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*, vol. 27, no. 4, pp. 403–418, 2006.

[28] J. M. Gevers, C. G. Rutte, and W. Van Eerde, "Meeting deadlines in work groups: Implicit and explicit mechanisms," *Applied psychology*, vol. 55, no. 1, pp. 52–72, 2006.

[29] P. J. Hu, P. Y. Chau, O. R. L. Sheng, and K. Y. Tam, "Examining the technology acceptance model using physician acceptance of telemedicine technology," *Journal of Management Information Systems*, vol. 16, no. 2, pp. 91–112, 1999.

[30] R. K. Yin, *Case study research and applications: Design and methods*. Sage publications, 2017.

[31] P. Van den Bossche, W. Gijselaers, M. Segers, G. Woltjer, and P. Kirschner, "Team learning: building shared mental models," *Instructional Science*, vol. 39, no. 3, pp. 283–301, 2011.