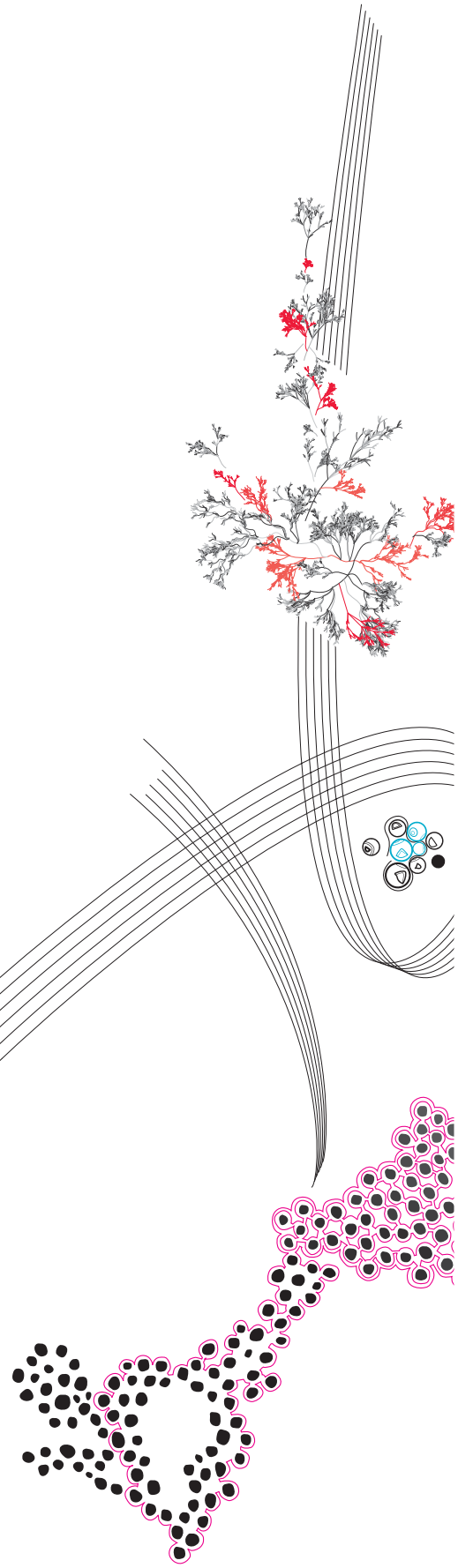


MSc Interaction Technology
Final Project

Exploring the Meeting Experiences in the Metaverse: A User Study on Immersive Interactions

Paolo Barzon

Supervisor: Dennis Reidsma
Supervisor: Jan Kolkmeier
Supervisor: Sylvie Dijkstra-Soudarissanane

October, 2023

Interaction Technology Programme
Faculty of Electrical Engineering,
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Contents

Abstract

Meetings in non-immersive platforms, such as MS Teams, present issues of usability, effectiveness, and social communication nature. As a consequence, employees complain that their online meeting experience is substandard. However, TNO proposes its own immersive communication systems as a substitute. Its immersive environment can compensate for the insufficient experience that non-immersive platforms provide. We conducted an experiment to test whether these claims, also supported by the literature at hand, apply as well to the system developed by TNO. The results show that the Pointcloud technology has potential, but the technological limitations prevent the system from being a valid alternative to MS Teams in the near future. We then highlighted the main issues in the technology that need to be addressed, to provide a meeting experience similar to or even better than the one on non-immersive platforms.

Keywords: User study, QoE, immersive communication systems.

Chapter 1

Introduction

The COVID-19 pandemic has forced many organizations to shift most of their day-to-day activities online, including meetings and presentations. The non-immersive, online meeting platforms chosen to host these activities were, for example, MS Teams and Zoom. Several aspects, such as communication and effectiveness, were negatively impacted by this change in the communication media, from face-to-face to online interactions. Meanwhile, the scientific studies presented in this thesis show that Immersive Communication Systems (ICSs) address said meeting issues by providing better social communication and meeting effectiveness thanks to the immersion given by their 3D environments. In particular, the immersiveness offered by ICSs enables users to feel as if they are communicating face-to-face and in a real-life context, which is reported to be better than using a personal device as a mediator. As a result, ICSs have gained increased attention from companies and researchers as replacements or additions to traditional online meeting platforms.

In this thesis, we describe the Thesis assignment conducted at the TNO company on understanding the feasibility of deploying ICSs for business meetings by evaluating the quality of experience provided. The name TNO stands for "Nederlandse Organisatie voor Toegepast Natuurwetenschappelijk Onderzoek"; in English: Netherlands Organisation for Applied Scientific Research. One of the research projects that TNO is involved in is developing immersive technologies that positively influence the overall online meeting experience and outcomes. In particular, the Social XR team of TNO is developing an ICS (shown in figure 1.1) in which their employees from different office locations can meet and collaborate in Virtual Reality (VR), taking advantage of the benefits of the immersive platform, and overcoming the limitations of traditional online meetings.

This thesis builds on the literature review completed previously, presented in chapter 2, which discussed how ICSs can provide a better meeting experience than traditional meeting platforms thanks to the enhanced quality of social communication provided, and the increased meeting effectiveness given by the shared collaborative tools. The problem statement that this thesis addresses is the unsatisfactory meeting experience given by traditional online meeting platforms. The high-level goal of the thesis is to assess whether the TNO ICS follows the premises of the literature review, thus providing better meeting experiences than non-immersive platforms, and to understand how feasible it would be to deploy such technology across the TNO locations in the Netherlands. The hypothesis of the Social XR team is that their ICS, thanks to its benefits mentioned before, will perform more or less favorably to traditional meeting platforms depending on the meeting type and the number of meeting participants, on measures of perceived effectiveness, social communication, and usability. We expect the TNO ICS not to be perfect, as the scientific literature collected highlighted a number of characteristics that can have a negative im-

Figure 1.1: Snapshots from the TNO ICS

pact on the meeting experience, including inducing a feeling of nausea and suffering from technological immaturity.

The next chapter of the thesis is the background research, which outlines the problem statement about the current non-immersive meetings' shortcomings, and the literature review, which examines studies to define the ICS technology, online meetings, the ICS added values, and the suggested evaluation methods; then, we present the previous study on the TNO ICS. The background research then concludes with a discussion and conclusion of the studies presented; we explain there that ICSs differ in technology, context, meeting type and their own characteristics, immersion and presence, and evaluation methods to assess the QoE. Chapter 3 defines the areas of interest for the Master's Thesis Assignment, namely understanding how the meeting experience changes under different meeting settings, and what needs to be improved to make ICSs a valid replacement for traditional online platforms. In chapter 4 we present the study design, the study population, the system design, and the procedure of research of our experiment, supported by the background research. Chapter 5 defines the comprehensive data analysis conducted to address the research questions and objectives of this study. The results of the data analysis on the interviews and questionnaire are presented in chapter 6; the chapter is divided into five sections, one for each research question. Then, in chapter 7 we provide explanations and interpretations of the results of the experiment. We describe our acknowledgment of the result limitations and give recommendations for future works in chapter 8. The last chapter summarizes the content and results of our research.

In summary, this thesis highlights that non-immersive platforms receive higher ratings for the majority of the meeting QoE factors than immersive platforms; the TNO ICS is not yet suitable for hosting meetings due to the low QoE it provides; nevertheless, we recognize the potential of this technology, and we believe it will perform at its highest capacity in the near future.

Chapter 2

Background

The purpose of the background research is to establish the context and relevance of our experiment. We present first the shortcomings of traditional online meetings, then the literature reviewed to better understand the context of ICSs and work meetings, and lastly the TNO ICS.

2.1 Problem Statement: Traditional Online Meetings Present Several Shortcomings

Traditional online meeting platforms are widely used nowadays and denote the most common means of communication between remote users. Examples of such platforms include MS Teams[1], Zoom[2], and Skype[3]. However, several limitations have been identified that negatively affect the user experience. The following paragraphs present six studies, describing the issues of traditional meetings in the factors of usability, effectiveness, and quality of social communication provided.

Karl et al. [22] categorized into six categories the causes of frustration that emerge during online meetings. The most common problems related to non-immersive meetings are an incorrect use of the video camera or microphone and annoying background noises. Participants of the study reported feeling higher levels of frustration and stress during and after such meetings, which compromised their quality of experience. Yankelovich et al. [43] described three types of issues of online meetings: audio issues, behavior issues, and technical issues. The audio problems, such as background noise and subpar microphone quality, were the most frequent and, together with behavior problems, had the largest negative impact on meeting effectiveness. Lastly, the main technical problems reported were "not everyone could view visual materials" and "necessary documents not available during the meeting".

The efficiency of meetings is reportedly negatively affected when meeting online. First, in the study of Brucks and Levav [13], participants in non-immersive online meetings generated significantly fewer total ideas and creative ideas than in-person pairs. This was due to fear of evaluation, dominance, social facilitation, social loafing, social sensitivity, perceptions of performance, and production blocking. Secondly, Kuzminykh and Rintel [23] analyzed the attention of participants during work meetings, revealing how remote participation is generally associated with lower motivation to engage both behaviorally and cognitively.

[1] https://www.microsoft.com/en-us/microsoft-teams/log-in
[2] https://zoom.us/
[3] https://www.skype.com/en/

Social communication aspects of online meetings are worse than in-person meetings. The research of Shoshan and Wehrt [31] identified the causes of exhaustion in the context of traditional online meetings. Participants in the study described difficulties relating to social aspects, namely struggling to read the social cues of others, while perceiving overwhelming pressure to provide such cues themselves. Fauville et al. [15] presented the term "zoom fatigue" as the sum of visual fatigue, vocal fatigue, and emotional fatigue. Long and frequent non-immersive online meetings tend to increase feelings of fatigue, which also leads to a negative attitude toward online meetings. The causes can be found in a failed trade-off between eye gaze and interpersonal distance, increased cognitive load, self-evaluation, and a lack of motion.

This chapter contains the most relevant empirically proven studies on drawbacks brought by 2D meetings. The issues discussed include usability, effectiveness, and social communication. The first category is intrinsic to technological systems. The last two categories of problems derive from the substandard effectiveness and quality of non-immersive online meeting channels, which hinder and limit users' communication and social interactions, due to the lack of social inclusion and presence. We will discuss again the issues related to traditional meeting platforms in chapter 7 and 9, to understand whether the TNO ICS prevents them, or whether we should report them in the list of improvements needed in the immersive platform.

2.2 Literature Review

We conducted the literature review to ground our decisions and hypothesis for the later phases of the assignment on scientific sources. The empirical data collected belongs to the context of our research, namely ICSs and work meetings. We researched four main concepts: ICSs, meetings, the added values of ICSs on meetings, and the evaluation of ICSs and meetings. In the following sections, we define four Literature Review Questions (LRQ) to guide our research on the studies conducted on ICSs and work meetings.

2.2.1 LRQ1: What Are Immersive Communication Systems?

Given the numerous differences between ICSs developed in recent years, the first LRQ aims at having a thorough understanding of ICSs by guiding our research on the definitions of ICSs and the aspects of immersion and presence they provide; the technologies used in different ICSs; the possible context of use of ICSs. Lastly, we present the State-of-the-Art.

We can define ICSs as platforms that enable users to connect and interact with each other in a virtual world while perceiving a sense of presence. In this thesis, we will commonly call the "Metaverse" any Virtual Environment (VE) accessible through ICSs. As defined in the seminal work of Slater [36], this sense of presence is a factor intrinsic to immersive systems, and it refers to a user's subjective psychological response to the system. When multiple users are utilizing the system at once, the feeling of presence they share is referred to as "social presence." The more immersive the system is, the higher the participants' sense of social presence. Lastly, immersion defines the degree to which ICSs provide realistic stimuli and experiences to simulate the real world.

The Virtuality Continuum (VC) spectrum defined by Milgram [29] categorizes the immersive technologies that provide a sense of presence or social presence. The ICSs are located in a spectrum that ranges from the real environment to a completely virtual environment. The author also defines six classes of display environments evenly distributed

across the spectrum; the least immersive class, closest to the real world, consists of monitor-based video displays; the most immersive class, closest to the virtual world, consists of complete graphic environments in which real physical objects play a role in the computer-generated scene.

ICSs can be defined by their context of use, and the most relevant examples are meetings, education, and training. In the empirical study of Fernandez Langa et al. [25], participants simulated the interactions that usually take place during meetings. Users rated their experience and interactions positively, reporting high satisfaction with regard to usability, presence, interaction quality, workload, scenario, and task effectiveness. Alhalabi's work [5] focused on the learning experience of students in engineering education. The questionnaire results showed how students who employed Head Mounted Displays (HMDs) performed better than their counterparts who used other technologies, and reported high levels of satisfaction with the immersive experience. Lastly, in the field of Undergraduate Nursing Education, Aebersold [4] concluded that virtual simulation provides an effective alternative to clinical training, as it allows students to practice caring for patients while remaining within a safety net.

This last paragraph will describe Meta Horizon Workrooms, shown in figure 2.1, which is the current leader in the field of immersive communication systems for work meetings. Its user-friendliness and accessibility make it the best immersive platform for communication. Meta's platform serves as the best example amongst many, such as Glue[4], Connec2[5], and Engage[6], each contributing unique immersive collaboration features. On Milgram's VC

Figure 2.1: Snapshot from the Meta Horizon Workrooms

continuum [29], the position of the Meta Horizon Workrooms is halfway through the scale, and the technology used falls into the fourth of the six categories, defined as "HMDs equipped with a see-through capability given by video representation". Its context of use, in a professional setting, is related to work meetings. Its most relevant features, which are described on the Meta website [1], include a Mixed Reality (MR) space that allows users to interact with their computer without disconnecting from the virtual environment.

[4]https://www.glue.work/
[5]https://connec2.nl/
[6]engagevr.io

The shared whiteboard facilitates synchronous collaboration as well as private note-taking. Other features include computer screen sharing, breakout rooms, and editable avatars that track and convey facial movements. Lastly, natural spatial audio facilitates communication, especially during meetings with more than two participants in the Meta Metaverse, to easily understand which user is talking. On the negative side, the avatars employed to represent the meeting participants do not properly convey non-verbal communication factors such as gaze, posture, and facial expressions, thus limiting the social cues shared during the meeting.

In this section we have discussed the findings of LRQ1, starting with defining ICSs as platforms to connect users in a virtual world, where the feelings of presence in users, i.e. to "be there" in the VE as if it was real, is given by the immersive nature of the system, i.e. providing realistic stimuli and experiences. Secondly, ICSs differ first in the technology used and their position on the RV Continuum. Then, ICSs differ in the context of use. Other than work meetings, immersive systems are employed for education and training. Lastly, the features and interactions of Meta Horizon Workrooms provide a great experience for users and set the standards that other ICSs should aim for, including the TNO ICS. To answer LRQ1, we can say that ICSs are platforms primarily used for work meetings, training, and educational purposes that convey a sense of presence through the use of technologies that digitally replicate the real environment at varying degrees.

Next, we will look in more detail at one of the contexts of use that these systems positively impact, i.e. online meetings.

2.2.2 LRQ2: What Are Online Meetings?

We define LRQ2 to explore the second topic of interest of our literature review, namely online meetings, by researching their different types, modes, and objectives.

Standaert et al. [39, 38, 37] discuss meeting modes, objectives, capabilities, and their relationships in three papers. The authors identified four meeting modes: Audio-Conferencing (AC), Video-Conferencing (VC), Tele-Presence (TP), and Face-to-Face (F2F). In the context of our thesis, VC is related to traditional online meeting platforms, while TP indicates immersive meeting platforms. The authors collected and ordered the fifteen most relevant and common meeting objectives of business meetings, with the first three being (a) to clarify a concept, (b) to exchange/share opinions or views, and (c) to build trust and relationships with one or more individuals. Additionally, the authors identified six capabilities as significant contributors to the meeting objectives: (a) hearing attendees' voices, (b) using shared computer screens and/or workspaces, (c) experiencing co-location, (d) seeing attendees' body language and gestures, (e) discerning attendees' facial expressions, and (f) observing what attendees are looking at. The authors related the capabilities to the meeting modes in a progressive sequence: AC provides the first two capabilities; VC provides the first four; TP and F2F provide them all. The authors also suggested the meeting mode that best fits the meeting objectives: exchanging information is associated with AC; making decisions and communicating sentiments are related to VC or TP; building relationships is best in TP or F2F. We graphically summarized the three studies of Standaert et al. in table 2.1.

Meeting Objective	Sub categories of meeting objectives	Meeting capabilities						Meeting mode	Pairwise Comparison	
		Hear attendees' voice	Use shared computer screens and/or work spaces	Experience co-location	See attendees' body language and gestures	Discern attendees' facial expressions	Observe what attendees are looking at		Eff TP > eff AC	Eff TP > eff VC
Exchanging Information	*Clarify a concept, issue or idea*							**Audio-conferencing**	3,63*	3,52*
	Routine exchange of information								-0,42	-0,14
	Non-routine exchange of information								3,02*	2,25
	Give or receive feedback								3,92*	3,47*
Making Decisions	*Exchange/share opinions or views on a topic or issue*							**Video-conferencing or telepresence**	2,58*	0,5
	Find a solution to a problem that has arisen								0,43	0,66
	Generate ideas on products, projects or initiatives								1,68	0,3
	Generate buy-in or consensus on an idea								2,76*	1,75
	Make a decision								1,42	1,04
Communicating Sentiments	*Show personal concern about or interest in a particular issue or situation*								3,34*	1,91
	Exchange confidential, private or sensitive information								2,72*	0,96
	Communicate positive or negative feelings or emotions on a topic or issue								5,57*	3,37*
Building relationships	*Build trust and relationships with one or more individuals*							**Telepresence or f2f**	7,94*	4,54*
	Maintain relationships with one or more other people and stay in touch								3,59*	2,11
	Assemble a team and/or motivate teamwork on a project								2,68*	0,61

Table 2.1: Summary of meeting objectives, capabilities, modes, and their rating

Expanding the work of Standaert, the Future of Work (FoW) team at TNO conducted a literature review [currently under submission] to understand meetings under different lenses. They first divided meetings into (a) regularly repeated meetings, (b) scheduled as needed, and (c) "learn and influence meetings". For each category, they provided examples and explained the intentions, participants, format, criticalities, capabilities, and meeting mode suggested. A relevant result of the study covers the meeting types that are not suggested to be held in ICSs: progress checks, governance cadence, planning, workshops, information gathering, community of practice, broadcast, and training meetings. Given the capabilities needed and the criticalities found, these meetings are more suitable to be held in VC or AC. The FoW team concluded that engaging participants and the flow of conversation play a significant role in reaching the goals of most meetings and positively impact their outcomes. Out of the ten meeting types that present engagement or flow as a criticality, eight of them are suggested to be held in immersive systems.

We introduced in these last two sections the term "meeting mode", used by Standaert et al. in their publications to indicate the technology used to host the meeting, i.e. immersive or non-immersive platforms. However, in the rest of the thesis, we will refer to meeting platforms as "communication media", as it is the appropriate term to use in the field of Human-Machine Interaction.

Yung et al. [44] designed the SPEL cube to graphically classify meetings based on: (a) the categories of location (ranging from physical to virtual), (b) virtuality of environment (ranging from real to virtual), and (c) social presence (ranging from low to high). Each of these three categories is positioned on one of the three dimensions of the cube (height, depth, and length); as a result, the eight vertexes of the cube present unique combinations of the values from the three categories mentioned above; figure 2.2 represents graphically the cube. According to the authors, ICSs would be positioned on vertex number 8, as they are characterized by a virtual location, high social presence, and high virtuality of the environment. On the other hand, traditional online meetings are located on vertex 5, as they share with ICS a virtual location, but the social presence is much lower due to the lack of immersion, and the environment not at all virtual.

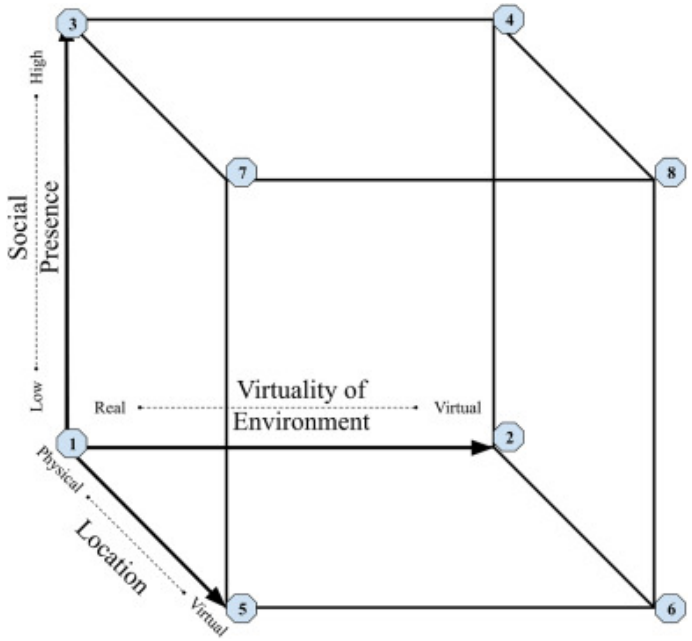


Figure 2.2: The SPEL cube by Yung et al.

This section concludes the research on the available literature to answer the question "what are online meetings?". To summarize the findings, we first identified the four main meeting objectives and fifteen subcategories that define the needs and goals of each meeting, which are reached when employing the related capabilities. In addition, the FoW team proposed a second framework, which explains how each meeting type can be defined according to its intentions, participants, format, criticalities, capabilities, and meeting mode. Lastly, we divided online meetings according to the SPEL cube: the only commonality of immersive and non-immersive meetings is the virtual location, while they are opposite in the virtuality of environment and social presence levels. To answer LRQ2, we can say that online meetings are defined by their objective, which is reached more easily if the meeting mode selected offers the appropriate meeting capabilities.

Until now, we have discussed the limitations of traditional online meetings and introduced the two main components of our literature review: ICSs and online meetings. In the following section, we will research how combining ICSs with online meetings can address the limitations of traditional online meetings, and investigate the added values that ICSs can bring to online meetings.

2.2.3 LRQ3: What Are The Added Values of ICSs That Enhance Online Meetings?

The literature review discussed in the previous sections proved that traditional meeting platforms do not always offer the best meeting experience to their users. Instead, TNO argues that ICSs are a valid alternative to online meetings thanks to the added values brought by enhanced social communication. This claim was already partially supported by the studies presented in the previous section, as ICSs provide more meeting capabilities than traditional meeting platforms, which are necessary to reach the goals of a number of meeting types. To further verify whether this assumption is correct, we formulated the third LRQ to better understand the added values of ICSs over non-immersive meeting platforms and to determine whether ICSs can compensate for their shortcomings. The literature at hand demonstrates that the main added values of ICSs for online meetings are: (a) better communication, (b) enhanced meeting effectiveness, (c) increased product creative quality, (d) shared real-time interactions, and (e) positive user perception of the technology and feeling of immersion, which are presented below.

Maloney et al. [28] studied the behavior of participants in social events held on VR platforms. Users considered non-verbal communication in social VR similar to offline F2F interactions. Nonverbal cues were reportedly relevant to engage in a realistic and immersive online interaction, as they help with expressing and sharing feelings. Providing more natural communication is the main added value of ICSs, which addresses the drawbacks described in section 2.1, where we discussed how users in non-immersive platforms complain about the lack of such non-verbal cues and communication.

Standaert et al. [37] observed in their research that the effectiveness of TP meetings is significantly higher than that of AC and VC meetings for four out of fifteen subcategories of meeting objectives: (a) build trust and relationships; (b) communicate feelings or emotions; (c) give or receive feedback; and (d) clarify a concept, issue or idea. In addition, TP meetings were found to be statistically more effective than AC meetings (and not for VC meetings) for seven more subcategories of meeting objectives. In only three subcategories, specifically, (a) finding solutions, (b) making decisions, and (c) generating ideas, TP was not significantly more effective than both AC and VC. Lastly, in only one category, namely, routine exchange of information, TP meetings perform worse than AC and VC, but not significantly. One possible explanation from the authors is that the life-size presence,

sense of shared space, and eye contact may help participants transmit cues, which convey trust, warmth, and attentiveness. As a result, TP is the suggested meeting mode for the four meetings mentioned above (build trust and relationships; communicate feelings or emotions; give or receive feedback; clarify a concept, issue or idea) because they provide a number of added values that aid meeting participants in reaching their goals. As for the comparison with F2F meetings, TP is not classified as statistically less effective for any of the subcategories of meeting types, but neither statistically more effective. These results show that despite the additional capabilities of F2F meetings compared to those that TP provides, the latter is found to be comparable in effectiveness for achieving objectives in meetings that might not be possible to be held in person. When introducing additional aspects of different natures than the categories mentioned above, the current analysis suggests that in situations where F2F meetings would require significant travel, time, and cost, TP provides an effective, possibly less costly, and more environmentally friendly alternative. Furthermore, the technological features of TP meetings could further enhance the experience, for example by allowing participants to record the meeting or easily share media, which would not always be as effective as in F2F meetings. Finally, within an ICS, conducting meetings is linked to higher levels of engagement and effectiveness in four types of meetings, thereby addressing the shortcoming regarding the unsatisfactory effectiveness provided by a non-immersive platform, as presented in section 2.1.

The study from Yang et al. [42] showed a significant difference in participants' individual product creative quality between the VR experimental condition and the paper-and-pencil control condition. More specifically, this benefit derived from the immersive VR technology, which allowed the participants to focus more, enter a better state of flow, and allowed for greater creative performance. The study validates that productivity is improved within ICSs, countering the limitation observed in traditional meeting platforms, as described in section 2.1.

The proof-of-concept study from Sadeghi et al. [34] aimed at evaluating the feasibility and efficacy of organizing remote VR meetings for medical teams. Participants appreciated the possibility of interacting together and in real-time on the same artifact shared between them. This added value reportedly improved the user experience and led to a positive attitude toward the use of VR as an alternative method for remote conferencing.

Lastly, the paper of Gunkel et al. [19] presents a modular web-based VR application, which was tested in a demo session at a conference space. Most participants appreciated the overall quality of the experience and felt positively involved in the VR environment. Furthermore, the evaluators noted a high degree of activity and interaction between users. The study shows that people have a positive opinion of this new technology and feel comfortable in such an immersive system. Although this study does not show that ICSs have better usability than traditional meeting platforms, it is surely a head start towards providing a meeting experience free of technical issues, such as those highlighted in section 2.1 regarding the problems of current meeting platforms.

This section concludes the research on the available literature to answer the question "what are the added values of ICSs that enhance online meetings?". To summarize the findings, it emerged that TP meetings are more effective in reaching four meeting objectives out of fifteen compared to AC and VC meetings; this insight is highly relevant to the goal of our research, as it helps us understand better for which meetings would it be more feasible to deploy the TNO ICS. Similarly, a study reported that people have a positive opinion of this new technology and feel comfortable in the immersive environment, thus deploying ICSs at TNO might also be met with a positive attitude by the employees. Then,

the other insights confirmed the initial claims of TNO on the potential of ICSs to address the drawbacks of non-immersive platforms. For example, VR technologies are found to help users focus better and perform more creatively. Another study reported that non-verbal communication in social VR is considered similar to offline face-to-face interaction, by allowing participants to easily share and recognize non-verbal cues. Regarding the learning experience, students using HMDs were more engaged in the lecture, were less anxious, and rated positively the overall experience. Lastly, cardiologists also found HMDs to be beneficial in their work lives; thanks to the immersive systems, they could easily work together on the same artifacts and rated positively VR as an alternative method for remote conferencing.

We have discussed the added values of ICSs for online meetings and the limitations of traditional online meeting platforms that they address. To ensure that users benefit from said added values, ICSs need to be evaluated before deployment; it might be the case that the technology is good in theory, but fails to deliver in the real world the values promised. The results of the evaluation will help the team uncover and address issues that could worsen the user experience with the system, and ensure that the technology can be deployed as a valid addition or substitution to non-immersive platforms. The theme of system evaluation will be researched in the next section.

2.2.4 LRQ4: How Can We Evaluate the Value of ICSs?

The evaluation of ICSs is essential to ensure they meet the standard expected by TNO and provide an adequate quality of experience, before deploying the technology across the company. In the following paragraphs, we will define the variables that impact the meeting experience in ICSs the most, and discuss the methods used for their evaluation.

QoS, QoE, and Their Components

We present in this section two scientific publications regarding the evaluation of a technological system. The first deals with multimedia services, while the second paper covers the evaluation of an ICS.

In the work of Alreshoodi and Wood [6], the concepts of Quality of Service (QoS) and Quality of Experience (QoE) are introduced in the context of evaluating the network quality of multimedia services. First, the authors explain how a QoS evaluation of networks focuses on objective and technical criteria to ensure their reliability and performance. Second, the paper emphasizes how the QoE evaluation takes into account the users' subjective experiences with the network.

Singh et al. [35] also provide a definition for QoS and QoE on the topic of evaluating an immersive system. According to the authors, the QoS is defined as "The quality of service comprises system performance measures that can impact the quality of experience", and the QoE is defined as "a multidisciplinary indicator that includes all the different aspects related to the experience that a user has in VR. It also includes the user experience".

To further understand how to evaluate the QoE of our ICS, we collected three of the most frequently mentioned subjective QoE factors included in the available scientific literature: (a) usability, (b) perceived effectiveness, and (c) quality of mediated social communication.

Usability We decided to employ usability as a measure for the meeting QoE because, according to the study of Singh et al. [35], it presents the strongest positive linear relationship with the QoE among the variables used in the study to evaluate the immersive

system, such as engagement, embodiment, and Quality of Interaction (QoI). Usability is popularly used for system evaluations and is defined by the International Organization for Standardization (ISO) [21] as "the effectiveness, efficiency, and satisfaction with which specified users can achieve goals in particular environments". Gabbard et al. [17] explain how usability problems prevent virtual applications from being useful, and only a careful usability evaluation can avoid such a problem.

Effectiveness Using Standaert et al. [37] methodology as our reference, we will investigate the perceived effectiveness of the meetings QoE, which the authors used as one of the criteria to determine which communication media to employ to reach the meeting's goals. Effectiveness is defined by the ISO as the "extent to which planned activities are realized and planned results are achieved"[2]. According to the literature review of Leach et al. [26], the employee's perspectives on meetings can impact not only their outcome, but also their capacity to accomplish their objectives. These perceptions might also influence broader job attitudes and the overall sense of well-being, consequently affecting more long-term choices like staying or not at their company. Lastly, the authors highlight that effective meetings should be encouraged to avoid undesired costs, considering the overall expenses associated with meetings, including salary and time employed.

Mediated Social Communication The third and last factor of the QoE that will be evaluated in this study is Mediated Social Communication. According to Toet et al. [40], the quality of a mediated social communication experience depends on the extent to which one feels like being physically together (spatial presence) and having an affective and intellectual connection (social presence) with another person. Mediated social communication is therefore highly related to the feeling of immersion provided uniquely by ICSs; for this reason, we decided to employ the quality of social communication as a measure for the meeting QoE. Mediated social communication refers to the way in which people exchange information through video conferencing software rather than face-to-face interactions to maintain their interpersonal relationships and well-being. To ensure that the capturing, modeling, and rendering techniques can convey a high-fidelity experience of the remotely connected partners and their physical environment, it is important to evaluate how the ICS can provide a coherent, realistic, and plausible social communication experience.

For the sake of simplicity, we will also refer to "mediated social communication" as just "social communication" throughout the rest of the thesis.

In the second part of the section dedicated to answering LRQ4, we described three of the QoE components that will be evaluated to assess the quality of an ICS, to understand whether it is ready to be deployed as an alternative to non-immersive platforms. Usability measures how well can users achieve their goals; effectiveness measures how much can users achieve their goals; social communication measures how immersed users feel during their activities. In our context, the "activities" mentioned in the section refer to work meetings, and the "goals" refer to the meeting objectives.

Next, we will present the most relevant methods, according to the literature, on how to evaluate the three QoE components.

Evaluation Methods

Regarding the best practices to evaluate usability, the systematic study of Fernandez et al. [16] analyzed the most relevant papers that employed Usability Evaluation Methods

(UEMs) for technological applications and artifacts. The authors determined that there is no single method that is most suitable for all circumstances and types, as it depends on the purpose of the evaluation and the type of artifact that is evaluated. Given the impossibility of determining the best UEM, Paz and Pow-Sang [33] showed that the most widely used UEMs by the scientific community are (a) usability tests, (b) heuristic evaluations, (c) questionnaires, and (d) interviews. We will discuss in the following paragraphs these four evaluation methods.

Best described by Bastien [8], usability tests are the most relevant evaluation approach, in which users directly participate, complete typical tasks with a product, or explore it freely, while their behaviors are observed and recorded to identify design flaws. The evaluators record the task completion time, task completion rates, and the number and types of errors. Lastly, based on the observations, design recommendations are proposed to improve the quality of the product.

As defined by Gabbard et al. [17] in their paper on the evaluation of VEs, heuristic evaluations involve experts assessing an interface against a set of established usability principles. They are useful when resources are limited and there is no time to conduct a usability test. Then, based on the findings, and especially the violations of such heuristics, the evaluators propose recommendations to improve the system.

As Bowman et al. [11] describe them, post hoc questionnaires are written sets of questions used to obtain demographic information, views, and interests of users after they have participated in a usability test. Also referred to as simply "questionnaires", they are good for collecting subjective data, reliably and consistently, on several aspects of an immersive experience, such as presence and fatigue.

As per Baxter et al. [9], interviews in the broadest sense are a guided conversation in which one person seeks information from another. Interviews differ greatly based on the goal of the evaluator: they could be (a) structured, (b) unstructured, or (c) semi-structured, and the data collected is either qualitative or quantitative. A structured interview is the most controlled type because the goal is to offer each interviewee the same set of possible responses; the interview may consist primarily of closed-ended questions, where the interviewee must choose from the options provided. An unstructured interview is the most similar to a normal conversation; the interviewer will begin with general goals but will allow the participant to go into each point with as much or as little detail and in the order he or she desires. A semi-structured interview is a combination of the structured and unstructured types; the interviewer may begin with a set of questions to answer (closed-ended and open-ended) but deviate from the order and even the set of questions from time to time. Interviews are flexible and can be used as a solo activity or in conjunction with another evaluation method.

We have discussed in this section how usability tests, questionnaires, heuristic evaluations, and interviews are the most popular evaluation methods chosen to evaluate the usability, effectiveness, and quality of social communication; which are part of the meeting QoE. The evaluation methods presented in this section can all be applied to the evaluation of ICSs. However, not all four usability evaluation methods can be used to assess the effectiveness and social communication as well; usability tests and heuristic evaluations are exclusive to assessing the usability of a system, while questionnaires and interviews can also be employed to evaluate the effectiveness and social communication in immersive technologies.

This section concludes the LRQ4, and with it, the Literature Review chapter. To summarize, the answer to the first literature review question ("what are ICSs?") is that we can

describe ICSs according to their technology employed (e.g. AR, Fish Tank VR, HMDs, CAVE), and their context of use (e.g. for meetings, social interactions, education, training). Then, the answer to LRQ2 ("what are online meetings?") is that online meetings are virtual gatherings described by their objectives, capabilities, modes, intentions, participants, format, and criticalities. The answer for LRQ3 ("what are the added values of ICSs for online meetings?") resides in the feeling of social presence given by the immersive nature of communication systems. Such factors are directly responsible for improving several aspects of meetings, such as (but not limited to) their effectiveness, product creative quality, communication, the learning experience, and real-time interactions. Lastly, for LRQ4, usability, effectiveness, and mediated social communication are the three main factors of the QoE needed to evaluate the quality of ICSs. The evaluation can be performed by conducting usability tests, questionnaires, heuristic evaluations, and interviews.

After the broad and comprehensive research on ICSs, we will now discuss specifically the study conducted on the ICS developed by TNO.

2.3 Previous Studies on ICSs at TNO

Our work is strongly connected to the research by Singh et al. [35], as it represents the initial effort by the Social XR team to investigate ICSs as a potential solution for the shortcomings of online meetings. In particular, their study investigated the elements that lead to meeting engagement in a social VR environment. The authors conducted multiple user testing sessions of the TNO ICS, followed by a questionnaire to assess the participants' experience. The research concluded that quality of communication is the most important element for creating engaging social VR meeting experiences. Furthermore, there is a strong linear relationship between the quality of communication, immersion, social presence, and embodiment in ICSs. Unfortunately, one of the main limitations encountered regarded the usability of the system, both on the hardware and software side. For example, the usability of several features, the navigation usability, and the hardware quality were considered to have decreased the Quality of Service (QoS), Quality of Interaction (QoI), and general usability of the system. Furthermore, the system was not deemed to be of adequate quality by the participants.

This section presented the first study on the evaluation of the TNO ICS. This work helped us understand how we can apply the insights gathered from the Literature Review to our evaluation of the system. The background research, aimed at establishing the context and relevance of our experiment, is now concluded.

2.4 Background Discussion and Conclusion

In this chapter, we first presented in this chapter the problem statement of our study, regarding the subpar meeting experience offered by traditional meeting platforms because of their several issues. We then reported the assumptions of the TNO team, which advanced the idea that ICSs can address said issues and offer a better meeting experience. Later, we formulated our high-level goal, i.e. understanding the feasibility of deploying ICSs for business meetings by evaluating the meeting experience provided. The scientific material we collected to answer LRQ1 provided us with extensive knowledge on the definition of ICSs that TNO proposed as a solution. The LRQ2 helped us understand in depth the context of the problem statement, namely online meetings. Our answer to LRQ3 shows that ICSs have the potential to perform better than traditional meeting platforms, thanks

to their immersive space, as the SocialXR team initially supposed. Lastly, in the section related to answering LRQ4 we explained that to reach the high-level goal of this study, we need to assess the meeting QoE by evaluating the system usability, effectiveness, and quality of social communication. Figure 2.3 summarizes the content and structure of the Background chapter.

We will define five research questions in the following chapter to understand whether the TNO ICS offers the added values required to address the drawbacks of non-immersive platforms, making it a viable alternative that can be deployed in the company.

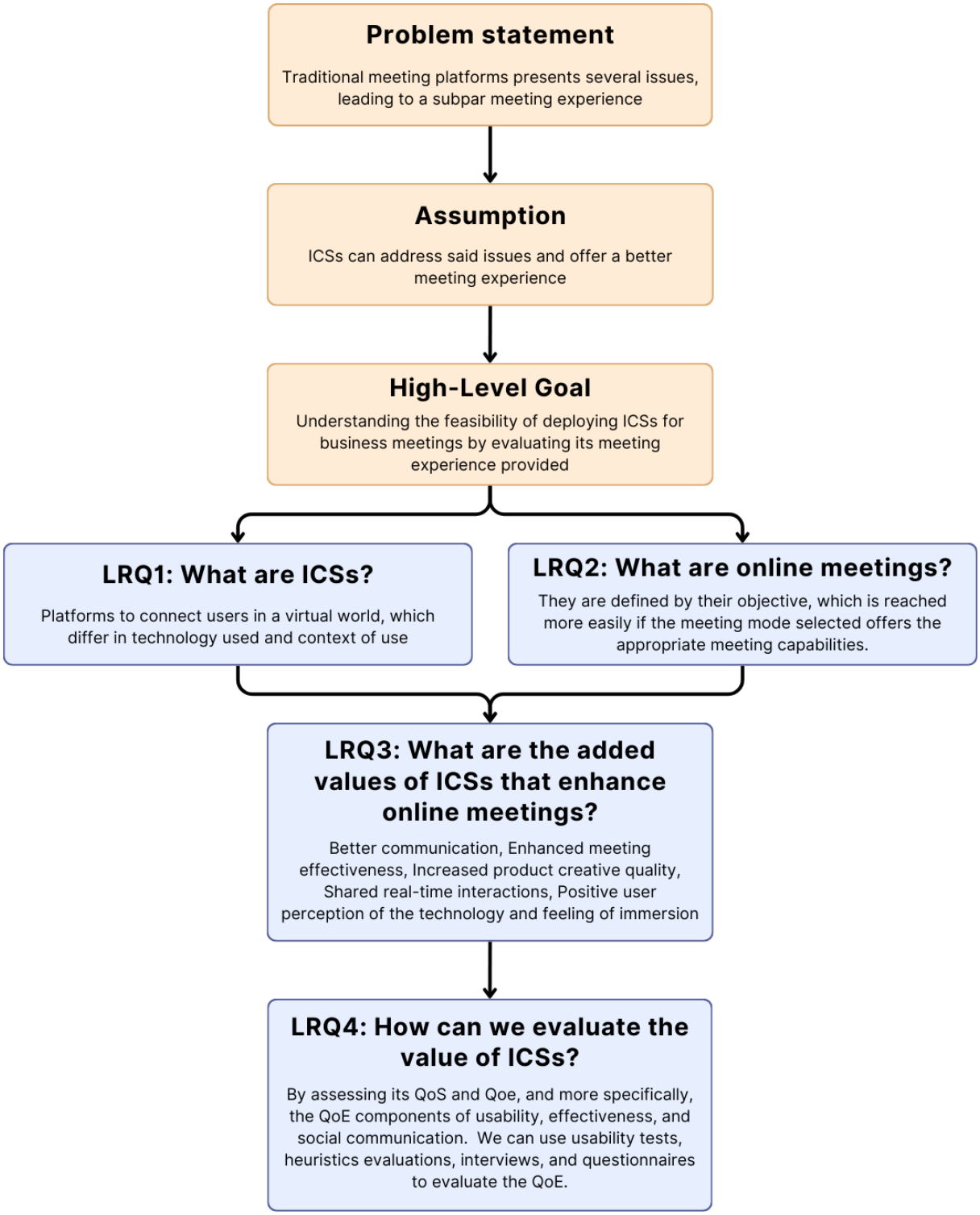


Figure 2.3: The graphical overview of the Background chapter.

Chapter 3

Research Question

After studying the characteristics of ICSs, we will now continue our research with the ICS designed by TNO. This chapter defines five main Research Questions (RQs): each of the first three RQs is related to one of the three subjective factors that are part of the meeting QoE; RQ4 and RQ5 are connected to other factors that, despite their minor impact, can still have an effect on the meeting's QoE. It is outside the scope of this thesis to evaluate the QoS of our ICS, which is presented in section 2.2.4, and includes the evaluation of objective factors such as network stability, data transfer speed, and latency. For each RQ we will present their goal and their rationale; for the first three RQs, we will also describe their hypotheses. Although this section should solely present the Research Questions, we decided to include their methodology as well, instead of presenting them after the chapter on the experiment design; we believe that by taking this approach, our study description would be easier to understand and less repetitive. It should also be noted that, although we describe the methodology used for each RQ, the experiment design does not change between them, as we prepared one experiment to answer all RQs.

The overarching RQ, which captures the collective goal of the five RQs, is defined as "how does the perceived QoE of meetings change for users in different conditions?". The goal is to understand how the different factors that influence the QoE play a role in experiencing the meetings in an ICS. It will help us understand the strengths and weaknesses of the ICS, the improvements needed, and the meeting conditions in which should or should not be used. We employ the five RQs to also investigate the meeting QoE of MS Teams, the platform used at TNO for online meetings; this allows us to compare the QoE of the two communication media, to better understand whether or immersive platform can be a valid substitute for MS Teams. Answering the overarching RQ concurs to reaching the high-level goal of the thesis, which is to understand the feasibility of deploying the TNO ICS for work meetings, by analyzing the meeting QoE.

3.1 RQ1: Usability

In this section we describe the RQ1; we investigate the first component of QoE in meetings, usability, in the TNO ICS ans MS Teams. We formulate RQ1 as "How does usability change in different communication media?"; our goal is to compare the users' perceived usability of the TNO ICS and MS Teams.

As mentioned in section 2.2.3, we can expect our ICS to receive positive usability ratings, comparably to what was reported in the study of Gunkel et al. [19] regarding a similar VR technology. However, our system is not yet fully developed, which might negatively influence its usability. Additionally, MS Teams has been the standard software

for online meetings at TNO for years, so employees might rate its usability higher because they are used to it.

For RQ1, the null hypothesis is that there is no difference in usability scores between the TNO ICS and MS Teams. Our alternative hypothesis is that there is a difference in the usability of communication media, as MS Teams is expected to perform better than the TNO ICS. If this null hypothesis is not rejected, or if the null hypothesis is rejected and MS Teams performed significantly better than the TNO ICS, we will analyze the results, draw conclusions, and present them to the team to understand where the system is lacking, and how it could be improved. However, if the null hypothesis is rejected and the TNO ICS performs significantly better than MS Teams, we will have tangible proof that the usability of the TNO ICS is better than the usability of MS Teams, addressing the technical issues of non-immersive platforms presented in the problem statement, and supporting the argument for the TNO ICS to be a viable alternative to MS Teams.

3.2 RQ2: Effectiveness

In this section we describe the RQ2; we investigate the second component of QoE in meetings, effectiveness, in the TNO ICS and MS Teams.

RQ2 is divided into two sub-research questions. The first states "How does the perceived effectiveness change in different communication media?". Our goal is to compare the effectiveness of meetings in the TNO ICS and MS Teams. We formulate the second sub-RQ2 as "How does the perceived effectiveness change in different meeting types?". Our goal is to compare the effectiveness of work meetings.

According to the studies of Standaert et al. [39, 38, 37], different capabilities (e.g. hear attendees' voice, discerning the attendee's facial expression) are required to reach the meeting goals (e.g. exchanging information, taking decisions), as seen in table 2.1. Additionally, depending on the meeting capabilities they offer, certain communication media are more effective in achieving specific goals. For example, meeting in a non-immersive meeting platform, such as MS Teams, will be sufficient for "exchanging information" meetings, as participants only need the capabilities of hearing the attendees' voices and sharing their screens to perform effectively; instead, an ICS might be needed for "making decisions" meetings, as participants need more capabilities to reach the meeting goals, such as to experience co-location and discern the attendees' facial expressions.

For RQ2 we defined two null and alternative hypotheses, one for each sub-RQ, and we expect the same effect for both. The first null hypothesis of RQ2 is that there is no relationship between perceived effectiveness and meeting type. Our alternative hypothesis is that meeting type has an effect on the meeting effectiveness; more specifically, we anticipate meetings that need more capabilities to receive higher ratings than meetings that require less capabilities in the TNO ICS condition, and worse ratings in the MS Teams condition. The second null hypothesis of RQ2 is that there is no relationship between perceived effectiveness and communication media. Our alternative hypothesis is that communication media has an effect on the meeting effectiveness; more specifically, we expect the TNO ICS to perform better than MS Teams in meetings that require more capabilities, and worse in meetings that require fewer capabilities. If the null hypotheses are rejected, we will have tangible proof that meetings are more effective if held in the TNO ICS rather than MS Teams, addressing the engagement and effectiveness issues of non-immersive platforms presented in the problem statement, and supporting the argument for the TNO ICS to be a viable alternative to MS Teams for meetings that require more capabilities (or less, depending on the results). In the other case, we will analyze the results, draw conclusions,

and present them to the team to understand where the system is lacking, and how it could be improved.

3.3 RQ3: Social Communication

In this section we describe the RQ3; we investigate the third component of QoE in meetings, social communication, in the TNO ICS and MS Teams.

As described by Toet et al. [40], social communication is divided into spatial presence and social presence. Therefore, RQ3 is divided into four sub-questions, two for each social communication factor: the first two sub-RQs cover the spatial presence factor, and the latter two cover social presence. The two sub-RQs of the spatial presence factor are expressed as follows: "How does the spatial presence change with different levels of immersion?" and "How does the spatial presence change in different communication media?". Our first goal is to compare the spatial presence perceived by participants in a meeting setting where they feel more or less immersed and together with their colleagues. Then, the second goal is to compare the spatial presence in the TNO ICS and MS Teams. The last two sub-RQs on social communication investigate instead the social presence in online meetings. The sub-RQs read as "How does the social presence change with different levels of immersion?" and "How does the social presence change in different communication media?". The goal of these sub-RQs is parallel to the one expressed before, but it deals with social presence instead. The rationale for RQ3 is that spatial and social presence are highly correlated to the feeling of immersion provided by the communication media, as explained in sections 2.2.1, 2.2.4, and 2.3.

For RQ3 we defined four null and alternative hypotheses, and we expect the same effect for all of them. The first null hypothesis of RQ3 is that there is no relationship between communication media and spatial presence. The alternative hypothesis is that the different communication media have an effect on the spatial presence perceived. We expect participants to perceive more spatial presence in meetings in the TNO ICS rather than in MS Teams. The second null hypothesis is that there is no relationship between the level of immersion and spatial presence. The alternative hypothesis is that different levels of immersion have an effect on the spatial presence perceived. We expect participants to experience more spatial presence in meetings with more immersion provided. The third and fourth null and alternative hypotheses are formulated identically to the first two, but they investigate social presence instead. If the null hypotheses are rejected, and the TNO ICS performs significantly better than MS Teams, we will have tangible proof that the TNO ICS provides a higher quality of social communication than MS Teams, addressing the communication issues of non-immersive platforms presented in the problem statement, and supporting the argument for the TNO ICS to be a viable alternative to MS Teams. In the other case, we will analyze the results, draw conclusions, and present them to the team to help them understand where the system is lacking, and how it could be improved.

3.4 RQ4: Extraneous Variables

We formulate RQ4 as "How do extraneous variables influence the subjective QoE factors"? The QoE might also be affected by other variables such as the user's age, education, or experience with VR systems. Our goal is to understand whether there are any hidden patterns, connections, and trends between the extraneous variables and the meeting QoE. We do not formulate individual hypotheses for these effects, however, we are still controlling for a number of these variables. We will conduct an exploratory analysis to understand

whether there are any patterns in the data between the extraneous variables and the meeting QoE. It will be up for future research to conduct an eventual explanatory analysis of these trends. Bandhari [10] defines such variables as "extraneous variables", factors not investigated that can potentially affect the outcomes of a research study. According to the author, there are four types of extraneous variables: (a) demand characteristics, cues that encourage participants to conform to researchers' behavioral expectations; (b) experimenter effects; unintentional actions by researchers that can influence study outcomes; (c) situational variables, which can alter participants' behaviors in study environments; and (d) participant variables, characteristic or aspect of a participant's background that could affect study results, even though it's not the focus of an experiment. We anticipate the participant variables to play a significant role in the perceived QoE.

3.5 RQ5: Other Explanatory Insights

We formulate RQ5 as "Which other subjective factors influence the meeting QoE?". Other subjective factors include, for example, Quality of Interactions (QoI) and embodiment, which were both highly positively correlated to the meeting QoE according to Singh et al. [35], but were left out of our study because there was not enough existing literature to support them as the main influencing factors of the QoE. Our goal is to conduct an exploratory analysis to understand whether there are any patterns in the data between factors different than those mentioned in the first four RQs and the meeting QoE. It will be up for future research to conduct an eventual explanatory analysis of these trends.

The chapter on the experiment RQs is now concluded. We defined five research questions to address the high-level goal of understanding the feasibility of deploying the TNO ICS for business meetings, by analyzing the subjective factors of the meeting QoE. The first RQ aims to evaluate the usability of the two communication media; the goal of the second RQ is to investigate the effectiveness of different meeting types in the TNO ICS and MS Teams; the third RQ evaluates the social communication in the two meeting platforms and between two and three participants; the fourth RQ covers our exploratory analysis; lastly, with the fifth research question we aim at understanding whether any other factors play a role in the QoE.

The next chapter will outline the methodology for applying these research questions in the experiment to gather the desired information.

Chapter 4

Experiment

In the previous chapters, we defined the problem statement of the proposed research, the literature review at the base of our assumptions, and the research questions for our investigation. Now, we will describe the experiment that will help us answer the RQs.

The participants of the experiment evaluated two communication media, the TNO ICS and MS Teams; we ground our design choices of the experiment regarding communication media on the studies discussed in LRQ1. Participants interacted with each other and with the communication media during meetings; we ground our design choices of the experiment regarding communication media on the studies discussed in LRQ2. Participants filled in questionnaires and participated in interviews to rate the subjective factors of the meeting QoE; we ground our design choices of the experiment regarding the evaluation of communication media on the studies discussed in LRQ3 and LRQ4.

This chapter is divided into four sections to describe the user study we conducted, focusing on (a) the study design, the variables used, the content and tasks of each meeting, the experiment conditions, and the evaluation methods; (b) the study population and their demographic; (c) the system design; and (d) the procedure of research, pilot testing, the user testing details, the issues that arose, and the questionnaire and interviews.

4.1 Study Design

The context and motivation of the RQs that we have discussed in chapter 3 provide a theoretical "framework" for the experiment, while the experiment methodology translates the framework into practical implementation. This section on the study design is divided into the variables used, the content of the meetings for the evaluation, the conditions, and the evaluation methods employed.

4.1.1 Variables

This section describes the dependent and independent variables used in the experiment, the rationale behind our selection of each, and how they related to each RQ.

To answer the first three RQs, we used the three main components of the QoE as dependent variables, and communication media, meeting type, and number of participants as independent variables. The first and second independent variables are within participants, while the third independent variable is between participants.

The goal of the first RQ is to compare the usability of communication media; the first dependent variable employed in the experiment is usability, and the independent variable

is the communication media. In our experiment, users participated in TNO ICS and MS Teams meetings to evaluate their usability.

The aim of RQ2 is to understand how the perceived effectiveness changes in communication media and in different meeting types. The dependent variable of the second RQ is "effectiveness", and the independent variables are meeting types and communication media. More specifically, participants took part in the meeting types "making decisions" and "exchanging information", defined by Standaert et al. [39, 38, 37]; since the authors defined the meeting types based on the existing literature and did not indicate which meeting type is the most frequent in companies, we chose "exchanging information" and "making decisions" as we believe them to be the two most popular objectives for company meetings in TNO. Furthermore, the first meeting type requires fewer meeting capabilities to reach the meeting goals, while the second meeting type requires more capabilities; for this reason, the authors suggested that non-immersive platforms should be used for "exchanging information" meetings, and immersive platforms should be used for "making decisions". The goal of RQ2 is also to verify this statement and understand whether immersive platforms could also be the preferred choice for "exchanging information" meetings. In our experiment, users participated in "exchanging information" and "making decisions" meetings in the TNO ICS and MS Teams to evaluate the meeting effectiveness.

The aim of RQ3 is to understand how the quality of social communication changes in different communication media and in different levels of immersion. The dependent variable is social communication; the independent variables are communication media and number of participants. According to Oh et al. [32], some communication media are more capable of delivering these verbal and nonverbal cues, while others are not, emphasizing that social presence is a quality of the communication media itself. Additionally, social presence is highly influenced by social cues and the presence of others; having more sources of social cues (i.e. more participants) is related to an increase in social presence. The goal of RQ3 is also to verify these two statements, thus understanding whether a change in the number of participants and social media has an effect on social communication. It should be noted that, as Toet et al. 2.2.4 discussed, the definitions of social presence and social communication are vague and not clearly defined. Therefore, we assume that we can generalize the results of the study of Oh et al. to spatial presence as well. In our experiment, we divided the users into two groups; the first participated in meetings with three users, to simulate a higher level of immersion, while those in the second group participated in meetings meetings with two users, to replicate lower levels of immersion. More specifically, each participant was assigned to the same meeting partner(s) for all meetings; half of the meetings were in the TNO ICS, and the other half in MS Teams.

Since RQ4 and RQ5 have a predominantly exploratory nature, they do not have specific variables or factors to investigate, and we are uncertain about the direction and magnitude of the effects described. In RQ4, other than studying the effect of the extraneous variables, we are interested in exploring the interaction effect between independent variables related to the other dependent variable (e.g. number of participants and effectiveness). Lastly, we analyze the effect that dependent variables have on each other.

Figure 4.1 reports the graphical summary of the RQs we defined to guide the experiment, their connection to the high-level goal, the dependent and independent variables, and the levels of each independent variable. The bold lines between the dependent and independent variables show the main relationship that we analyzed in RQs one to three; the thin dotted lines between the independent and dependent variables indicate the potential interaction effect that we are researching in RQ4.

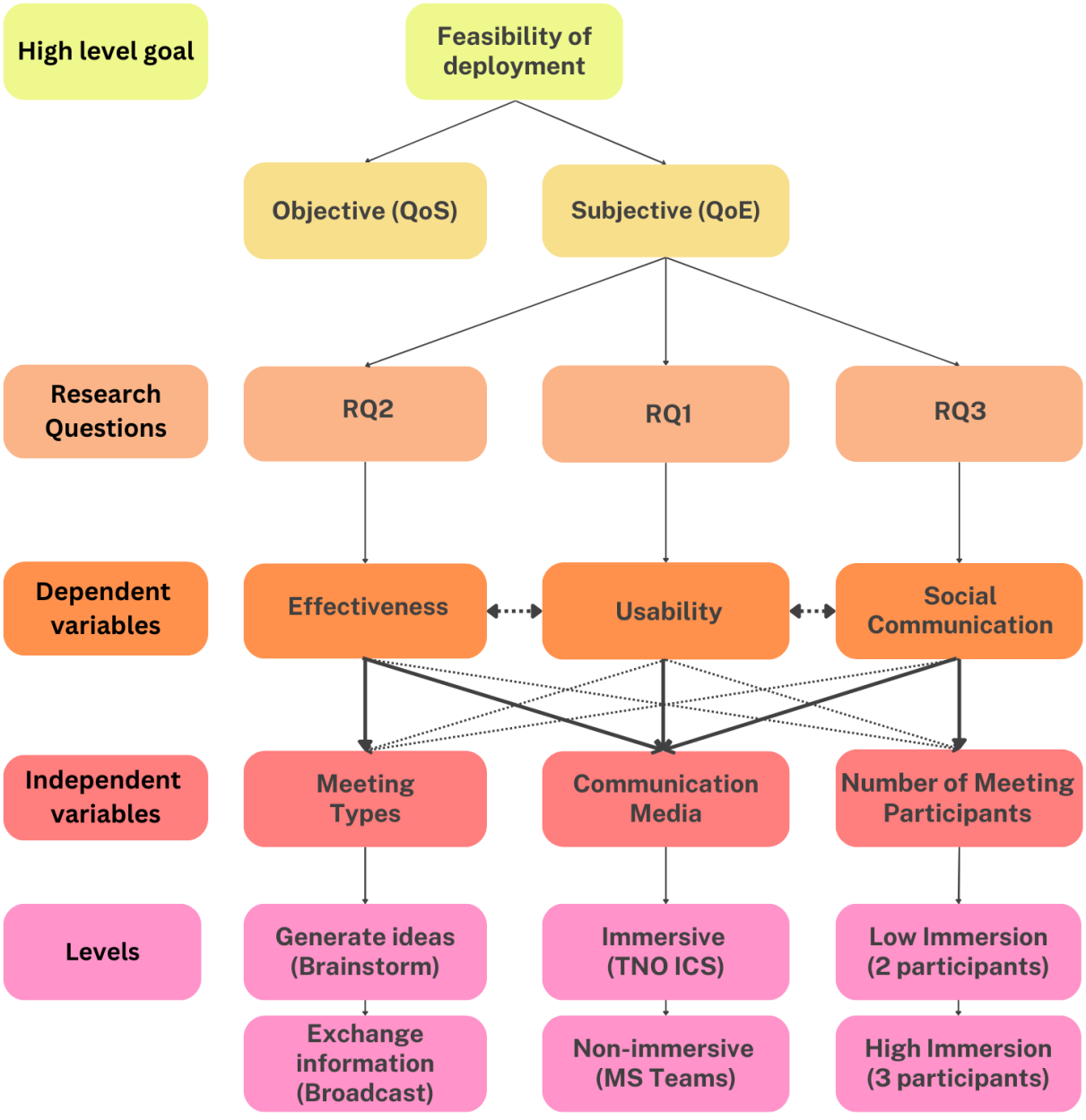


Figure 4.1: The relationship between the high-level goal and the variables.

4.1.2 Meeting Content and Tasks

This section describes the tasks that participants completed during their meetings. The first goal of the tasks is to allow the users to actively experience the Metaverse and interact with it and their colleagues; the second goal of the tasks is to simulate brainstorm and broadcast meetings, without having to ask employees to talk about their work, which might be confidential. We use "broadcast" and "brainstorm" to indicate the subcategories of meeting objectives "exchange information" and "generate ideas", respectively.

All participants completed six tasks with their partner(s), across four meetings. The goal of each task is related to the meeting type. More precisely, for the "brainstorming meetings" participants had to generate as many ideas as possible for the following four topics: (a) tiny people problem, (b) thumb problem, (c) tourist problem, and (d) hanger problem. Task (a) deals with listing all the consequences of people waking up and finding themselves very tiny. Task (b) involves thinking about all the implications of having an additional thumb on each hand. Task (c) is about generating ideas on ways to lure more American tourists to The Netherlands. Lastly, task (d) requires users to come up with as many ways as possible of using a normal hanger. All these topics are taken from the book of Isaksen [20], which reviewed 50 empirical studies that employed brainstorming techniques and listed the tasks used in their experiments. The four topics chosen for this study are among those that were employed the most in previous studies. The codes T1, T2, T3, and T4, for practicality, were used in the study to identify the four tasks mentioned.

Then, for the "broadcast meetings", participants discussed, chose, and agreed on the importance of given items to survive in particular situations. The tasks are called "desert survival" and "lost at sea". In these tasks, participants share their ideas and collaborate to order a given list of objects according to the importance of having them in such dangerous situations (namely being on a boat in the ocean and being stranded in the desert), to increase their chances of surviving while searching for help. "Lost at sea" was first defined by Nemiroff and Pasmore [30], while "desert survival" was designed by Lafferty et al. [24]; both were described in the work of Littlepage et al. [27] as they are tasks extensively used in previous studies for similar purposes. The codes we used in the experiment to identify these two tasks are respectively T5 for "desert survival" and T6 for "lost at sea". For both tasks, their complete description and their list of items that participants had to rank are reported in chapter A of the Appendix.

Table 4.1 summarizes the tasks, their codes, and their description.

4.1.3 Meeting Conditions

This section describes the Each meeting of our experiment is defined by the conditions related to it. Our goal is to evaluate the QoE for each of these meeting conditions to understand which one is the TNO ICS most feasible to be deployed.

The independent variables "communication media" and "meeting type" are within subjects. The combination of the two levels of these two variables creates four different meeting conditions. More in detail, condition 1, or C1, represents a brainstorming meeting in MS Teams; C2 represents a brainstorming meeting in the TNO ICS; C3 represents a broadcast meeting in MS Teams; lastly, C4 represents a broadcast meeting in the TNO ICS. All users participated in four meetings, each characterized by a different condition.

The third independent variable, "number of participants", is instead between subjects, as seven groups are consistently formed by two employees, as opposed to six groups with three users for all meetings. Consequently, users assigned to groups of two people participated in meetings characterized by conditions C1 to C4. Users assigned to groups of

Meeting type	Code	Name	Description
Brainstorm	T1	Tiny People Problem	Discuss the consequences if all people, upon waking up, would find themselves very tiny
	T2	Thumb Problem	Discuss the consequences if all people wake up with an additional thumb on each hand
	T3	Tourist Problem	Discuss the ways to lure more American tourists to The Netherlands
	T4	Hanger Problem	Discuss all possible ways to use a coat hanger
Broadcast	T5	Desert Survival	Choose the items to survive in the desert after a plane crash
	T6	Lost at Sea	Choose the items to survive on a raft after the main boat sunk

Table 4.1: A summary of the experiment tasks.

three people participated in meetings characterized by the same conditions, but we used the codes C5 to C8 for these conditions to differentiate them between those with groups of two participants and groups of three participants.

Grouping and naming the meeting conditions under specific codes was crucial for the experiment management, as it helped in providing a quick and clear overview of the details for each meeting.

Table 4.2 provides an overview of the eight meeting conditions, divided by the levels of each independent variable. To provide an example, meeting condition C6 indicates a brainstorm meeting with three participants in the TNO ICS.

				Communication Media	
				MS Teams	TNO ICS
Meeting Participants	2	**Meeting Type**	Brainstorm	C1	C2
			Broadcast	C3	C4
	3	**Meeting Type**	Brainstorm	C5	C6
			Broadcast	C7	C8

Table 4.2: A summary of the experiment conditions.

A	B	C	D
B	C	D	**A**
C	D	**A**	B
D	**A**	B	C

Table 4.3: A latin square of four elements.

4.1.4 Meeting Order

Each of the 13 groups, six with three people, and seven with two people, was assigned to one of four letters: A, B, C, D. These four letters each symbolize a unique order of the four meetings that groups had to follow, taking into consideration the experiment conditions (described in 4.1.3) and tasks (described in 4.1.2). The experiment conditions and tasks are ordered according to the latin square design, to have a controlled randomization of the user experience with the different meetings. A Latin square is an n × n array filled with n different elements, each occurring exactly once in each row and exactly once in each column [1]. In the context of our study, it implies that participants participate in each meeting condition exactly once and in a different order. Latin squares are useful to reduce order effects when designing experiments with multiple conditions. Table 4.3 represents a 4x4 latin square, the same one we used in our experiment, highlighting how the element A occurs exactly once in each row and column.

Distributing the meetings according to the latin square counterbalanced the "first timer-effect" and "habituation effect", ensuring that they would have not interfered with the meeting experiences, thus preventing other extraneous variables from influencing the study. As Singh et al. [35] present them, the first-timer effect can create a "wow" effect and lead to a higher feeling of excitement when first using a VR system. However, the excitement quickly decreases in the following interactions, as a result of the habituation effect. A similar phenomenon is related to VR-induced symptoms, such as motion sickness; these symptom levels are highest on the first immersive meeting and reduced to negligible by the third experience.

Table 4.4 summarizes the meeting task. the meeting conditions, and meeting order. For example, the four groups assigned to the letter A had their first experience to be a brainstorm meeting in MS Teams (conditions C1 or C5, depending on the number of participants), and had to discuss tasks T1 (tiny people problem) and T2 (thumb problem).

4.1.5 Evaluation Methods

In this section, we present the evaluation methods used to gather the data needed to answer the five RQs. We answered the first four RQs through the quantitative data gathered from the questionnaires; instead, to answer this research question we collected and analyzed the qualitative data of the interviews. We used the SUS questionnaire to collect the data about usability to answer RQ1. The System Usability Scale (SUS) questionnaire designed by Brooke [12] provides a measure of people's subjective perceptions of the usability of a system. The evaluation score determines whether the usability of the system is excellent, good, ok, poor, or the worst. We used the "Perceived Usefulness" questionnaire as a

[1] https://en.wikipedia.org/wiki/Latin_square

Team	Meeting and Task ID			
A	C1/C5. T1, T2	C2/C6. T3, T4	C3/C7. T5	C4/C8. T6
B	C2/C6. T2, T4	C4/C8. T6	C1/C5. T1, T3	C3/C7. T5
C	C4/C8. T5	C3/C7. T6	C2/C6. T4, T3	C1/C5. T2, T1
D	C3/C7. T6	C1/C5. T3, T1	C4/C8. T5	C2/C6. T4, T2

Table 4.4: A summary of the order of the experiment conditions and meeting tasks for each team.

measure to assess the meeting's effectiveness to answer RQ2. The Measurement Scales for Perceived Usefulness and Perceived Ease of Use [14] were initially developed to be valid measurement scales for predicting user acceptance of computers, and have been used in countless empirical research studies on different technical systems. The ten questions of the questionnaire are divided into the sections of Perceived Usefulness and Perceived Ease of Use; since the second component is fairly similar to some of the SUS questionnaire, we decided to only employ the first five questions to avoid redundancy. It is important to notice that the effectiveness (or usefulness) of a system is hereby considered a subjective factor, and it does not take into consideration the objective nature of meeting effectiveness, such as the outcome of said meeting. Therefore, we will be referring to its subjective part when mentioning effectiveness in the following chapters of the thesis. Lastly, we used the H-MSC-Q (Holistic Mediated Social Communication Questionnaire) from Toet et al. [40] as a measure to assess the social communication provided in the meeting to answer RQ2. the H-MSC-Q is aimed at evaluating the two components of social communication, namely spatial and social presence, provided by immersive platforms. The questionnaire focuses on the perceived realism, plausibility, and coherence of the experience, and is specifically designed to determine the quality of a mediated experience.

We employed interviews as an evaluation method to gather the data needed to answer RQ5. In RQ5, our goal is to conduct an exploratory analysis to understand which other subjective factors influence the QoE. We also expect to gather insights to understand the major pains and needs of our users, and how to address them; the design suggestions, resulting from the data analysis, will help the team in improving the immersive system. There are no predetermined sets of interview questions for the evaluation of ICSs. To best investigate the meeting QoE, the interviews focused on uncovering influential factors beyond usability, effectiveness, and social presence. The interview questions and the three questionnaires are in chapter B and C of the Appendix.

This chapter is summarized in table 4.5. The RQ definitions in this table are more detailed than the one we presented in Chapter 3, as they now include the specific dependent variables that we evaluate. The fourth column, "Hypothesis/Description", presents the hypothesis of the first three RQs, however, for RQs 4 and 5 we included their description, as they are part of our exploratory analysis, and we are not expected to formulate their hypothesis.

ID	RQ definition	Instrument used	Hypothesis/Description
Overarching RQ	How does the perceived QoE of meetings change for users in different conditions?	All of the below	The independent and extraneous variables will have an effect on the QoE
RQ1	How does usability change in different communication media?	SUS [12]	Usability of MS Teams >Usability of TNO ICS
RQ2.1	How does the perceived effectiveness of meeting types change in different communication media?	Perceived Usefulness [14]	TNO ICS: Eff. Brainstorm >Eff. Broadcast. MS Teams: Eff. Broadcast >Eff. Brainstorm
RQ2.2	How does the perceived effectiveness of communication media change in different meeting types?	Perceived Usefulness [14]	Brainstorm: Eff. TNO ICS >Eff. MS Teams. Broadcast: Eff MS Teams >Eff. TNO ICS
RQ3.1.1	How does the spatial presence of communication media change with different numbers of meeting participants?	H-MSC-Q [40]	3 and 2 participants: S.P. of ICS >S.P. of MS Teams
RQ3.1.2	How does the spatial presence of meeting participants change in different communication media?	H-MSC-Q [40]	TNO ICS and MS Teams: S.P. with 3 >S.P. with 2
RQ3.2.1	How does the social presence of communication media change with different numbers of meeting participants?	H-MSC-Q [40]	3 and 2 participants: S.P. of ICS >S.P. of MS Teams
RQ3.2.2	How does the social presence of meeting participants change in different communication media?	H-MSC-Q [40]	TNO ICS and MS Teams: S.P. with 3 >S.P. with 2
RQ4	How do extraneous variables and further interaction effects influence the subjective QoE factors?	All the above	The extraneous variables or further interaction effects might explain better the variations in the data
RQ5	Which other subjective factors influence the meeting QoE?	Interviews	Other factors that could not be observed through a questionnaire might affect negatively or positively the meeting experience

Table 4.5: A summary of the five RQs.

We discussed in this section the three dependent and three independent variables used in the experiment and their relationships. Secondly, we proposed the six tasks, two for broadcast and four for brainstorm, that participant completed during their meetings. Then, we described the eight meeting conditions, labeled from C1 to C8, which indicate the type of meeting each participant took part in, given by the combination of communication media, meeting type, and number of participants. The next section will discuss the study population of the experiment.

4.2 Study Population

In this section, we report the number of participants in the experiment, the participants' recruitment, and the participants' demographic.

4.2.1 Recruiting Participants and Sample Size

The empirical study of Tullis and Stetson [41] aimed at understanding the variation of the statistical significance of evaluation methods depending on different sample sizes. In their experiment, they asked participants to perform significant but simple tasks on two websites and evaluate them by filling in different questionnaires (e.g. SUS, QUIS[2], and CSUQ[3]). The goal of the questionnaires was to determine which website did the user prefer. As one would expect, the accuracy of the questionnaires increases as the sample size gets larger. With a sample size of only 6, all of the methods yield an accuracy of only 30-40%, meaning that 60-70% of the time, at that sample size, one would fail to find a significant difference between the two systems analyzed. On the other hand, most of the questionnaires appear to reach an asymptote at a sample size of 12, where questionnaires yield an accuracy between 70% and 100%. The improvement observed by going to a sample size of 14 is small, or even non-existing. We chose to have 32 participants in our study, divided into six groups of three people, and seven groups of two, for a total of thirteen groups, which should be enough to test for statistical significance in the results, at least for the SUS questionnaire, according to the study of Tullis and Stetson. We expect that the Perceived Effectiveness and H-MSC-Q questionnaires of our experiment follow the trends described in their study. The number of participants needed was also decided by considering the limitation of the budget at hand. More specifically, to entice participants to participate in all meetings, the team decided to compensate them for their time with a coupon valid on bol.com of the value of 60 euros.

We recruited our participants by designing a poster (figure E.1 in the Appendix) and sharing it in the company, together with sending a message on the company platform for employee communication. The inclusion criteria included being over 18 years of age, being proficient in English, being a TNO employee, having access to the offices of TNO Den Haag New Babylon or TNO Leiden, and being prepared to do four meetings. The exclusion criteria included having a history of motion sickness or other physical conditions that would make using a VR headset problematic, having a history of psychiatric or neurological conditions that would interfere with their ability to participate in the study, and not allowing us to collect their data or record their interviews. For the sake of simplicity, we used the "convenience sampling" method to recruit our participants rather than selecting them based on their demographic information.

[2]https://digital.ahrq.gov/health-it-tools-and-resources/evaluation-resources/workflow-assessment-health-it-toolkit/all-workflow-tools/questionnaire

[3]https://garyperlman.com/quest/quest.cgi

In total, 41 employees from different TNO locations signed up for the study, but only the first 32 were selected, based on the principle of "first come first serve"; the others were added to a waiting list. There were 4 dropouts in total: two of them quit before the start of the experiment, and were replaced with two other employees from the waiting list. Unfortunately, two participants were forced to leave the experiment due to lack of time after the start; since they were in the same group, which did not participate in a meeting yet, they were substituted with as many employees, thus the group managed to conduct all meetings and interview sessions before the deadline and with the same set of participants.

4.2.2 Participants' Demographic

For the purpose of this experiment, the personal data of the 32 participants we collected was their name, surname, email, user ID, sex, age, location, ability to use digital devices, level of experience with VR technologies prior to the experiment, and whether they knew their meeting partner(s) before the meeting. The TNO Ethical Commission supervised and approved the experiment plan and ensured that all data was processed according to the company regulations. All participants signed and read the documents regarding the informed consent and participant information before starting the experiment. The data was confidential, as each participant received privately an ID code that would have identified them in the questionnaires they filled out and the interviews they participated in. For statistical purposes, the variable "participants' age" was divided into two intervals, based on the median age. The first group consisted of 17 participants under the age of 28 (not included), while the second group included 15 participants aged 28 or older. Overall, our participants were relatively young (n = 32, mean = 33.65, median = 28, SD = 11.8, min = 22, max = 61). We believe that our participants consisted of mostly young people for two main reasons: first, interns had more availability in their agenda, and were more appealed by the monetary compensation; secondly, young people are typically more willing to experiment with and welcome new technologies. Nine employees participated in the experiment with a colleague they already knew, meaning that four of the thirteen groups had employees who were already used to communicating with each other. On the other hand, eleven participants indicated that they did not know the partner(s) they had been grouped with. The remaining twelve individuals answered "partially knew each other", as they either knew only one of the two other group members or knew their partner(s) to a limited extent. The majority of participants were located in Den Haag New Babylon, which was to be expected since we only distributed physical posters there. Only two employees reported working in Leiden, while five others were mainly located in other offices. This last group of participants had to select one of the two locations, and they all chose Den Haag. The office locations we mentioned were the only ones in which participants could hold their Metaverse meetings. Then, most participants admitted that they had little experience with VR technologies. By assigning numerical values to the verbal descriptors of the questions, where 1 = none and 5 = very much, we retrieved the following data about the participants' VR experience: n = 32, mean = 2.34 (between little and moderate), median = 2, SD = 1.12. Regarding the participants' gender identity, nine participants identified as women, while the other 23 identified as men. Lastly, fourteen participants responded that they were "confident" when using digital devices, and the remaining eighteen answered "extremely confident", on a five-item Likert scale.

Figure 4.2 shows a graphical overview of the most relevant participants' demographic, namely gender, age, VR experience, and relationship.

After discussing the combinations of variables and meeting tasks of each meeting and

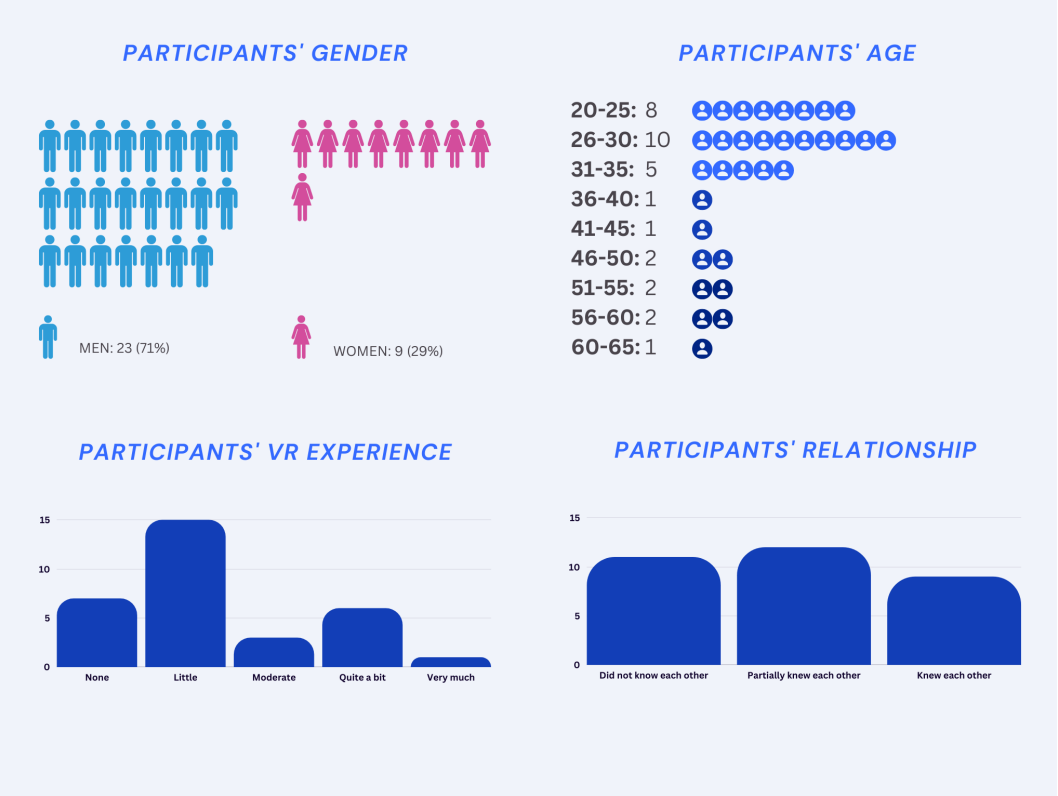


Figure 4.2: The participants' demographic.

how to evaluate them, we presented in this chapter the employees who participated in said meetings. Overall, the 32 participants were predominantly men, under 35 years of age, with little to no VR experience, located in Den Haag, and confident in their digital skills; this imbalance in the demographic can be attributed to the convenience sampling. The study of Tullis and Stetson provided us with empirical evidence that a sample size of 32 would yield accurate results for our questionnaires. The following chapter will discuss the immersive and non-immersive systems used in the experiment.

4.3 System Design and Locations

The current section describes the technology and software used in meetings with the TNO ICS, and the office locations where participants held their meetings. Given that the Social XR team had seven HMDs and computers available for use, we prepared seven rooms for the meetings in the TNO ICS technology to ensure that, even in the possibility that two distinct groups planned a VR meeting at the same time, their request could be accommodated. The systems were deployed in the TNO locations of Den Haag New Babylon, and Leiden Sylviusweg. More specifically, the first has six rooms reserved for the experiments, while the latter location only has one. Therefore, we ensured that the participants who selected Leiden as their preferred location would not be added to the same group. Although we had initially planned to deploy the system in additional locations, to allow more employees to participate in the experiment without having to travel long distances, only the two locations mentioned had rooms available for us to use. As a result, the majority of meetings were held between participants located in the same building. Although the intended use of the platform is to host meetings between employees located in different locations, our

solution not only did not compromise the outcome of the experiment, but made it also more convenient for us to set up the rooms and help the participants. Each of the seven rooms dedicated to the experiment had a laptop or computer, an Oculus Meta Quest 2 with its controllers and charging cable, and a depth camera. Figure 4.3 represents the standard system setup.

Figure 4.3: The researcher wearing the HMD, holding the controllers, and sitting on a chair facing the ZED Camera.

Each computer had installed the software "TNOCaptureStarter" to access the ZED depth camera, and participants could access the Metaverse through its website tno90.westeurope.cloudapp.azu from both the computer and the headset. The camera captures the user and sends the image to the computer, which streams it on the website. The Oculus of the second participant receives the image of the user from the website, and displays it in the virtual room, thus giving the feeling of being with another person in the same space. In this way, each participant sends their image to their partners and receives back their Pointcloud representation. Figure 4.4 depicts a schematic representation of three colleagues, connected from separate rooms, experiencing the feeling of sitting together at a table in the VR environment.

Figure 4.5 shows a user, wearing the HMD, in the dedicated Metaverse room. The laptop at the bottom of the image shows how others would perceive him in the virtual environment. Figure 4.6 shows the other users in the Metaverse room, as seen by the first participant.

Describing in depth the hardware and software components of the TNO ICS is not in the scope of this thesis. For further research, the paper by Gunkel et al. [18] explains in depth the web-based VR framework used.

Regarding the MS Teams meetings, employees could meet either from home or from their offices, as this meeting did not require any specific hardware other than the personal laptop of the participants. Given that MS Teams is the standard meeting platform at TNO, no participant had to download additional software for the experiment. To replicate as closely as possible the meeting conditions the Metaverse provides, users were instructed not to share any material or browse other tabs during the meeting. Figure 4.7 represents

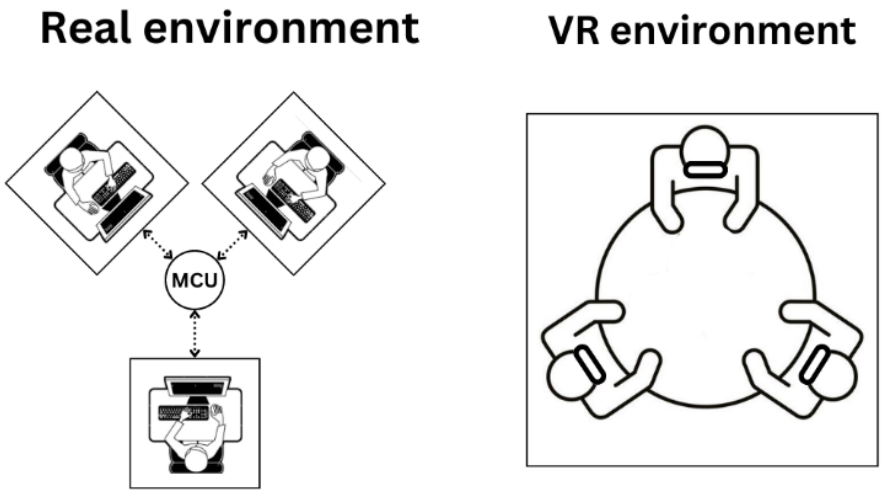


Figure 4.4: The schematic representation of the real world and virtual environment disposition.

Figure 4.5: A participant and his Pointcloud representation.

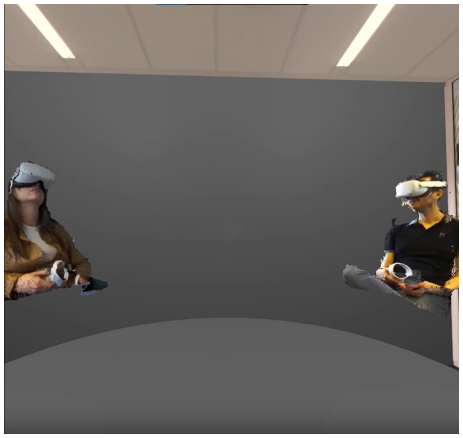


Figure 4.6: What the participant sees in the Metaverse room.

what a standard meeting in MS Teams, with three employees, would look like[4].

Lastly, the human resources needed for the experiment, other than the participants, consisted of only the author of this document, as under the guidance of my supervisors, managed to prepare and carry out the entire experiment independently.

After the study design and population, we presented in this chapter the design of the TNO ICS, comprising a computer, a depth camera, and an HMD. Moreover, the office

[4]Image taken from https://www.computerworld.com/article/3542389/microsoft-teams-video-meetings-best-practices.html

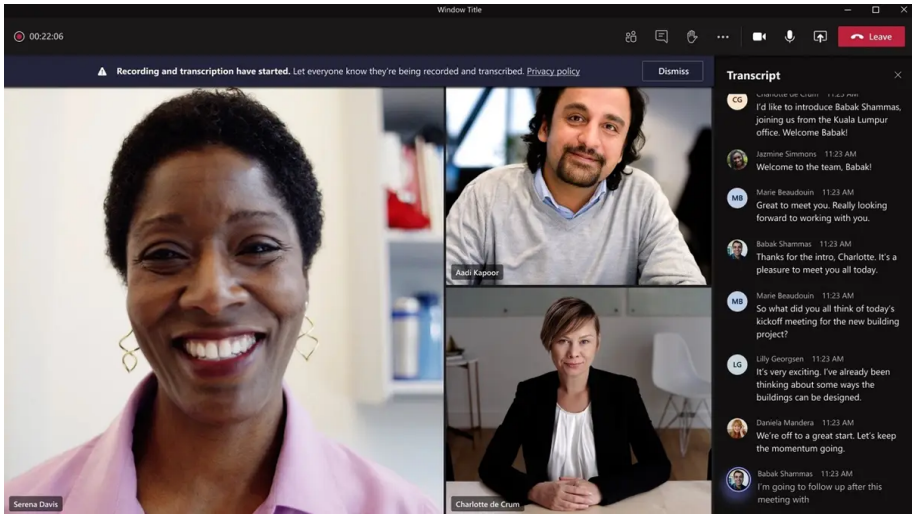


Figure 4.7: A meeting in MS Teams

locations for the Metaverse meetings were in Den Haag and Leiden. Next, we will outline the several stages of the experiment.

4.4 Procedure of research

This section includes the preparatory phase of the experiment, the actual experimental procedures, the distribution of surveys, and the subsequent interview sessions.

4.4.1 Preparation Details

The participant recruitment started on March 13th by word of mouth, a post on the TNO's platform for employee communication, and the posters left in the company break rooms. On March 31st, after having received enough participation requests, we organized an onboarding meeting with the participants to provide them with further information about the experiment and answer their questions.

Each participant was then paired with either one or two colleagues. Each group had to participate in four meetings, two in the Metaverse and two in MS Teams.

4.4.2 Pilot Testing

On May 1st, together with two other colleagues from TNO, who were not part of the participants in the user testing, we conducted the pilot testing to ensure that the experiment was ready to begin. The pilot testing consisted of running the complete process, from booking a room for the Metaverse meeting to the interview session. With no more instruction than those given to the actual participants, and under our strict supervision, they went through the steps that their colleagues would have done later. The testing produced mostly positive results, as they were able to autonomously conduct the meeting in the Metaverse and fill in the questionnaire, which were the most critical parts of the process. The main insights gathered from the pilot were related to the instructions, as we realized that they needed to be more precise and detailed, especially since the technology was new for most of the users. Additionally, it showed how a single T1-T4 brainstorm task took too little time to complete in the meeting, while two broadcast tasks were too long to do

in one sitting. Therefore, it leads us to assign two brainstorm tasks or one broadcast task for each meeting, regardless of the communication media.

4.4.3 User Testing

The first meeting took place on May 8th, while the last interview session was on June 21st, meaning that the experiment took 44 days to complete, against the 49 (7 weeks) initially hypothesized. Participants were instructed verbally, at the beginning of each meeting in the TNO ICS, on how to set up the equipment and hold the meeting in the Metaverse; additionally, they would receive an explanation of the tasks they had to complete. After that, including for MS Teams meetings, participants were not recorded, nor controlled by any external observer, to avoid biasing or influencing them; we valued participants having a more realistic meeting experience than having the experiment in a strictly controlled setting.

At every meeting we always reminded participants to spend at least 20 minutes on their task(s), as we thought it was the minimal amount of time for them to meaningfully experience the VR system, and no more than 60 minutes especially for the meetings in the ICS, as the guidelines on VR use suggest. To our knowledge, there is still no study on the minimum amount of minutes that a person needs to be in the Metaverse to get acquainted with the virtual environment. Although we could not verify this ourselves, most participants mentioned during the interview sessions that the tasks were interesting and stimulating enough to make them work for more than 20 minutes.

4.4.4 Issues During User Testing

The pilot test run was effective in uncovering issues with the instructions, but several other issues only emerged later, during the user test.

First, after the first few meetings, users reported that the function to share slides in the VR room would not work. We had initially planned the broadcast tasks to be a moment of discussion and confronting each other to order the provided items in a list based on relevance, as explained in section 4.1.2. We expected participants to read the meeting instructions and the items to order from the slides, displayed on virtual screens on the VR room walls. However, the failure of the slides feature forced us to change the task. First, we had to explain at the beginning of every meeting the meeting objective; then, participants had to come up on their own with the items that they thought were the most important in such situations and rank them together.

Secondly, several issues related to the technology tainted the participants' experience during the experiments. One of the most common problems reported was the Pointcloud being either severely mispositioned or not at all shown in the Metaverse meeting room; this problem was often linked to computers that suddenly froze or became unresponsive; in total, 15 users reported such an issue. Then, 7 employees mentioned their HMDs running out of battery upon entering the Metaverse; this issue happened three times in meetings with three participants, meaning that the other two participants could still continue with their meeting; for statistical analysis purposes, we counted their questionnaires as if they were related to a meeting with two participants. Since they could not participate in the meeting, the questionnaire they compiled was therefore discarded, to ensure the data quality of the study, its validity, and integrity. Similarly, their interview was not coded. Lastly, two users could not activate the immersive experience, thus seeing the room and the colleagues in a 2D window in front of them. Unlike the other cases, where the issues prevented participants from participating in the meeting, these two employees still completed the task

given. However, we decided to drop their questionnaire data as well, as we doubt that they could have given a fair and valuable rating to the system's usability, effectiveness, and quality of social communication perceived.

Missing and Removed Data

Interviews were then analyzed first to understand whether participants experienced such technical issues during their meeting that influenced the experience more than expected. The statistical analysis conducted on the questionnaire scores, reported in section 6.1.1, revealed a statistically significant difference between the meetings with and without technical issues, de facto skewing unfairly the score to lower values. As a consequence, to present the data in an unbiased way, and to ensure the integrity and reliability of the data analysis, we dropped the scores collected from 25 questionnaires relating to experience with major technical problems. The exclusion criteria include the four issues mentioned previously. More participants also mentioned other technical issues, such as hearing an echo or having to restart the software. However, we believe that their overall Metaverse experience proceeded as intended, thus their data was not excluded.

4.4.5 Questionnaires and Interviews

After each meeting, participants had to fill in the questionnaire provided online, which included questions on their demographic, the meeting information, and the three questionnaires mentioned on usability, effectiveness, and social communication. The complete questionnaire consisted of 39 questions and can be accessed in Appendix B. The questionnaires were designed for simplicity of use on Google Forms.

Then, before the following meeting experience, participants had to individually participate in the interview session, held on MS Teams, which usually lasted around 30 minutes. The interview questions can be found in Appendix C. The topics of the initial 15 questions were the overall experience, the system usability, the perceived effectiveness, and the social communication factor. We expected that users would experience the meetings in different ways, and thanks to their diverse backgrounds, they would have been able to analyze the systems in widely different ways. In addition, the qualitative data was meant to be part of an exploratory study, without having to answer any hypothesis. Therefore, we designed the interview to be semi-structured, and the questions were meant to be open-ended, to allow the interviewees to express themselves freely, and open up to other topics of discussion. Following each question, we consistently inquired further by asking "why," which allowed us to delve deeper into the participants' opinions and gain richer insights. As expected, several participants mentioned topics, ideas, and factors that we initially did not contemplate, allowing us to take inspiration from their comments and add new items to the question pool. At the end of the interview series, the interview document counted 38 questions; having a broad range of items allowed us to ask only those that were relevant to each user's experience they had.

Overall, TNO employees only filled in 118 questionnaires out of the 128 expected, as two participants were sick on the day of the meeting and did not participate, four others forgot to fill in the questionnaire before the following experiment, and four people decided not to complete the questionnaire because of the major issues experienced during the meeting.

In this section, we first summarized the eight meeting conditions and order of the six tasks, divided into four groups of participants. Secondly, we discussed the pilot testing, which helped us improve the meeting instructions. Then, we described the user testing in-

formation. Subsequently, we reported the issues faced during the experiment, i.e. the slides malfunction. Lastly, we talked about the filling in of questionnaires and the methodology used to conduct the interviews.

This section concludes the chapter on the experiment. It explained how we applied the theoretical knowledge from the literature review and the research questions to gather the data regarding meetings in the Metaverse. Next, we will describe the analysis we conducted on the data collected during the experiment.

Chapter 5

Data Analysis

This chapter describes the comprehensive data analysis conducted to address the research questions and objectives of this study, namely evaluating the three main QoE components of the TNO ICS to understand its feasibility to be deployed for online meetings. The two sections of this chapter cover the methodology used to analyze the questionnaires and interviews. The results will be presented in the following chapter.

5.1 Questionnaires Analysis

The Google Form website, which hosted the questionnaires, converted the data collected to a comprehensive Excel file. Afterward, we conducted a data cleaning process and formatted the file to enhance clarity and practicality.

First, based on questions 4 to 8 of the questionnaire, we analyzed the participants' demographic, which is reported in section 4.2.2. Then, the procedure to calculate the SUS score, described in [12], proceeds as follows: for each of the SUS questionnaires collected, we subtracted 1 from the numerical value (1 to 5) assigned to items 1, 3, 5, 7, and 9; for items 2, 4, 6, 8, and 10 we subtracted their score from 5. Lastly, we multiplied by 2.5 the sum of the scores to obtain the overall SUS value on a range from 0 to 100. Successively, we assigned numerical values to the verbal descriptors of each question from the "Perceived Usefulness" questionnaire to run the statistical analysis. More specifically, we converted to 1 the lowest item of the scale, "Extremely unlikely", to 7 the highest item, "Extremely Likely; the intermediate labels were correspondingly mapped to their respective numerical values. We calculated the final score of each questionnaire, ranging from 1 to 7, by averaging the value of the six questions. We repeated the calculation for the H-MSC-Q as well, where the lowest value "strongly disagree" was converted to 1, the highest value "strongly agree" to 7, and the intermediate to appropriate numerical values. Then, calculating the average of the first five questions produced the final score related to spatial presence, and we calculated the score of the social presence by averaging the last ten questions. Both scores range from 1 to 7. At the end of this process, each of the 118 questionnaires contained six important values for the following data analysis: the participant code, the communication media, the meeting type, the number of participants, and the overall score for the three questionnaires.

5.1.1 R Functions and Code Snippets

This section explains the functions used for the data analysis. We used the software "R" to conduct the statistical analysis and create the graphs. The packages used in the statistical

analysis are reported in the footnotes throughout this section. The complete code used for the statistical analysis is in the additional documents submitted with this Thesis.

First, for each of the three questionnaires we created a data frame "reg_data" containing four variables: first, the overall scores related to the correspondent questionnaire, calculated following the procedure explained in the previous section; secondly, a binary variable "communication media", indicating either a meeting in the TNO ICS or in MS Teams; then, the variable "subject", associating each score to the participant ID; lastly, excluding for the SUS questionnaire analysis, the fourth variables indicates either the meeting type (brainstorm or broadcast) for the analysis on effectiveness, or the number of participants (two or three) for the analysis on social communication.

First, we conducted the linear mixed-effects regression analysis using the "lmer" function from the package "lme4"[1]; the package "lmerTest"[2] provides the p-values in type I, II or III ANOVA and summary tables. The function we employed is written as

H_MSC_Q_Social_Presence_Scores ~ Number_of_participants + Communication_Media + (1|Subject), where the first argument indicates the questionnaire scores, the second and third arguments indicate the two (or eventually more) variables used, and "Subject" indicates the list containing the participants' ID codes. We also used the models with only one variable at a time, and models using the operators "*" or ":" instead of or in combination with "+", to study the interactions instead of the combinations between variables. The results were displayed using the function "summary(model)". We chose the Linear Mixed-Effect Regression Models (LMMs) to test for the statistical significance of our quantitative data findings. LMMs are a statistical model, extending linear regression models, for data that are collected and summarized in groups, non-independent, multilevel or hierarchical, longitudinal, or correlated[3][4]. These models are "mixed" because they contain both fixed effects and random effects. Fixed-effects terms are usually the conventional linear regression part; the independent variable has a fixed relationship with the dependent variable across all observations. On the other hand, random effects are associated with individual experimental units drawn at random from a population. In our experiment, the communication media, meeting type, and number of participants represent the fixed effects, as they directly influence the dependent variables. Instead, the participants constitute the random effect; their unique psychological characteristics may cause them to react differently to various stimuli, leading to factors affecting the dependent variable that change randomly between participants.

We used the function "aictab" from the package "AICcmodavg"[5] to rank in a table the models, based on the Akaike Information Criterion (AIC) score they received. The AIC is a method for evaluating how well a model fits the data; the function employed contains the list of LMM models used and their associated names.

We performed an analysis of variance (ANOVA) on the lmer results to assess the significance of the effects of the variables. We used the function anova(comm.eff) from the native functions of R, where the argument in parenthesis is the name associated with the model of interest.

We used the function "predictmeans" from the package "predictmeans"[6] to calculate the statistical values of our variables; the first argument of the function is the model and

[1] https://cran.r-project.org/web/packages/lme4/index.html

[2] https://cran.r-project.org/web/packages/lmerTest/index.html

[3] URL: https://it.mathworks.com/help/stats/linear-mixed-effects-models. html.

[4] Ajitesh Kumar. Fixed vs random vs mixed effects models - examples, Mar 2023. URL: https://vitalflux.com/ fixed-vs-random-vs-mixed-effects-models-examples/.

[5] https://cran.r-project.org/web/packages/AICcmodavg/index.html

[6] https://cran.r-project.org/web/packages/predictmeans/index.html

the second argument is the variable to test.

We used the function "TukeyHSD", native to R, to run the multiple comparison tests; the function used is TukeyHSD(table), where the argument contains the results of the ANOVA.

We used the function "pairwise.t.test", native to R, to run the Bonferroni correction for the Multiple Test Correction; the first argument of the function contains the scores from the questionnaire, the second argument contains an independent variable, and the last argument specifies to use the Bonferroni method.

We used the function "cohensD" from the package "lsr"[7] to calculate the effect size using Cohen's d; the function used is cohensD(group1, group2), where the two arguments indicate the scores associated to the two levels of an independent variable.

5.2 Interviews Analysis

The interview analysis started with recording and transcribing each interview session with the MS Teams software. Then, the ATLAS.ti program was used to code them, following the Thematic Analysis process. According to the paper's suggestions, every meaningful comment (e.g. actions, causes, outcomes, etc.) was given a relevant code to categorize and classify the different parts of the data. Our goal was to find meaningful patterns, themes, and concepts in the participants' comments. Lastly, we merged codes that were duplicates or nearly identical.

Instead of the expected 128 interviews, only 118 were carried out, for four main reasons; first, two participants were suddenly ill on the day of the experiment, and could not participate; secondly, three participants preferred not to participate in an interview session because the major technical issues prevented them to experience the Metaverse meaningfully, and anticipated that they had no feedback on the experience; then, one participant did not have the time to participate in three interviews; lastly, a couple of participants could not complete their two meetings in MS Teams before the deadline.

Coding the first 60 interviews, related to the meeting with the TNO ICS, led to data saturation, as barely any new code had been found in the last interviews coded. We then decided not to code the interviews related to meetings in MS Teams, but to go through them to find novel insights. Although interesting, the few insights emerged were mostly not related to the Metaverse, but to MS Teams; for example, participant 1A1Y mentioned that "when [I am] on teams I do not feel pressured into talking and I can just disappear in the background". Together with similar insights gathered throughout the whole interview analysis process, which are reported later, we presented them to the team to help them understand the users' opinions about the Metaverse experience.

The chapter on data analysis ends here. We explained how, out of the 128 interviews, only 118 were held, and 60 were coded following the guidelines of the Thematic Analysis process. As for the questionnaires, we mainly presented the steps followed to calculate the overall score of the three questionnaires; then, we reported the R functions used to compute the AIC, linear mixed effect regression, ANOVA, predicted means, Tukey HSD, Bonferroni correction, and Cohen's d. After having discussed the Research Questions we followed, the experiment designed to apply them, and the analysis process used to gather information from the data collected, we will present next the outcomes of our research.

[7] https://cran.r-project.org/web/packages/lsr/index.html

Chapter 6

Results

This chapter contains the quantitative and qualitative results of our experiment, divided into sections related to each one of the five RQs. We list in the footnotes the sources used throughout the chapter that, despite not being published in a journal article, were scientific and relevant enough to help us delve deeper into the various statistical analysis topics relevant to this chapter.

6.1 Questionnaires Results

In this section, we present the results of the Akaike Information Criterion, the linear mixed effect regression model, the ANOVA applied to the model, the predicted means and confidence intervals, the Tukey HSD post hoc test on the ANOVA, the Bonferroni correction, and the Cohen's d. These statistical analyses test the data collected from the three questionnaires. Most results are then summarized in plots and tables. We will report the significance codes of the p-values as follows: "***" for p-values between 0 and 0.001; "**" for values between 0.001 and 0.01; "*" between 0.01 and 0.05; "." to indicate p-values from 0.05 and 0.1; lastly, no code is applied to p-values above 0.1. Following the standard practice, we also set an alpha level of .05 for all statistical tests to determine which predictors are significant[1]. As mentioned in section 4.4.4, for our statistical analysis we did not account for the data related to meetings that had technical issues.

6.1.1 Excluding the Data

Considering the unexpected number of meetings tainted by technical problems, we decided to exclude the data related to those experiences, to have a fair comparison between communication media working as intended. Of course, there is no need to divide the MS Teams level between the experiences with and without issues, as all meetings in this communication media were completed smoothly. From the 60 meeting experiences in the TNO ICS, 24 presented severe technical issues that prevented the participants from fully acknowledging the system's usability, effectiveness, and social communication as intended. From those 24 experiences, only 22 questionnaires were completed, as two participants did not bother to complete the questionnaire. We then conducted statistical analysis to understand whether there was a significant difference in the Metaverse experiences completed in normal conditions against those with major issues. Regarding the usability questionnaire, the ANOVA indicates a highly significant distinction ($p < .001$) among the Communication Media conditions of MS Teams, ICS with errors, and ICS with no errors. More specifically,

[1] https://www.statology.org/significance-codes-in-r/

the Tukey HSD post hoc test reveals a difference of -16.27 points between the ICS with and without errors conditions, with a lower bound of -24.72, an upper bound of -7.83, and a p-value $<.001$ affirming its statistical significance. Then, we repeated the analysis of the perceived effectiveness questionnaire data. As before, the role of Communication Media remains significant in influencing the scores of the questionnaire ($p < .001$). More specifically, the Tukey HSD post hoc test reveals a difference of -1.31 points between the ICS with and without errors conditions, with a lower bound of -1.99, an upper bound of -0.63, and p-value $<.001$ affirming its statistical significance. The same applies to spatial and social presence, as the differences in communication media present a statistically significant effect. Furthermore, in the analysis of spatial presence, the Tukey HSD post hoc test reveals a difference of -1.05 points between the ICS with and without errors conditions, with a lower bound of -1.67, an upper bound of -0.43, and p-value $<.001$ affirming its statistical significance. Regarding social presence, the Tukey HSD post hoc test reveals a difference of -0.73 points between the ICS with and without error conditions, with a lower bound of -1.22, an upper bound of -0.25, and p-value $<.001$ affirming its statistical significance. In conclusion, we demonstrated that the meeting experiences concerning the TNO ICS with technical issues are significantly different from meetings without issues for all three QoE factors. We decided then to discard the data from the 22 questionnaires related to these events, as they do not provide an accurate description of the TNO value. After excluding the skewed data, we can now proceed further with the data analysis process by analyzing the answers to three questionnaires.

6.1.2 RQ1: Usability

We present in this section the results of the statistical analysis conducted to answer the research question "How does usability change in different communication media?". The categorical, independent variable of this part of the study is "communication media", containing the two levels "ICS" and "MS Teams", indicating the meeting platform used in the meeting; the continuous dependent variable is usability, and its values represent the scores obtained from the SUS questionnaire, on a scale from 0 to 100.

In total, 32 participants filled out 93 "SUS" questionnaires; 35 questionnaires are related to meetings in the TNO ICS, and 58 are associated with meetings in MS Teams. As mentioned before, we collected 60 questionnaires in total related to meetings in the ICS, but 25 of them were discarded. Having to exclude the questionnaires related to meeting experiences with technical issues is the reason why, for RQs one to three, their related questionnaires have more instances of MS Teams meetings than TNO ICS meetings.

We did not need to use the Akaike Information Criterion, as there was only one variable to analyze, and one model to choose.

We then fit our linear mixed-effect model with the independent and dependent variables, namely SUS score and communication media, and subjects as random effects. Regarding the fixed effects, the estimated intercept returned a value of 52.75 for the dependent variable "usability". The t-test performed to compare the means of communication media on SUS scores showed that there was a significant difference ($t(df) = 16.37$, $p < .001$). The mean of MS Teams meetings was 36.64 points (standard error: 2.23) higher than the mean of meetings in the TNO ICS. The random effect is given first by the subjects in our study, which are responsible for a variance of 50.87 points in the data (standard deviation: 7.13). Then, the residuals are responsible for a variance of 98.73 points in the data (standard deviation: 9.94).

Applying the ANOVA method for linear mixed effect regression (lmer) model produces a type III ANOVA table for fixed-effect, which confirms that there was a significant difference

($F(1, 74.50) = 267.84$, $p < .001$) between communication media.

We used the function "predictmeans" to calculate the statistical values of the levels TNO ICS (estimated mean = 52.75, standard error = 2.21, lower confidence limit = 48.3, upper confidence limit = 57.2) and MS Teams (estimated mean = 89.39, standard error = 1.85, lower confidence limit = 85.66, upper confidence limit = 93.12), from the variable communication media. Figure 6.1 shows the distribution of the usability scores of the two communication media and their predicted mean (shown by a dotted line). In the plot, it appears that there is a noticeable difference in the means between the communication media.

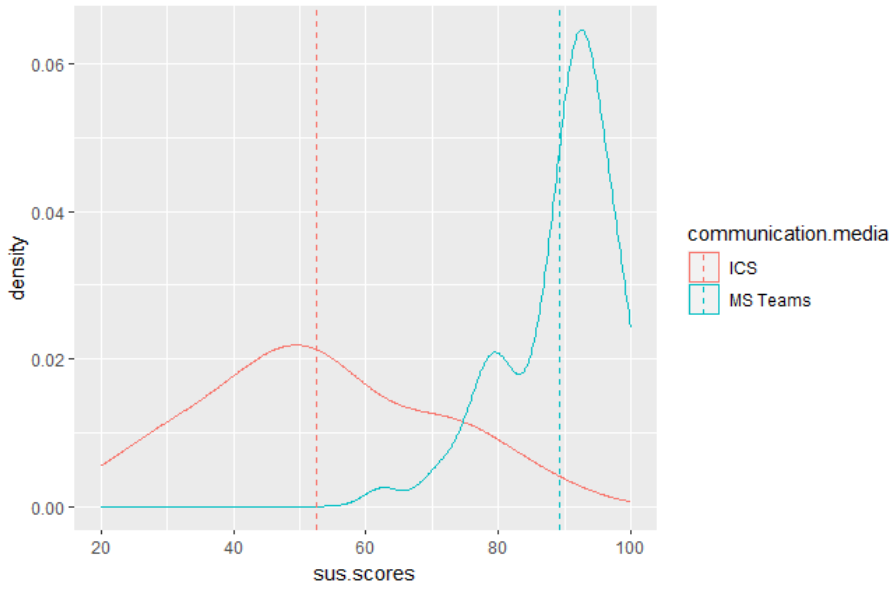


Figure 6.1: The distribution of SUS scores and their predicted mean.

According to the SUS score interpretation, described in the widely used paper of Bangor et al. [7], the TNO ICS obtained an "F", the lowest of the grades, with an adjective rating "ok", which according to the author signifies a marginally low acceptable technology. For comparison, the grade assigned to MS Teams is "B", which translates to "excellent" in the adjective scale.

We chose the Tukey HSD for our multiple comparison test. The post hoc comparisons using the Tukey HSD test indicated that the mean score for the ICS meeting condition was significantly different than the MS Teams meeting condition (mean difference = 36.74, lower interval = 31.48, upper interval = 42, $p < .001$).

We chose the Bonferroni correction for the Multiple Test Correction. This test attempts to prevent data from incorrectly appearing to be statistically significant by making an adjustment during comparison testing. The adjusted p-value for the mean difference in SUS scores between TNO ICS and MS Teams is still $< .001$. Based on the output, the difference between communication media is confirmed to be statistically significant.

To show whether the effect of our variables on the meeting effectiveness is large enough to be meaningful in the real world, we calculated the effect size using Cohen's d. The effect size of different communication media, as measured by Cohen's d, was d = 2.97, indicating a big effect. We follow the commonly used rule of thumb to interpret Cohen's d: a value of 0.2 represents a small effect size; a value of 0.5 represents a medium effect size; a value of 0.8 represents a large effect size.

	Broadcast	Brainstorm	TOTAL
ICS	18 (15)	17 (10)	35 (25)
MS Teams	29	29	58
TOTAL	47 (15)	46 (10)	93 (25)

Table 6.1: The number of Perceived Effectiveness questionnaires representing each level.

6.1.3 RQ2: Effectiveness

We present in this section the results of the statistical analysis conducted to answer the research questions "How does the perceived effectiveness of meeting types change in different communication media?" and "How does the perceived effectiveness of communication media change in different meeting types?". The two categorical, independent variables of this part of the study are "communication media" and "meeting type". The first contains the two levels "ICS" and "MS Teams", indicating the meeting platform used in the meeting; the second contains the two levels "brainstorm" and "broadcast", indicating the tasks to complete in the meeting. The continuous dependent variable is effectiveness, and its values represent the scores obtained from the perceived usefulness questionnaire.

Given that the questionnaire allowed users to give their answers on a 7-point Likert scale, the results presented in this section will be in a range from 1 to 7. In total, 32 participants filled out 93 "perceived usefulness" questionnaires; 46 of them rated broadcast meetings, while the remaining 47 evaluated brainstorm meetings; lastly, 35 questionnaires are related to meetings in the TNO ICS, and 58 are associated with meetings in MS Teams. Table 6.1 summarizes the questionnaires divided per meeting conditions. The numbers in parenthesis indicate how many questionnaires, per meeting type, were discarded.

First, we used the Akaike Information Criterion to compare different possible models and determine which one is the best fit for the data. First, we test how each variable performs separately. Then, we want to know if the combination of communication media and meeting type is better at describing variation in the effectiveness scores. Finally, we check whether the interaction of the two variables can explain the effectiveness better than any of the previous models. The results show that the best-fit model, carrying 71% of the cumulative model weight, included only the parameter communication media. The second best-fit model includes the combination (but no interaction) between the two variables; this model differs from the AIC score of the best model by 2.93 points. The best-fit model is the one that only includes the communication media parameter, and will be used in the lmer and ANOVA. However, for the goal of the analysis, we will also include the meeting type parameter in the model for the statistical analysis related to such variables. To cover every aspect of the analysis, we will also present the interaction effects of the two variables, given by the third model using the operator "*". Table 6.2 shows, for each of the four models, the number of parameters in the model (K), the information score of the model (AICc), the difference in AIC score between the best model and the model being compared (Delta_AICc), the proportion of the total amount of predictive power provided by the full set of models contained in the model being assessed (AICcWt), the sum of the AICc weights (Cum.Wt), and the value describing how likely the model is, given the data (LL)[2]. The model we choose is the one that has the lowest score in the AICc column.

[2]https://www.scribbr.com/statistics/akaike-information-criterion/

Model	K	AICc	Delta_AICc	AICcWt	Cum.Wt	Res.LL
Communication Media	4	274.84	0.00	0.71	0.71	-133.19
Communication Media + Meeting Type	5	277.77	2.93	0.16	0.87	-133.54
Communication Media * Meeting Type	6	278.26	3.42	0.13	1.00	-132.64
Meeting Type	4	365.54	90.70	0.00	1.00	-178.54

Table 6.2: The Model selection based on AICc.

We then fit our linear mixed-effect model with the independent variable communication media first, to test its significance using the best-fit model, and then we fit the model with the combinations of communication media and meeting type; for both, we used the participants (also called "subjects") as random effects. Linear mixed models are an extension of simple linear models to allow both fixed and random effects; fixed effects have a constant effect on the dependent variable, while random effects have a varying effect on the dependent variable across groups or individuals[3][4][5]. Regarding the fixed effects, the estimated intercept returned a value of 3.46 for the dependent variable "effectiveness". The t-test performed to compare the means of meeting type on effectiveness showed that there was no significant difference ($t(df) = -0.94$, $p > .1$). The mean of broadcast meetings was 0.17 points (standard error: 0.18) lower than the mean of brainstorm meetings. Successively, the t-test performed to compare the means of communication media on effectiveness showed that there was a significant difference ($t(df) = 13.41$, $p < .001$). The mean of meetings in MS Teams was 2.6 points (standard error: 0.19) higher than the mean of meetings in ICS. Lastly, to cover every aspect of the analysis, we fit our linear mixed-effect model with combinations between the independent variables and the interactions of their factors; the t-test performed to compare the means of the interaction effects between meeting type and communication media on effectiveness showed that there was not a significant difference ($t(df) = 1.4$, $p > 0.1$). The mean of Broadcast meetings in MS Teams was 0.52 points (standard error: 0.37) higher than the mean of brainstorm meetings in ICS. Table 6.4 displays the data of the fixed effects in our model; the "Estimate" column shows the estimated values of the coefficients for each fixed effect; the "Std. Error" column represents the standard errors associated with each coefficient estimate; the "df" column shows the degrees of freedom associated with each coefficient; the "t value" column shows the t-statistic associated with each coefficient estimate; lastly, the "Pr(>|t|)" column shows the p-value associated with each coefficient estimate[6]. The random effect is given first by the subjects in our study, which are responsible for a variance of 0.4 points in the data (standard deviation: 0.63). Then, the residuals are responsible for a variance of 0.74 points in the data (standard deviation: 0.86). Table 6.3 displays the variance and standard deviation of the two random groups; the "Subject (Intercept)" random effect captures the variability in the intercepts of different subjects; the "Residual" random effect accounts

[3]https://stats.stackexchange.com/questions/29329/what-is-the-meaning-of-operators-in-regression-or-anova-formulas-in-r

[4]https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-linear-mixed-models/

[5]https://vitalflux.com/fixed-vs-random-vs-mixed-effects-models-examples/

[6]https://www.linkedin.com/advice/1/how-do-you-interpret-report-results-t-test

Groups	Name	Variance	Std.Dev.
Subject	(Intercept)	0.4013	0.6335
Residual		0.7390	0.8596

Table 6.3: The random effects in the linear mixed effect model.

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	3.4594	0.2118	76.1277	16.337	< .001 ***
Meeting Type (Broadcast)	-0.1704	0.1804	63.0662	-0.944	0.349
Communication Media (MS Teams)	2.6015	0.1940	71.0509	13.412	< .001 ***
Meeting Type (Broadcast) : Communication Media (MS Teams)	0.5190	0.3732	63.8609	1.391	0.1692

Table 6.4: The fixed effects in the linear mixed effect model.

for the remaining variability within groups that cannot be explained by the fixed effects or the "Subject (Intercept)" random effect. In this model, the random effects do not seem to add too much of a variance.

Applying the ANOVA method for linear mixed effect regression model fits produces a type III ANOVA table for fixed-effect, which confirms that there was no significant difference ($F(1, 63.07) = 0.89$, $p > .1$) between meeting type; similarly, it confirms that there was a significant difference ($F(1, 71.05) = 179.89$, $p < .001$) between communication media; lastly, the ANOVA on the linear mixed-effect model with combinations between the independent variables and the interactions of their factors confirms that there was not a significant difference ($F(1, 63.86) = 1.93$, $p > .1$) between the interaction effects of communication media and meeting type. As before, the results are given by the linear mixed model using the communication media variable only, then the combination of the two variables for the insights on the meeting types, while the last test is taken from the model using the interaction between the two variables. Table 6.5 represents the complete results of the ANOVA. The Df column displays the degrees of freedom for the independent variable (the number of levels in the variable minus 1), and the degrees of freedom for the residuals; the Sum Sq column displays the sum of squares (i.e. the total variation between the group means and the overall mean); the Mean Sq column is the mean of the sum of squares, calculated by dividing the sum of squares by the degrees of freedom for each parameter; the F value column is the test statistic from the F test; the Pr(>F) column is the p-value of the F statistic[7].

We used the function "predictmeans" to calculate the statistical values of the levels TNO ICS (estimated mean = 3.37, standard error = 0.19, lower confidence limit = 2.98, upper confidence limit = 3.76) and MS Teams (estimated mean = 5.98, standard error = 0.16, lower confidence limit = 5.65, upper confidence limit = 6.3), from the variable

[7]https://cran.r-project.org/web/packages/lmerTest/lmerTest.pdf

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
Meeting Type	0.659	0.659	1	63.066	0.8919	0.3486
Communication Media	132.930	132.930	1	71.051	179.8891	$< .001$ ***
Meeting Type : Communication Media	1.411	1.411	1	63.861	19.336	0.1692

Table 6.5: The type III Analysis of Variance Table.

Meeting Type	Mean	SE	Df	LL(95%)	UL(95%)
Brainstorm	4.7602	0.17328	52.16861	4.4125	5.1079
Broadcast	4.5898	0.17590	52.16861	4.2369	4.9428

Table 6.6: The expected mean, standard error, and confidence limits of the meeting types.

communication media. The expected mean is what would result from the long term of doing an experiment over and over[8]. Secondly, we analyzed the two levels of the second variable "meeting type", namely brainstorm meetings (estimated mean = 4.76, standard error = 0.17, lower confidence limit = 4.41, upper confidence limit = 5.11) and broadcast meetings (estimated mean = 4.59, standard error = 0.18, lower confidence limit = 4.24, upper confidence limit = 4.94). Table 6.6 and 6.7 show the complete data returned by the function "predictmeans". The first column contains the predicted means; the standard error on the predicted mean is in the second column; the degrees of freedom are in the third column; the lower and upper limits of the confidence interval are in the fourth and fifth columns. Figures 6.2 and 6.3 show the distribution of the effectiveness scores of the two communication media and their predicted mean (shown by a dotted line). In the plots, it appears that there is a noticeable difference in the means between the communication media, but not between the meeting types.

We chose the Tukey HSD for our multiple comparison test. This procedure is used to find means that are significantly different from each other[9]. The post hoc comparisons using the Tukey HSD test indicated that the mean score for the ICS meeting condition was significantly different than the MS Teams meeting condition ($p < .001$). Considering

[8]https://openstax.org/books/statistics/pages/4-2-mean-or-expected-value-and-standard-deviation
[9]https://en.wikipedia.org/wiki/Tukey%27s_range_test

Communication Media	Mean	SE	Df	LL(95%)	UL(95%)
ICS	3.3742	0.19297	41.91	2.9848	3.7637
MS Teams	5.9758	0.16208	41.91	5.6487	63.029

Table 6.7: The expected mean, standard error, and confidence limits of the communication media.

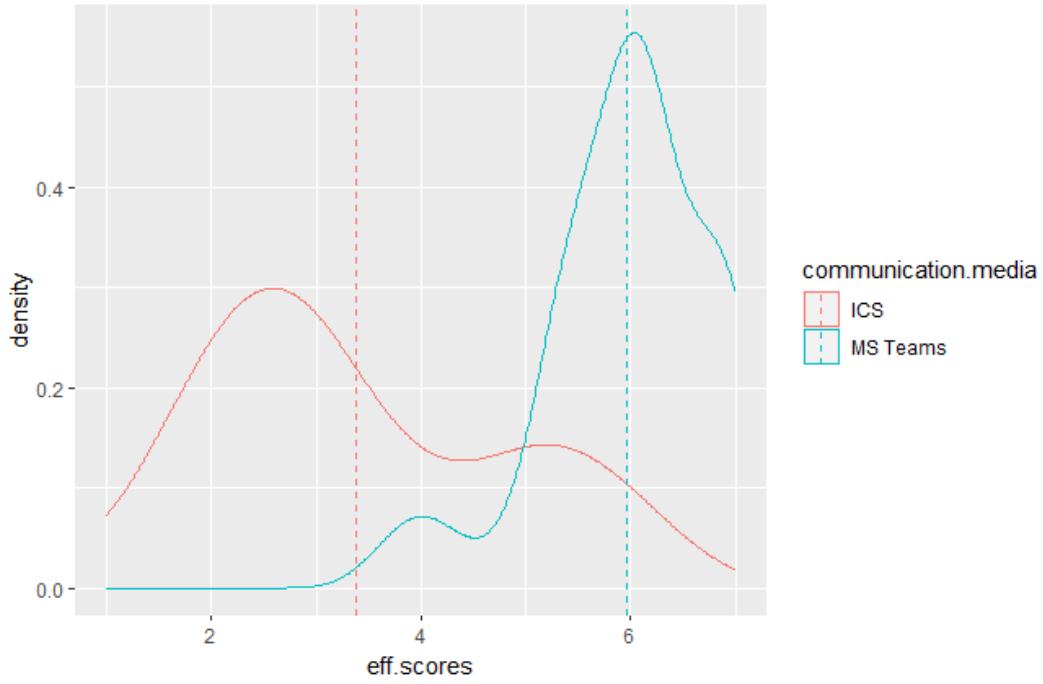


Figure 6.2: The distribution of effectiveness scores and the predicted mean of the two communication media.

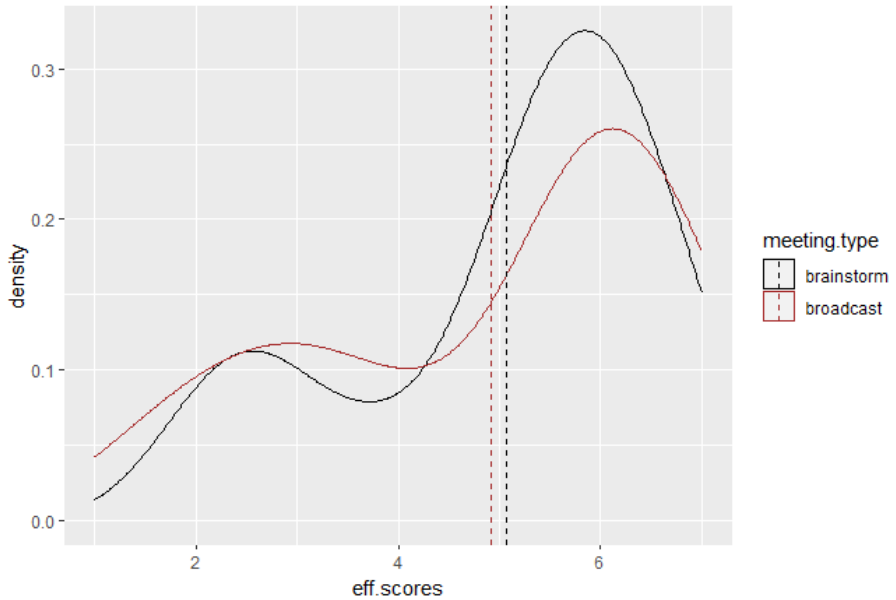


Figure 6.3: The distribution of effectiveness scores and the predicted mean of the two meeting types.

	diff	lwr	upr	p
Broadcast - Brainstorm	-0.1507092	-0.5906066	0.2891881	0.4978489
MS Teams - ICS	2.579082	2.125108	3.033057	$< .001$ ***
Broadcast : ICS - Brainstorm : ICS	-0.52124183	-1.4645.342	0.4220505	0.4739244
Brainstorm : MS Teams - Brainstorm : ICS	2.31513410	1.4782088	3.1520594	$< .001$ ***
Broadcast : MS Teams - Brainstorm : ICS	2.33237548	1.4954502	3.1693007	$< .001$ ***
Brainstorm : MS Teams - Broadcast : ICS	2.83637593	1.9843978	3.6883541	$< .001$ ***
Broadcast : MS Teams - Broadcast : ICS	2.85361731	2.0016392	3.7055955	$< .001$ ***
Broadcast : MS Teams - Brainstorm : MS Teams	0.01724138	-0.7152272	0.7497100	0.9999153

Table 6.8: The estimated mean difference, confidence interval limits, and p-value from the Tukey HSD test.

the interaction effects, both comparisons between brainstorm meetings (mean difference = 2.32, lower interval = 1.48, upper interval = 3.15, $p < .001$) and broadcast meetings (mean difference = 2.85, lower interval = 2, upper interval = 3.71, $p < .001$) in the two communication media are statistically significant. As explained above, there was no significant difference between meeting types, so this variable did not require a post hoc test. Table 6.22 summarizes the results of the post hoc test. The first column represents the differences in means between different levels; the second and third columns represent the lower and upper confidence interval for the difference in means; the last column represents the p-value. The first two rows result from the Tukey test on the combination of the communication media and meeting type; the other rows result from the interaction effects of the same variables[10][11]. To cover every aspect of the study, here and in the following post hoc tests, we report the data of every combination between variables, but we are not interested in the results of each of them. For example, we are not interested in the comparison between a broadcast meeting in the TNO ICS and a brainstorm meeting in MS Teams.

We chose the Bonferroni correction for the Multiple Test Correction. This test attempts to prevent data from incorrectly appearing to be statistically significant by making an adjustment during comparison testing[12]. The adjusted p-value[13] for the mean difference

[10] https://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/preport

[11] https://rstudio-pubs-static.s3.amazonaws.com/$266172_6ffd3f44e0f845459ae265684bfd2168.html$

[12] https://www.investopedia.com/terms/b/bonferroni-test.asp

[13] https://www.statology.org/bonferroni-correction-in-r/

in effectiveness scores between brainstorm and broadcast meetings is .66. Instead, the adjusted p-value for the mean difference in effectiveness scores between meetings in ICS and Ms Teams is $< .001$. Based on the output, the only significant difference is between communication media.

To show whether the effect of our variables on the meeting effectiveness is large enough to be meaningful in the real world, we calculated the effect size using Cohen's d[14]. The effect size of different meeting types, as measured by Cohen's d, was d = 0.09, indicating a small effect. Instead, the effect size of different communication media was d = 2.42, indicating a large effect[15][16].

It should be noted that, unlike the SUS questionnaire, both the Perceived Usefulness and H-MSC-Q questionnaires are not associated with an adjective rating scale to provide absolute judgment. Instead, we have to rely on the limited intuitiveness of a 7-point scale to present and compare the results.

6.1.4 RQ3: Social Communication

We present in this section the results of the statistical analysis conducted to answer the research questions on spatial presence "How does the spatial presence of communication media change with different numbers of meeting participants?" and "How does the spatial presence of meeting participants change in different communication media?", and the research questions on social presence "How does the social presence of meeting participants change in different communication media?" and "How does the social presence of communication media change with different numbers of meeting participants?". The two categorical, independent variables of this part of the study are "communication media" and "number of participants". The first contains the two levels "ICS" and "MS Teams", indicating the meeting platform used in the meeting; the second contains the two levels "2" and "3", indicating how many participants participated in the meeting. The continuous dependent variable is social communication, divided into social and spatial presence, and their values represent the scores obtained from the homonym sections of the H-MSC-Q questionnaire.

Given that the questionnaire allowed users to give their answers on a 7-point Likert scale, the results presented in this section will be in a range from 1 to 7. In total, 32 participants filled out 93 "H-MSC-Q" questionnaires; 50 of them rated meetings with two participants, while the remaining 43 evaluated meetings with three participants; lastly, 35 questionnaires are related to meetings in the TNO ICS, and 58 are associated with meetings in MS Teams. Table 6.9 summarizes the questionnaires divided per meeting conditions. The numbers in parenthesis indicate how many questionnaires, for both meeting conditions, were discarded.

RQ3.1: Spatial Presence

First, we used the Akaike Information Criterion to compare different possible models and determine which one is the best fit for the data. First, we test how each variable performs separately. Then, we want to know if the combination of communication media and the number of participants is better at describing variation in the spatial presence scores. Finally, we check whether the interaction of the two variables can explain the spatial

[14]https://www.scribbr.com/statistics/effect-size/
[15]https://statisticseasily.com/2023/04/06/how-to-report-cohens-d-in-apa/
[16]https://www.statology.org/interpret-cohens-d/

	2 participants	3 participants	TOTAL
ICS	26 (4)	9 (21)	35 (25)
MS Teams	24	34	58
TOTAL	50 (4)	43 (21)	93 (25)

Table 6.9: The number of H-MSC-Q questionnaires representing each level.

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	Res.LL
Communication Media	4	320.82	0.00	0.46	0.46	-156.19
Number of participants	4	321.43	0.61	0.34	0.79	-156.49
Communication Media + Numb. of participants	5	323.18	2.36	0.14	0.93	-156.25
Communication Media * Numb. of participants	6	324.68	3.86	0.07	1.00	-155.85

Table 6.10: The Model selection based on AICc.

presence scores better than any of the previous models. The results show that the best-fit model, carrying 46% of the cumulative model weight, included only the parameter communication media. The second best-fit model includes only the parameter "number of participants"; this model differs from the AIC score of the best model by 0.61 points. To cover every aspect of the analysis, we will also present the interaction effects of the two variables, given by the third model using the operator "*". Table 6.10 shows, for each of the four models, the number of parameters in the model (K), the information score of the model (AICc), the difference in AIC score between the best model and the model being compared (Delta_AICc), the proportion of the total amount of predictive power provided by the full set of models contained in the model being assessed (AICcWt), the sum of the AICc weights (Cum.Wt), and the value describing how likely the model is, given the data (LL).

We then fit our linear mixed-effect model with combinations between the independent variables and the combination of their factors, and subjects as random effects. Regarding the fixed effects, the estimated intercept returned a value of 4.07 for the dependent variable "spatial presence". The t-test performed to compare the means of number of participants on spatial presence showed that there was no significant difference ($t(df) = 0.22$, $p > 0.1$). The mean of meetings with three participants was 0.07 points (standard error: 0.32) higher than the mean of meetings with 2 people. Successively, the t-test performed to compare the means of communication media on spatial presence showed that there was not a significant difference ($t(df) = -0.991$ $p > 0.1$). The mean of meetings in MS Teams was 0.26 points (standard error: 0.27) lower than the mean of meetings in ICS. Lastly, to cover every aspect of the analysis, we fit our linear mixed-effect model with combinations between the independent variables and the interactions of their factors; the t-test was performed to compare the means of the interaction effects between number of participants and communication media on social presence showed that there was not a significant difference ($t(df) = 0.14$, $p > 0.1$). The mean of meetings with three participants in MS Teams was

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.0732	0.2303	68.1748	17.686	<.001 ***
Communication Media (MS Teams)	-0.2626	0.2650	74.7061	-0.991	0.325
Number of participants (3)	0.06909	0.31516	30.59366	0.219	0.828
Number of participants (3) : Communication Media (MS Teams)	0.08155	0.58717	74.18019	0.139	0.890

Table 6.11: The fixed effects in the linear mixed effect model.

Groups	Name	Variance	Std.Dev.
Subject	(Intercept)	0.2576	0.5075
Residual		1.4531	1.2054

Table 6.12: The random effects in the linear mixed effect model.

0.08 points (standard error: 0.59) higher than the mean of meetings with two participants in ICS. Table 6.11 displays the data of the fixed effects in our model; the "Estimate" column shows the estimated values of the coefficients for each fixed effect; the "Std. Error" column represents the standard errors associated with each coefficient estimate; the "df" column shows the degrees of freedom associated with each coefficient; the "t value" column shows the t-statistic associated with each coefficient estimate; lastly, the "Pr(>|t|)" column shows the p-value associated with each coefficient estimate. The random effect is given first by the subjects in our study, which are responsible for a variance of 0.26 points in the data (standard deviation: 0.51). Then, the residuals are responsible for a variance of 1.45 points in the data (standard deviation: 1.21). Table 6.12 displays the variance and standard deviation of the two random groups; the "Subject (Intercept)" random effect captures the variability in the intercepts of different subjects; the "Residual" random effect accounts for the remaining variability within groups that cannot be explained by the fixed effects or the "Subject (Intercept)" random effect.

Applying the ANOVA method for linear mixed effect regression model fits produces a type III ANOVA table for fixed-effect, which confirms that there was no significant difference ($F(1, 30.59) = 0.05$, $p > .1$) between number of participants; similarly, it confirms that there was no significant difference ($F(1, 74.71) = 0.98$, $p > .1$) between communication media; lastly, the ANOVA on the linear mixed-effect model with combinations between the independent variables and the interactions of their factors confirms that there was not a significant difference ($F(1, 74.18) = 0.02$, $p > .1$) between the interaction effects of communication media and number of participants. As before, the results are given by the linear mixed model using the communication media variable only, then the combination of the two variables for the insights on the number of participants, while the last line is taken

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
Communication Media	14.269	14.269	1	74.706	0.982	0.3249
Number of participants	0.069255	0.069255	1	30.594	0.0481	0.8279
Number of participants : Communication Media	0.02817	0.02817	1	74.180	0.0193	0.8899

Table 6.13: The type III Analysis of Variance Table.

Number of participants	Mean	SE	Df	LL(95%)	UL(95%)
2	3.8712	0.22322	25.46997	3.4118	4.3305
3	3.9402	0.22721	25.46997	3.4727	4.4078

Table 6.14: The expected mean, standard error, and confidence limits of the "number of participants" variable.

from the model using the interaction between the two variables. Table 6.13 represents the complete results of the ANOVA. The Df column displays the degrees of freedom for the independent variable (the number of levels in the variable minus 1), and the degrees of freedom for the residuals; the Sum Sq column displays the sum of squares (i.e. the total variation between the group means and the overall mean); the Mean Sq column is the mean of the sum of squares, calculated by dividing the sum of squares by the degrees of freedom for each parameter; the F value column is the test statistic from the F test; the Pr(>F) column is the p-value of the F statistic.

We used the function "predictmeans" to calculate the statistical values of the levels MS Teams (estimated mean = 3.81, standard error = 0.18, lower confidence limit = 3.44, upper confidence limit = 4.18) and TNO ICS (estimated mean = 4.07, standard error = 0.23, lower confidence limit = 3.61, upper confidence limit = 4.54), from the variable communication media. Secondly, we analyzed the two levels of the second variable "number of participants", namely three (estimated mean = 3.94, standard error = 0.23, lower confidence limit = 3.47, upper confidence limit = 4.41) and two participants (estimated mean = 3.87, standard error = 0.22, lower confidence limit = 3.41, upper confidence limit = 4.33). Table 6.14 and 6.15 show the complete data returned by the function "predictmeans". The first column contains the predicted means; the standard error on the predicted mean is in the second column; the degrees of freedom are in the third column; the lower and upper limits of the confidence interval are in the fourth and fifth columns. Figures 6.4 and 6.5 show the distribution of the effectiveness scores of the two communication media and their predicted mean (shown by a dotted line).

There was no need for a post hoc test, as we did not observe any statistical significance in the model.

We chose the Bonferroni correction for the Multiple Test Correction. The adjusted p-value for the mean difference in spatial presence scores between meetings with two and three participants is .96. Instead, the adjusted p-value for the mean difference in spatial presence scores between meetings in ICS and MS Teams is .33. Based on the output, there is once again no statistical difference.

Communication Media	Mean	SE	Df	LL(95%)	UL(95%)
ICS	4.0732	0.23031	50.68074	3.6108	4.5357
MS Teams	3.8106	0.18398	50.68074	3.4412	4.1800

Table 6.15: The expected mean, standard error, and confidence limits of the communication media.

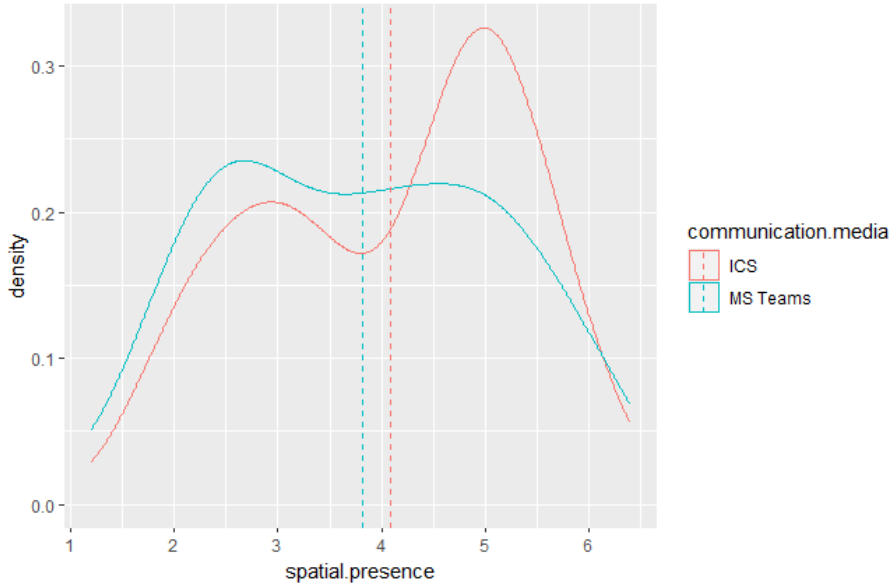


Figure 6.4: The distribution of spatial presence scores and the predicted mean of the two communication media.

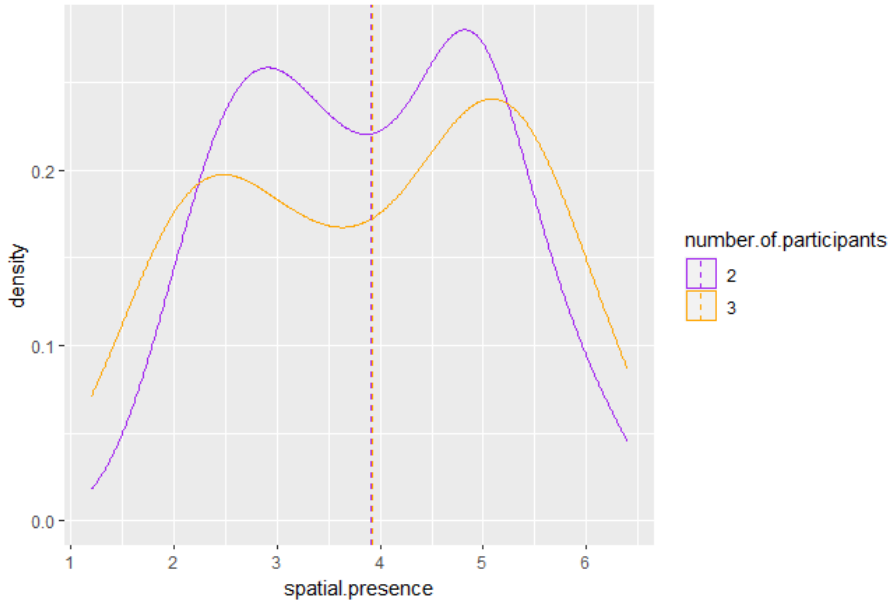


Figure 6.5: The distribution of spatial presence scores and the predicted mean for meetings with two and three participants.

Model	K	AICc	Delta_AICc	AICcWt	Cum.Wt	Res.LL
Communication Media	4	210.28	0.00	0.85	0.85	-100.91
Communication Media + Numb. of participants	5	214.29	4.01	0.11	0.96	-101.80
Communication Media * Numb. of participants	6	216.63	6.36	0.04	1.00	-101.83
Number of participants	4	222.50	12.23	0.00	1.00	-107.02

Table 6.16: The Model selection based on AICc.

To show whether the effect of our variables on the spatial presence is large enough to be meaningful in the real world, we calculated the effect size using Cohen's d. The effect size of different numbers of participants, as measured by Cohen's d, was d = 0.01, indicating a small effect. Additionally, the effect size of different communication media was d = 0.21, indicating a small effect. We follow the commonly used rule of thumb to interpret Cohen's d: a value of 0.2 represents a small effect size; a value of 0.5 represents a medium effect size; a value of 0.8 represents a large effect size.

RQ3.2: Social Presence

First, we used the Akaike Information Criterion to compare different possible models and determine which one is the best fit for the data. First, we test how each variable performs separately. Then, we want to know if the combination of communication media and the number of participants is better at describing variation in the social presence scores. Finally, we check whether the interaction of the two variables can explain the social presence scores better than any of the previous models. The results show that the best-fit model, carrying 85% of the cumulative model weight, included only the parameter communication media. The second best-fit model includes the combination (but no interaction) between the two variables; this model differs from the AIC score of the best model by 4.01 points out of 210.28 Although the best-fit model only includes the communication media parameter, for the goal of the analysis we will also include the combination with the "number of participants" parameter; to cover every aspect of the analysis, we will also present the interaction effects of the two variables, given by the third model using the operator "*". Table 6.16 shows, for each of the four models, the number of parameters in the model (K), the information score of the model (AICc), the difference in AIC score between the best model and the model being compared (Delta_AICc), the proportion of the total amount of predictive power provided by the full set of models contained in the model being assessed (AICcWt), the sum of the AICc weights (Cum.Wt), and the value describing how likely the model is, given the data (LL).

We then fit our linear mixed-effect model with combinations between the independent variables and the combination of their factors, and subjects as random effects. Regarding the fixed effects, the estimated intercept returned a value of 4.75 for the dependent variable "social presence". The t-test performed to compare the means of the number of participants on social presence showed that there was no significant difference ($t(df) = 0.09$, $p > .1$). The mean of meetings with three participants was 0.01 points (standard error: 0.16) higher than the mean of meetings with 2 people. Successively, the t-test performed to compare

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.7759	0.1213	76.3772	39.374	« .001 ***
Communication Media (MS Teams)	0.5706	0.1483	81.6575	3.848	0.000235 ***
Number of participants (3)	0.01396	0.16467	39.64664	0.085	0.932857
Number of participants (3) : Communication Media (MS Teams)	-0.1903	0.3284	81.6131	-0.580	0.56384

Table 6.17: The fixed effects in the linear mixed effect model.

Groups	Name	Variance	Std.Dev.
Subject	(Intercept)	0.02698	0.1642
Residual		0.46985	0.6855

Table 6.18: The random effects in the linear mixed effect model.

the means of communication media on social presence showed that there was a significant difference (t(df) = 3.62 p < .001). The mean of meetings in MS Teams was 0.57 points (standard error: 0.15) higher than the mean of meetings in ICS. Lastly, to cover every aspect of the analysis, we fit our linear mixed-effect model with combinations between the independent variables and the interactions of their factors; the t-test was performed to compare the means of the interaction effects between the number of participants and communication media on social presence showed that there was not a significant difference (t(df) = -0.58, p > .1). The mean of meetings with three participants in MS Teams was 0.19 points (standard error: 0.33) lower than the mean of meetings with two participants in ICS. Table 6.17 displays the data of the fixed effects in our model; the "Estimate" column shows the estimated values of the coefficients for each fixed effect; the "Std. Error" column represents the standard errors associated with each coefficient estimate; the "df" column shows the degrees of freedom associated with each coefficient; the "t value" column shows the t-statistic associated with each coefficient estimate; lastly, the "Pr(>|t|)" column shows the p-value associated with each coefficient estimate. The random effect is given first by the subjects in our study, which are responsible for a variance of 0.03 points in the data (standard deviation: 0.16). Then, the residuals are responsible for a variance of 0.47 points in the data (standard deviation: 0.69). Table 6.18 displays the variance and standard deviation of the two random groups; the "Subject (Intercept)" random effect captures the variability in the intercepts of different subjects; the "Residual" random effect accounts for the remaining variability within groups that cannot be explained by the fixed effects or the "Subject (Intercept)" random effect.

Applying the ANOVA method for linear mixed effect regression model fits produces a type III ANOVA table for fixed-effect, which confirms that there was no significant difference (F(1, 39.65) = 0.01, p > 0.1) between number of participants; similarly, it confirms that there was a significant difference (F(1, 78.23) = 13.10, p < .001) between communication media; lastly, the ANOVA on the linear mixed-effect model with combinations between

Model	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
Communication Media	69.572	69.572	1	81.657	14.807	0.0002353 ***
Number of participants	0.0034	0.0034	1	39.647	0.0072	0.9328569
Number of participants : Communication Media	0.1594	0.1594	1	81.613	0.3358	0.563839

Table 6.19: The type III Analysis of Variance Table.

the independent variables and the interactions of their factors confirms that there was not a significant difference ($F(1, 81.61) = 0.34$, $p > 0.1$) between the interaction effects of communication media and number of participants. As before, the results are given by the linear mixed model using the communication media variable only, then the combination of the two variables for the insights on the number of participants, while the last line is taken from the model using the interaction between the two variables. Table 6.19 represents the complete results of the ANOVA. The Df column displays the degrees of freedom for the independent variable (the number of levels in the variable minus 1), and the degrees of freedom for the residuals; the Sum Sq column displays the sum of squares (i.e. the total variation between the group means and the overall mean); the Mean Sq column is the mean of the sum of squares, calculated by dividing the sum of squares by the degrees of freedom for each parameter; the F value column is the test statistic from the F test; the Pr(>F) column is the p-value of the F statistic.

We used the function "predictmeans" to calculate the statistical values of the levels MS Teams (estimated mean = 5.35, standard error = 0.1, lower confidence limit = 5.16, upper confidence limit = 5.54) and TNO ICS (estimated mean = 4.78, standard error = 0.12, lower confidence limit = 4.53, upper confidence limit = 5.02), from the variable communication media. Secondly, we analyzed the two levels of the second variable "number of participants", namely three (estimated mean = 5.07, standard error = 0.12, lower confidence limit = 4.81, upper confidence limit = 5.32) and two participants (estimated mean = 5.05, standard error = 0.11, lower confidence limit = 4.83, upper confidence limit = 5.28). Table 6.20 and 6.21 show the complete data returned by the function "predictmeans". The first column contains the predicted means; the standard error on the predicted mean is in the second column; the degrees of freedom are in the third column; the lower and upper limits of the confidence interval are in the fourth and fifth columns. Figures 6.6 and 6.7 show the distribution of the social presence scores of the two communication media and their predicted mean (shown by a dotted line).

We chose the Tukey HSD for our multiple comparison test. The post hoc comparisons using the Tukey HSD test indicated that the mean score for the ICS meeting condition was significantly different than the MS Teams meeting condition ($p < .001$). Considering the interaction effects, only the comparison between meetings with two participants (mean difference = 0.64, lower interval = 0.11, upper interval = 1.16, $p < .05$) in the two com-

Number of participants	Mean	SE	Df	LL(95%)	UL(95%)
2	5.0546	0.10828	22.72168	4.8305	5.2788
3	5.0686	0.12321	22.72168	4.8135	5.3236

Table 6.20: The expected mean, standard error, and confidence limits of the "number of participants" variable.

Communication Media	Mean	SE	Df	LL(95%)	UL(95%)
ICS	4.7759	0.12129	57.44391	4.5330	5.0187
MS Teams	5.3465	0.09501	57.44391	5.1563	5.5367

Table 6.21: The expected mean, standard error, and confidence limits of the communication media.

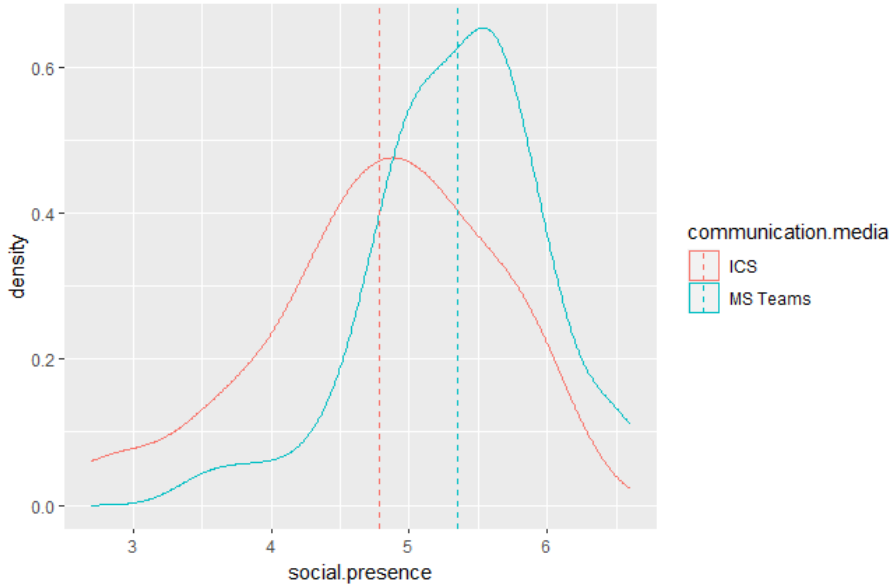


Figure 6.6: The distribution of social presence scores and the predicted mean of the two communication media.

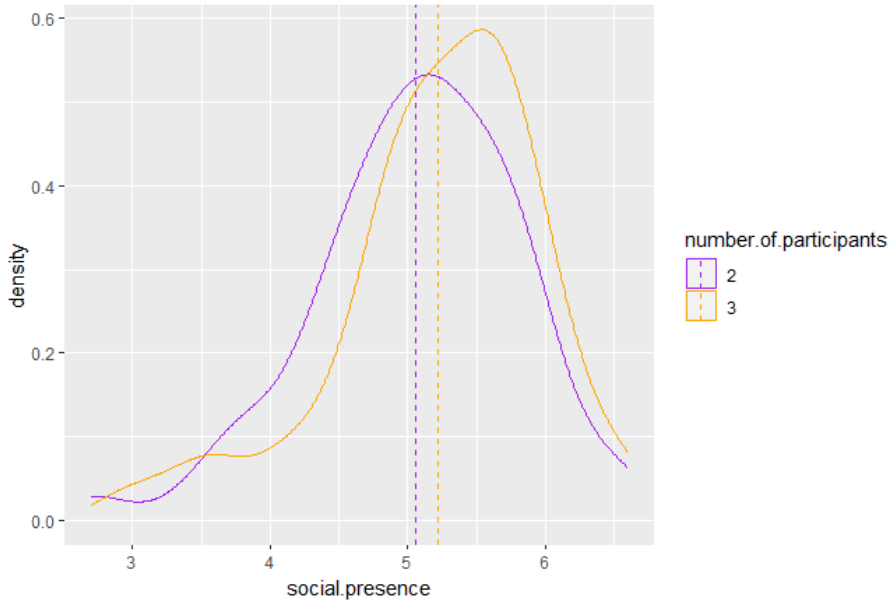


Figure 6.7: The distribution of social presence scores and the predicted mean for meetings with two and three participants.

munication media is statistically significant. As explained above, there was no significant difference between meeting types, so this variable did not require a post hoc test. Table 6.22 summarizes the results of the post hoc test. The first column represents the differences in means between different levels; the second and third columns represent the lower and upper confidence interval for the difference in means; the last column represents the p-value. The first two rows result from the Tukey test on the combination of the communication media and number of participants; the other rows result from the interaction effects of the same variables.

We chose the Bonferroni correction for the Multiple Test Correction. The adjusted p-value for the mean difference in social presence scores between meetings with two and three participants is .28. Instead, the adjusted p-value for the mean difference in social presence scores between meetings in ICS and Ms Teams is $< .001$. Based on the output, the only significant difference is between communication media.

To show whether the effect of our variables on social presence is large enough to be meaningful in the real world, we calculated the effect size using Cohen's d. The effect size of different numbers of participants, as measured by Cohen's d, was d = 0.23, indicating a small effect. Instead, the effect size of different communication media was d = 0.81, indicating a large effect. We follow the commonly used rule of thumb to interpret Cohen's d: a value of 0.2 represents a small effect size; a value of 0.5 represents a medium effect size; a value of 0.8 represents a large effect size.

6.1.5 RQ4: Extraneous Variables

We present in this section the results of the exploratory analysis conducted to answer the research question "How do extraneous variables and further interaction effects influence the subjective QoE factors?". We did not employ any independent or dependent variable for this part of the study, as it is an exploratory analysis. Our goal is instead to find trends in the data, with the help of summary statistics and graphical representations, that might be of interest for future research. We will discuss in the following sections the results of

	diff	lwr	upr	p adj
3 - 2	0.1695814	-0.1244874	0.4636502	0.2549347
MS Teams - ICS	0.5124727	0.209837	0.8151085	0.0011323 **
3 : ICS - 2 : ICS	0.11666667	-0.60389195	0.8372253	0.9742487
2 : MS Teams - 2 : ICS	0.63750000	0.11010509	1.1648949	0.0112201 *
3 : MS Teams - 2 : ICS	0.57058824	0.08519612	1.0559804	0.0144696 *
2 : MS Teams - 3 : ICS	0.52083333	-0.20740464	1.2490713	0.2472347
3 : MS Teams - 3 : ICS	0.45392157	-0.24449855	11.523.417	0.3289320
3 : MS Teams - 2 : MS Teams	-0.06691176	-0.56363233	0.4298088	0.9848533

Table 6.22: The estimated mean difference, confidence interval limits, and p-value from the Tukey HSD test.

the extraneous variables and interaction effects for each of the three dependent variables.

Usability

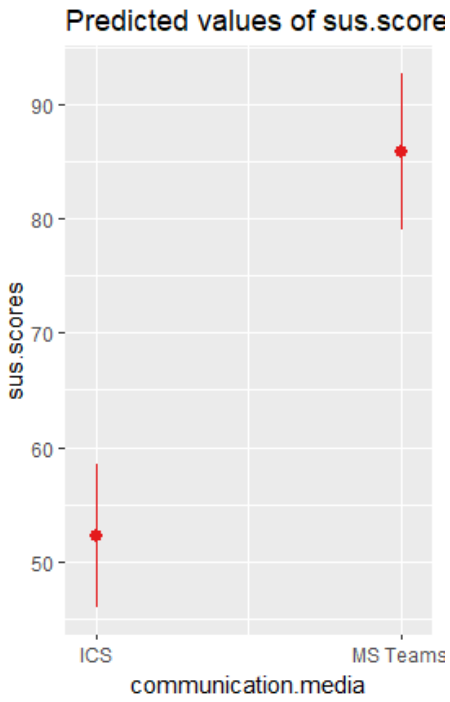
First, we used the Akaike Information Criterion to compare different possible models and determine which one is the best fit for the data. First, we test how each independent variable (i.e. communication media, meeting type, number of participants) performs separately. Then, we want to know if the combination of the three variables is better at describing variation in the effectiveness scores. Finally, we check whether the interaction of the three variables can explain the effectiveness better than any of the previous models. The results show that the best-fit model, carrying 100% of the cumulative model weight, included the interaction between the three independent variables. Then, we conducted additional model testing, beginning with the best-fitting model and progressively incorporating the five variables obtained from the initial questionnaire section (i.e. age, digital literacy, participants' relationship, level of VR experience, and meeting order). The results show that the best-fit model, carrying 100% of the cumulative model weight, included the interaction between the three variables and the level of VR experience participants had.

Figure 6.8 shows the predicted means of the SUS scores depending on the three independent variables and the level of VR experience of participants.

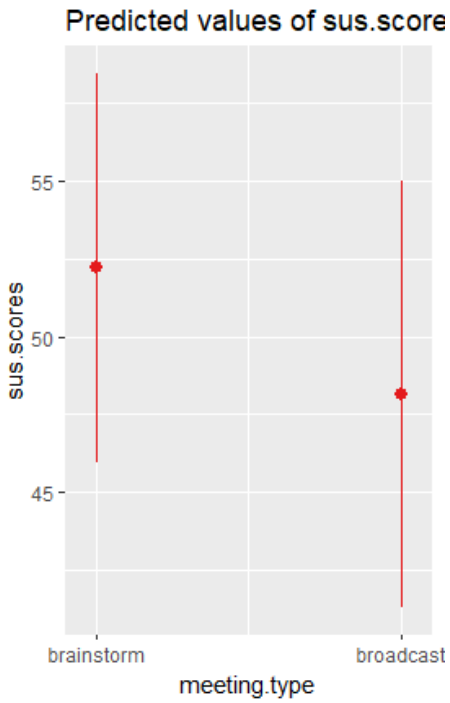
We then grouped the effectiveness scores into ranges of one unit, creating 7 levels (1, 2, 3, ..., 7). We repeated this process for the social and spatial presence scores. We then used these ranges as an independent variable to investigate whether they had an effect on the SUS scores. Figure 6.9 shows the expected mean in SUS scores depending on the ranges of effectiveness, spatial, and social presence scores given for the same meeting experience.

Effectiveness

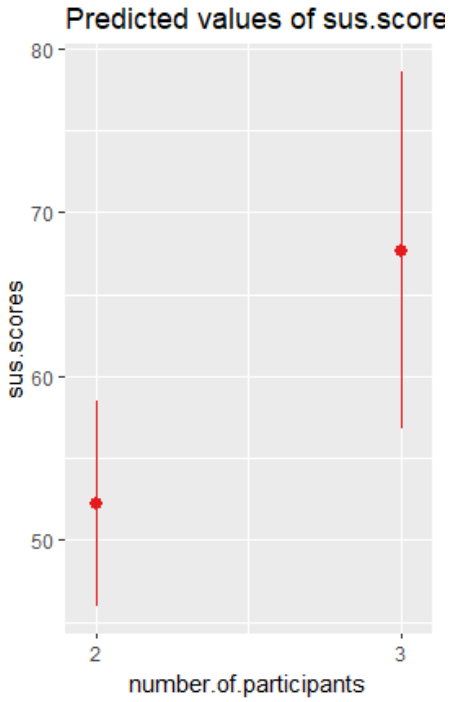
First, we used the Akaike Information Criterion to compare different possible models and determine which one is the best fit for the data. First, we test how each variable (i.e. communication media, meeting type, number of participants) performs separately. Then, we want to know if the combination of the three variables is better at describing variation in the effectiveness scores. Finally, we check whether the interaction of the three variables



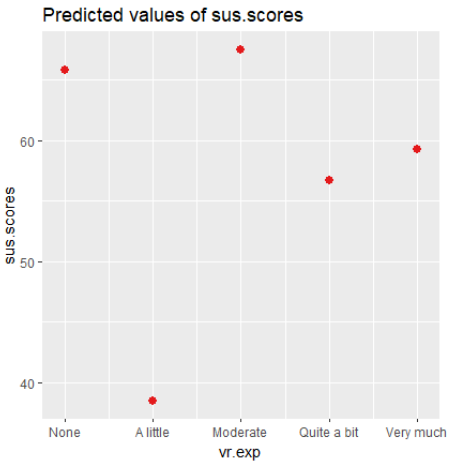
(a) Communication media



(b) Meeting type

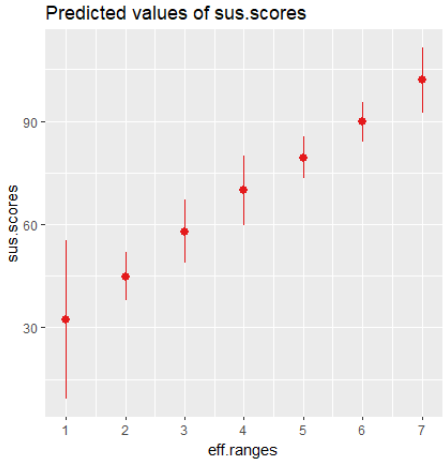


(c) Number of participants

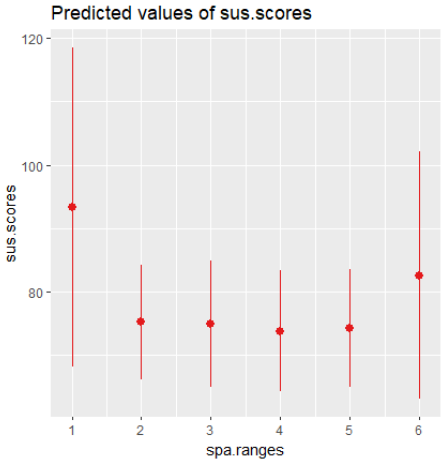


(d) Level of VR experience

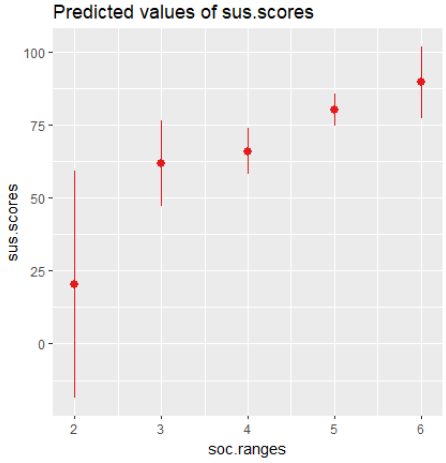
Figure 6.8: The predicted usability scores based on the three independent variables and the level of VR experience.



(a) Effectiveness



(b) Spatial Presence



(c) Social Presence

Figure 6.9: The predicted usability scores based on the ranges of the three dependent variables.

can explain the effectiveness better than any of the previous models. The results show that the best-fit model, carrying 84% of the cumulative model weight, included only the ranges of usability scores as a fixed effect, and subject ID as a random effect. Then, we conducted additional model testing, beginning with the best-fitting model and progressively incorporating the five variables obtained from the initial questionnaire section (i.e. age, digital literacy, participants' relationship, level of VR experience, and meeting order). The results show that the best-fit model is the same as before.

Figure 6.10 shows the predicted means of the effectiveness scores depending on the ranges of the usability scores.

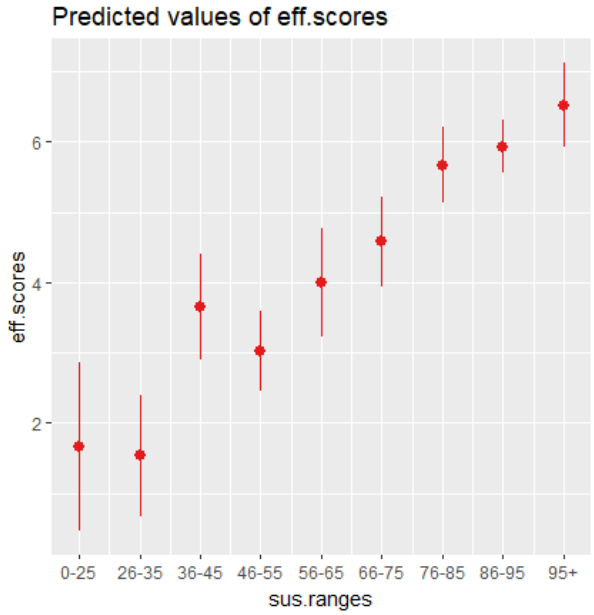


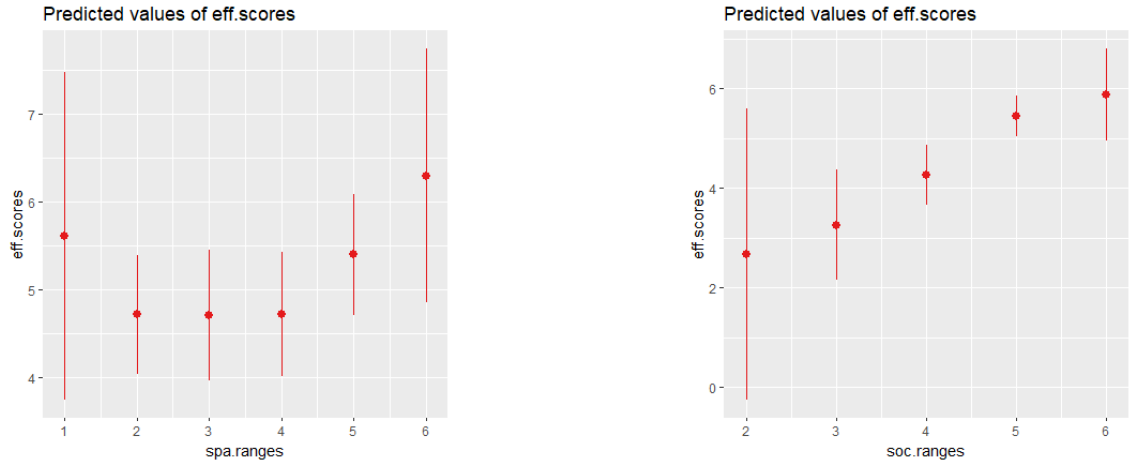
Figure 6.10: The predicted means of effectiveness scores per each SUS range.

We then grouped the spatial and social presence scores into ranges of one unit, creating 7 levels (1, 2, 3, ..., 7). We finally used these ranges as an independent variable to investigate whether they had an effect on the effectiveness scores.

Figure 6.11 shows the expected mean in SUS scores depending on the ranges of effectiveness, spatial, and social presence scores given for the same meeting experience.

Spatial Presence

First, we used the Akaike Information Criterion to compare different possible models and determine which one is the best fit for the data. First, we test how each variable (i.e. communication media, meeting type, number of participants) performs separately. Then, we want to know if the combination of the three variables is better at describing variation in the effectiveness scores. Finally, we check whether the interaction of the three variables can explain the effectiveness better than any of the previous models. The results show that the best-fit model, carrying 23% of the cumulative model weight, included only the social presence variable. Then, we conducted additional model testing, beginning with the best-fitting model and progressively incorporating the five variables obtained from the initial questionnaire section (i.e. age, digital literacy, participants' relationship, level of VR experience, and meeting order). The results show that the best-fit model is the same as before. Figure 6.12 shows the predicted means of the spatial presence scores depending



(a) Spatial Presence

(b) Social Presence

Figure 6.11: The predicted effectiveness scores based on the ranges of social and spatial presence scores.

on the ranges of the social presence scores.

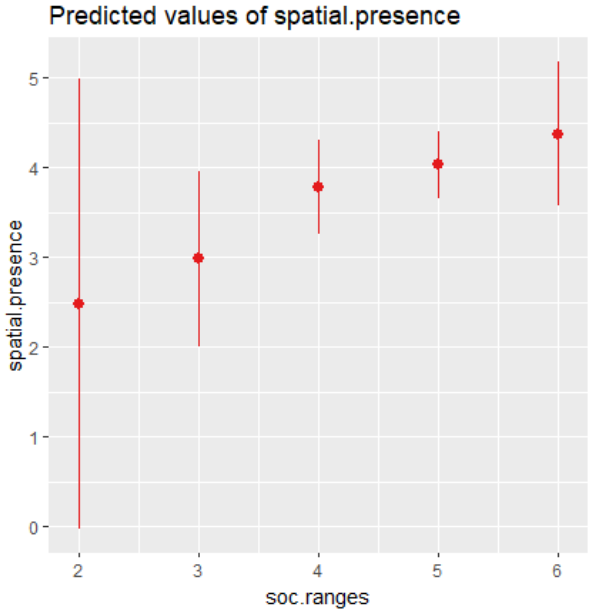
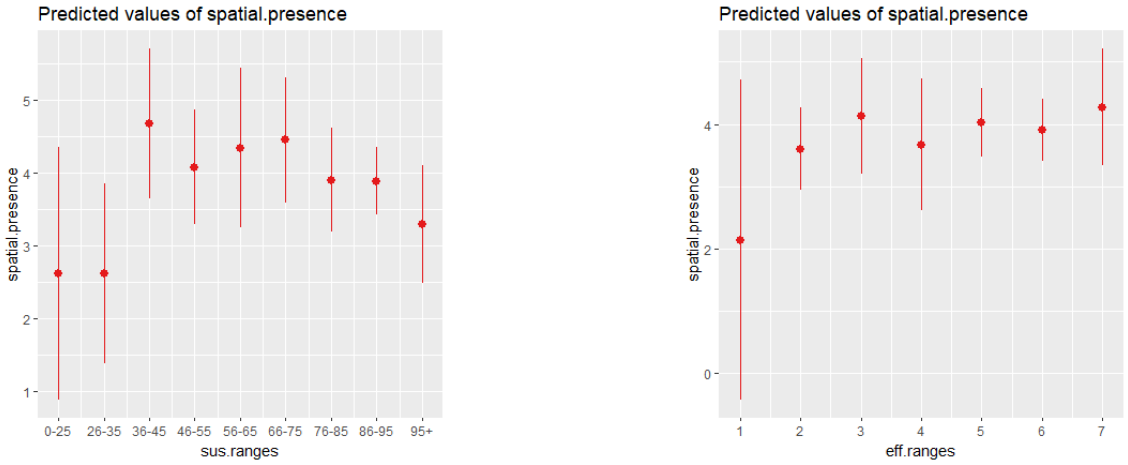


Figure 6.12: The predicted means of spatial presence scores for each range of the social presence scores.

We then grouped the SUS scores into ranges spanning 10 points, creating 9 levels (0-25, 26-35, 36-45, ..., 95+). We also grouped the effectiveness scores into ranges of one unit, creating 7 levels (1, 2, 3, ..., 7). We then used these ranges as independent variables to test whether they had an effect on the spatial presence scores. Figure 6.13 shows the expected mean in spatial presence scores depending on the ranges of usability and effectiveness scores given for the same meeting experience.



(a) Usability

(b) Effectiveness

Figure 6.13: The predicted spatial presence scores based on the ranges of usability and effectiveness scores.

Social Presence

First, we used the Akaike Information Criterion to compare different possible models and determine which one is the best fit for the data. First, we test how each variable (i.e. communication media, meeting type, number of participants) performs separately. Then, we want to know if the combination of the three variables is better at describing variation in the effectiveness scores. Finally, we check whether the interaction of the three variables can explain the effectiveness better than any of the previous models. The results show that the best-fit model, carrying 90% of the cumulative model weight, included only the effectiveness scores as a fixed effect, and subject ID as a random effect. Then, we conducted additional model testing, beginning with the best-fitting model and progressively incorporating the five variables obtained from the initial questionnaire section (i.e. age, digital literacy, participants' relationship, level of VR experience, and meeting order). The results show that the best-fit model is the same as before.

Figure 6.14 shows the predicted means of the social presence scores depending on the ranges of the effectiveness scores.

We then grouped the SUS scores into ranges spanning 10 points, creating 9 levels (0-25, 26-35, 36-45, ..., 95+). We also grouped the spatial presence scores into ranges of one unit, creating 7 levels (1, 2, 3, ..., 7). We then used these ranges as independent variables to test whether they had an effect on the spatial presence scores. Figure 6.15 shows the expected mean in spatial presence scores depending on the ranges of usability and effectiveness scores given for the same meeting experience.

The section presenting the results of the questionnaires is now concluded. We have presented the results for each of the four RQs. The next chapter will present the results of the interviews, related to RQ5.

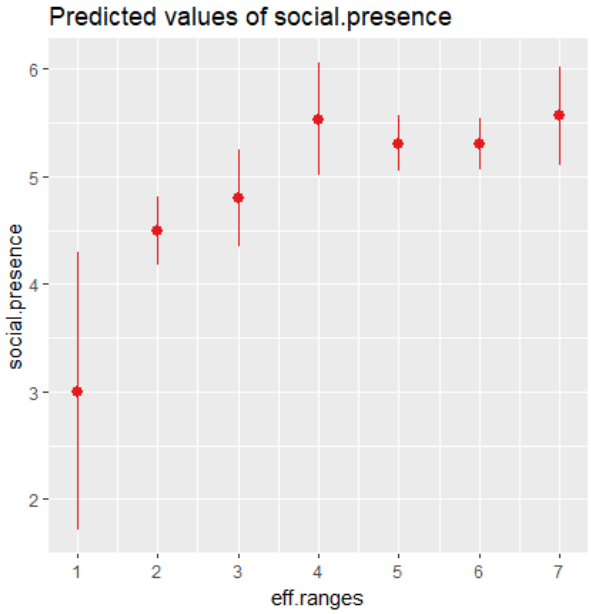
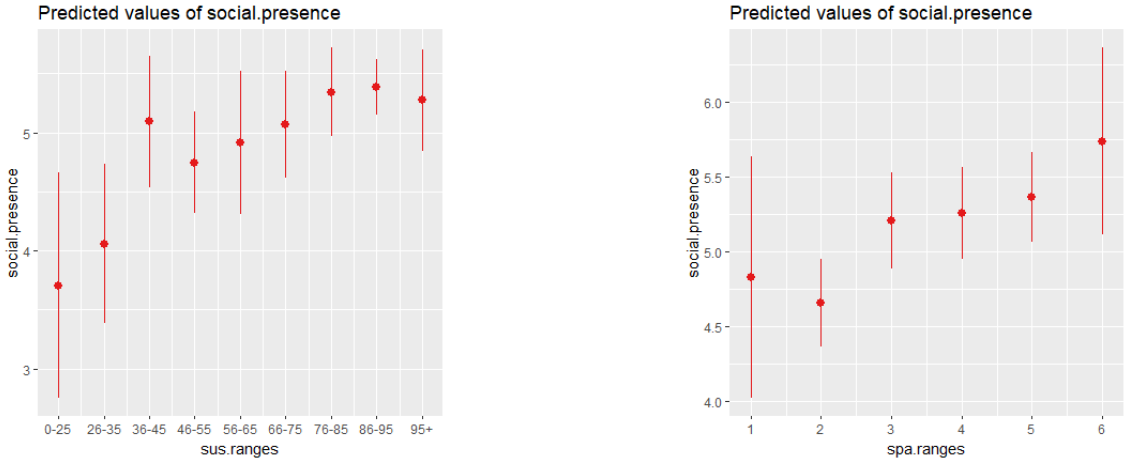


Figure 6.14: The predicted means of social presence scores for each range of the effectiveness scores.



(a) Usability

(b) Spatial Presence

Figure 6.15: The predicted social presence scores based on the ranges of usability and spatial presence scores.

6.2 Interviews Results

6.2.1 RQ5: Other Explanatory Insights

We present in this section the results of the thematic analysis conducted to answer the research question "Which other subjective factors influence the meeting QoE?" from the qualitative data collected from the interviews. We did not employ any independent or dependent variable for this part of the study, as it is an exploratory analysis. Our goal is instead to find factors that negatively or positively affected the meeting experience, which could not be observed through a questionnaire. We summarized the comments into three groups, depending on the topics or subjects that they address, namely the systems, the QoE factors, and the meeting preferences; then, all comments are grouped based on their valence, namely, positive, negative, or constructive (neutral). We did not divide the dependent variable "social communication" into its component social and spatial presence because, unlike the questionnaire data, it was hard to assign comments to either social or spatial presence, as some of them related to both of them.

Starting from the comments related to the QoE factors, it emerged that the main issues are related to the HMD covering the partner's eyes, the Pointcloud being displayed badly or in the wrong position, and the presence of technical issues as mentioned before. The first two issues are related to social presence, while the third is related to usability. As for the suggestions, participants mostly expressed their desire to interact with the environment, to have a whiteboard, and to be able to share their screen and notes. The first idea is part of the social presence cluster, while the other two are part of the effectiveness cluster. Lastly, TNO employees appreciated seeing each other in the Metaverse the most, followed by the feeling of being there, and the ability to see the hands and body of their partner.

For all three factors, we can divide the comments into those related to the hardware and comments related to the software. Examples from the first group are the HMD being too heavy and the virtual environment making participants uncomfortable or dizzy; this category of comments is related to the hardware we chose, and they address in particular the Oculus Meta Quest 2. On the other hand, examples of comments related to the software are the system being unintuitive, and the setups being long and complex.

In the third group, we collected comments on what would be the ideal and not ideal meetings to have in the two communication media. The majority of the interviewees indicated brainstorm meetings as the most feasible type of meeting to be held in the Metaverse, followed by meetings with a large group of people. On the other hand, most users reportedly would not use the Metaverse for 1-on-1 meetings. Contrary to expectations, we find it challenging to define this data as reliable. Most participants provided feedback that indicates the system is not yet at an optimal stage, making them hesitant about holding any type of meeting in the current system. Furthermore, their responses regarding the preferred meeting to be held in the TNO ICS seem to refer to an envisioned, desired system, which may differ from the final product. Additionally, when asked how many employees would be in a "large group of people", as they mentioned, many respondents could only give a vague answer (e.g. "more than four", or "just big groups").

As mentioned earlier, some comments were too interesting to be merely coded. Therefore, we collected together the most captivating observations, and shared them with the team, to spark discussions and generate insights that could potentially contribute to the refinement of our study and the enhancement of the TNO ICS. For example, participant 2B1X was satisfied with the ICS effectiveness and mentioned that the feeling of being together in a meeting entices him into being active and not just disappearing in the background like on Teams. On the contrary, user 1D1X was severely dissatisfied with the

system, and commented on the HMD blocking the participants' faces saying that "Meta's avatars are much better than a Pointcloud without a face".

Lastly, all comments were coded based on their valence: positive feedback, comprising positive impressions of the system, appreciated features, and positive experiences; then, constructive feedback, consisting of suggestions for improvement and desired features; lastly, negative feedback, encompassing dislikes, concerns, and negative experiences.

Figure 6.16 depicts a graphical overview of the qualitative data analysis from the interviews. The graph presents the themes most frequently coded in the interviews, divided by the QoE factors they are referring to, and distinguished according to their valence. Each code on the x-axis is grouped into its related QoE factor and is colored green, orange, or red, depending on their belonging to either the Positive, Negative, or Constructive Feedback group. The numbers on the y-axis reflect the number of interviews in which each code appeared at least once. We decided not to show the total number of times a code was applied to user comments, as participants sometimes mentioned the same occurrence repeatedly in the same session.

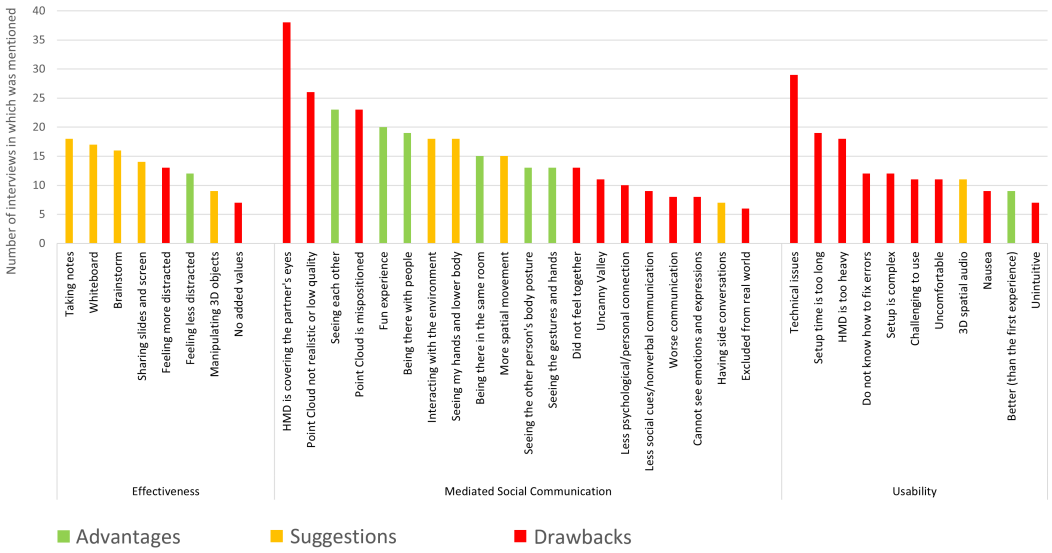


Figure 6.16: The main codes collected during the interview analysis.

Figure 6.17 shows the relationship between the groups of codes: all comments have a valence, meaning that they all either represent a positive comment, a negative comment, or a need from the participants; then, comments could either refer to one of the three subjective QoE factors, or to the meeting conditions in which participants would or would not use the Metaverse for; in the first case, comments could also be associated to either the software or the hardware.

This section concludes the chapter on the results. To summarize, the results of the statistical analyses on the quantitative data show that MS Teams received significantly better scores than the TNO ICS in usability, effectiveness, and social presence, but not in spatial presence. Then, the differences in meeting type and number of participants did not have a statistically significant effect on any dependent variables. Regarding the thematic analysis of the qualitative data, the usability factor received mostly negative comments; the effectiveness factor received mostly constructive comments from the needs of the participants; and lastly, social communication received a balanced amount of positive, negative, and constructive comments.

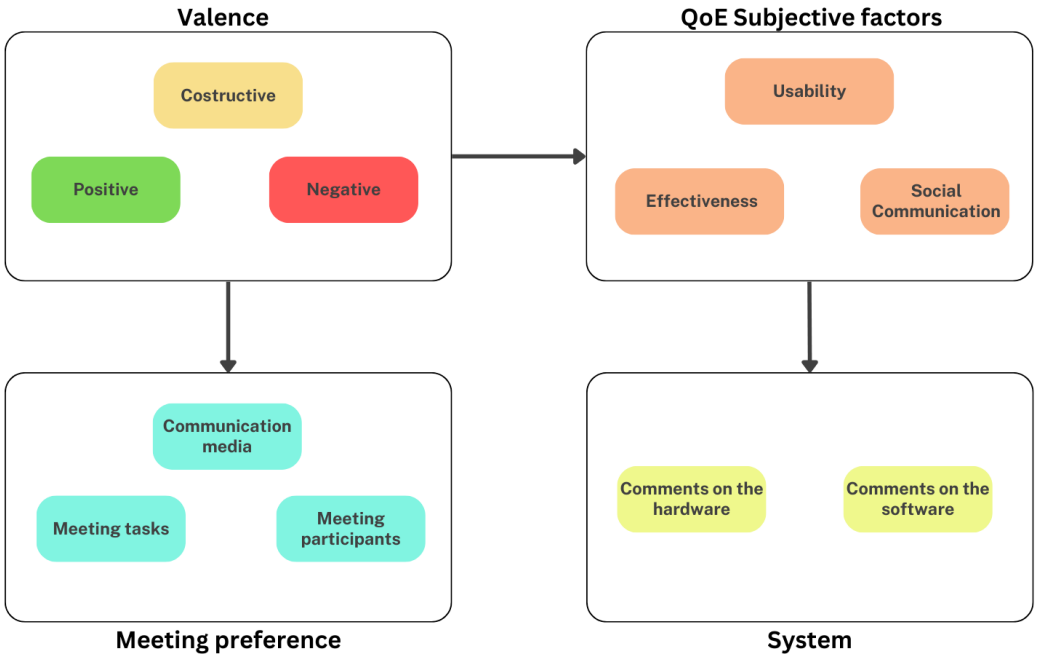


Figure 6.17: The relationship between codes collected during the interview analysis.

The next chapter will discuss the results presented in this chapter.

Chapter 7

Discussion

7.1 RQ1: Usability

RQ1 is written as "How does usability change in different communication media?". For RQ1, the null hypothesis is that there is no difference in usability scores between the TNO ICS and MS Teams. Our alternative hypothesis is that there is a difference in the usability of communication media, as MS Teams is expected to perform better than the ICS. The results of the lmer model and ANOVA showed a difference in SUS scores of 36.64 points; furthermore, the p-value is lower than our significance level of 0.05, so we can consider our results statistically significant. The post hoc test confirmed the significance, and the results indicated that MS Teams is expected to receive higher scores. Furthermore, Cohen's d confirmed the large effect size, thus proving it would be meaningful in the real world. These results reject the null hypothesis; the difference in usability can be attributed to the different communication media; more specifically, MS Teams is the meeting platform that received the higher grades. It should be noted that the variance of the random effects is large; this result indicates that the usability ratings also vary considerably between subjects, according to their perception of the system.

We assume that the user interface and the software components of the TNO ICS did not perform at their best capacity as they are still under development, thus providing participants with a negative user experience. Instead, software such as MS Teams, which are the standard in modern companies and are constantly updated and refined, thus providing users with better tools and interfaces to complete their goals. As seen in the results of the interview sessions, the main usability pain points (i.e. technical issues, long setup time, and troubleshooting problems) derive from the "technical immaturity" of the technology. We are confident that the future improvements made to the software will greatly improve the system's usability, especially now that we have highlighted what needs to be improved the most.

We also noticed an additional discussion point for the usability of the TNO ICS in table 6.9, where we reported that 21 questionnaires related to meetings with three participants had to be discarded, meaning that seven meetings were affected by serious technical issues. On the other hand, only 4 questionnaires related to meetings with two participants had to be discarded, meaning that only two meetings were affected by serious technical issues. Such a discrepancy shows how the system is less feasible to be deployed for meetings other than "1-on-1", which should be one of the strengths of the system. We suppose that issues are caused first by the instability of the system, which struggles to connect more than two users at once, and by a simple matter of probability, as connecting more users together increases the likelihood of having at least one defective system that might

negatively influence the experience for all participants.

7.2 RQ2: Effectiveness

RQ2 states "How does the perceived effectiveness of meeting types change in different communication media?" and "How does the perceived effectiveness of communication media change in different meeting types?". For simplicity, we label the first question "RQ2.1", and the latter "RQ2.2". The first null hypothesis of RQ2 is that there is no relationship between perceived effectiveness and meeting type. Our alternative hypothesis is that meeting type has an effect on the meeting effectiveness; more specifically, we anticipated brainstorm meetings to receive higher ratings than broadcast meetings in the TNO ICS condition, and worse ratings in the MS Teams condition. The results of the lmer model and ANOVA showed a difference in effectiveness ratings of 0.17 points between meeting types; furthermore, the p-value is higher than our significance level of 0.05, so we can consider our results not statistically significant. The variance of the random effects is small; this result indicates that the effectiveness ratings do not vary between subjects, according to their perception of the system. Then, the estimated mean difference between different meeting types in the same communication media equaled 0.5 and 0.01 points, and both had a p-value higher than the significance level. The Bonferroni correction test confirmed the non-significance of the results. Lastly, Cohen's d returned a small effect size, thus proving that a difference in meeting types would not be meaningful in the real world. These results do not reject the null hypothesis; the difference in effectiveness, both in the TNO ICS and MS Teams, cannot be attributed to the different meeting types. We believe that this result is given by the tasks being too similar. Although they are different on a theoretical level, we could not put them into practice as we wanted to, because the TNO ICS did not support sharing slides, as explained in section 4.4.4. Therefore, we had to change the tasks to account for this unexpected issue; as a result, the broadcast task became a brainstorm task as well, where participants, instead of only ranking items, had to come up with items on their own before discussing them. This issue was confirmed in the interviews, where several participants admitted that they both perceived the two tasks to be similar to a brainstorming exercise. Nevertheless, brainstorm was the task that was indicated by the most participants in the interview to be the most feasible in the Metaverse. For example, participant 1B1Z stated "When I have brainstorm sessions, it's all about chemistry. You feed off each other. So if you were 2-3 people and you start brainstorming and suddenly the ideas start flowing. And that's difficult if you have to compete for getting into the meeting or having a slot [in MS Teams]". This result from the interviews supports the effectiveness scores being slightly higher for the brainstorm conditions.

The second null hypothesis of RQ2 is that there is no relationship between perceived effectiveness and communication media. Our alternative hypothesis is that communication media has an effect on the meeting effectiveness; more specifically, we expected the TNO ICS to perform better in brainstorm meetings, and worse in broadcast meetings. The results of the lmer model and ANOVA showed a difference in effectiveness ratings of 2.6 points between communication media; furthermore, the p-value is lower than our significance level of 0.05, so we can consider our results statistically significant. Then, the estimated mean difference between MS Teams and the TNO ICS in the same meeting types equaled 2.32 and 2.85 points, and both had a p-value lower than the significance level. The Bonferroni correction test confirmed the significance of the results. Lastly, Cohen's d returned a large effect size, thus proving that a difference in communication would be meaningful in the real world. These results reject the null hypothesis; the difference

in effectiveness, both in the brainstorm and broadcast meeting, can be attributed to the different communication media; more specifically, MS Teams is the meeting platform that received the higher grades. As seen in section 6.1.5, there is a trend in participants who rated highly the system usability to also rate highly its effectiveness. This second result supports the results of RQ2.2. We believe that proper usability allows users to effectively reach the meeting goals. In addition, the interviews also made us understand that the TNO ICS lacks several tools for productivity and collaboration, such as a virtual notebook for taking notes, and a whiteboard to sketch or attach post-its. We believe that the effectiveness of the TNO ICS would greatly improve if these tools were added to the system, as the Meta Horizon Workrooms do.

7.3 RQ3: Social Communication

7.3.1 RQ3.1: Spatial Presence

RQ3.1 states "How does the spatial presence of communication media change with different numbers of meeting participants?" and "How does the spatial presence of meeting participants change in different communication media?". For simplicity, we label the first question "RQ3.1.1", and the latter "RQ3.1.2". The first null hypothesis of RQ3.1 is that there is no relationship between spatial presence and communication media. Our alternative hypothesis is that the communication media has an effect on the meeting spatial presence; more specifically, we anticipated both groups with two and three employees to experience a higher level of spatial presence in the TNO ICS rather than in MS Teams. The results of the lmer model and ANOVA showed a difference in effectiveness ratings of 0.07 points between meetings with different communication media; furthermore, the p-value is higher than our significance level of 0.05, so we can consider our results not statistically significant. The variance of the random effects is small; this result indicates that the spatial presence ratings do not vary between subjects, according to their perception of the system. Then, the estimated mean difference between different numbers of participants in the same communication media equaled 0.5 and 0.01 points, and both had a p-value higher than the significance level. The Bonferroni correction test confirmed the non-significance of the results. Lastly, Cohen's d returned a small effect size, thus proving that a difference in the number of participants would not be meaningful in the real world. These results do not reject the null hypothesis; the difference in the perceived spatial presence, both in meetings with two and three participants, cannot be attributed to the difference in communication media. We believe that this result is given by the Metaverse room being small, grey, and anonymous, as commented by several participants in the interviews. For example, participant 1D1Y said "...the gray [wall], I don't know, it didn't feel that much realistic"; participant 1A2Z instead said, having also tried the competitor's Metaverse called Connec2: "After the first time I thought, okay I'm in this room again and there's nothing else you know, but I like Connec2 more because it's a big building you can walk around and you can see things. There's a cool environment around it". Nevertheless, despite the technical limitations, the TNO ICS managed to convey a higher sense of spatial presence, even if not statistically significant. The virtual environment and the Pointcloud representation seem promising technologies for conveying higher levels of spatial presence. We are positive that designing a better and more welcoming virtual environment would make participants feel more in the same place together.

The second null hypothesis of RQ3.1 is that there is no relationship between spatial presence and number of participants. Our alternative hypothesis is that the number of par-

ticipants has an effect on the meeting spatial presence; more specifically, we anticipated both groups in the TNO ICS and MS Teams to experience a higher level of spatial presence in meetings with three people rather than with two people. The results of the lmer model and ANOVA showed a difference in spatial presence ratings of 0.26 points between meetings with different communication media; furthermore, the p-value is higher than our significance level of 0.05, so we can consider our results not statistically significant. The Bonferroni correction test confirmed the non-significance of the results. Lastly, Cohen's d returned a small effect size, thus proving that a difference in communication media would not be meaningful in the real world. These results do not reject the null hypothesis; the difference in the perceived spatial presence, both in meetings in the TNO ICS and MS Teams, cannot be attributed to the different numbers of participants. We believe that this result is given by the fact that the number of social cues shared did not vary enough between the two conditions, thus yielding the same results. We suppose that there might be a significant difference in spatial presence if we were to compare the perceptions of people in a meeting with two participants, against a meeting with a bigger group of people.

7.3.2 RQ3.2: Social Presence

RQ3.2 states "How does the social presence of communication media change with different numbers of meeting participants?" and "How does the social presence of meeting participants change in different communication media?". For simplicity, we label the first question "RQ3.2.1", and the latter "RQ3.2.2". The first null hypothesis of RQ3.2 is that there is no relationship between social presence and communication media. Our alternative hypothesis is that the communication media has an effect on the meeting spatial presence; more specifically, we anticipated both groups with two and three employees to experience a higher level of social presence in the TNO ICS rather than in MS Teams. The results of the lmer model and ANOVA showed a difference in social presence ratings of 0.57 points between meetings with different communication media; furthermore, the p-value is lower than our significance level of 0.05, so we can consider our results statistically significant. The variance of the random effects is small; this result indicates that the social presence ratings do not vary between subjects, according to their perception of the system. Then, the estimated mean difference between different communication media with the same number of participants equaled 0.45 and 0.64 points, and only the latter (related to meetings with two participants) had a p-value lower than the significance level. The Bonferroni correction test confirmed the statistical significance of the results. Lastly, Cohen's d returned a large effect size, thus proving that a difference in communication media would be meaningful in the real world. These results reject the null hypothesis; the difference in the perceived social presence, both in the TNO ICS and MS Teams, can be attributed to the difference in the communication media only for meetings with two participants; however, the results follow our expectations only partially, as MS Teams is the meeting platform that received the higher grades in both cases. We believe that this result is given by the faulty and not yet properly working Pointclouds. Several participants commented on them, saying that they were either misplaced or of low quality. For example, participant 1B1Z stated that "you could actually see a little bit about body language but the image is very coarse, pixelated". Participant 2C2X, who had previously tried the Meta Horizon Workrooms, even admitted that their avatars are "superior" to the Pointcloud technology. In addition, low scores of social presence might be a cause of the HMD covering the eyes of participants, thus making it impossible for participants to look each other in their eyes, or see where they are looking. Participant 1C2X said: "On teams, I can see your whole face, right? I can see your eyes, I can see your emotions a bit better there". We are positive

that improving the quality of the Pointclouds and applying the HMD-removal technique, already in development at TNO, could greatly increase the feeling of social presence in the TNO ICS.

The second null hypothesis of RQ3.2 is that there is no relationship between social presence and number of participants. Our alternative hypothesis is that the number of participants has an effect on the meeting social presence; more specifically, we anticipate both groups in the TNO ICS and MS Teams to experience a higher level of spatial presence in meetings with three people rather than with two people. The results of the lmer model and ANOVA showed a difference in social presence ratings of 0.003 points between meetings with different numbers of participants; furthermore, the p-value is higher than our significance level of 0.05, so we can consider our results not statistically significant. Then, the estimated mean difference between meetings with different numbers of participants with the same communication media equaled 0.11 and 0.07 points, and none had a p-value lower than the significance level. The Bonferroni correction test confirmed the non-significance of the results. Lastly, Cohen's d returned a small effect size, thus proving that a difference in the number of participants would not be meaningful in the real world. These results do not reject the null hypothesis; the difference in the perceived social presence, both in meetings in the TNO ICS and MS Teams, cannot be attributed to the different numbers of participants. We believe that this result is given by the fact that the number of social cues did not vary enough between the two conditions, thus yielding the same results. We suppose that there might be a significant difference in spatial presence if we were to compare the perceptions of people in a meeting with two participants, against a meeting with a bigger group of people.

7.4 RQ4: Extraneous Variables

RQ4 is written as "How do extraneous variables and further interaction effects influence the subjective QoE factors?". Given the exploratory nature of RQ4, we did not assign a hypothesis to it, but we try to find as many insights as possible for future research. In this section, we discuss the graphical representations reported in section 6.1.5 that present interesting patterns and anomalies that we cannot explain, and should be further tested in future studies. Figure 6.8c, shows a considerable difference in usability scores expected by groups of two and three participants. Figure 6.8d shows an unexpected pattern between the level of VR experience of participants and the usability scores they gave to the communication media, as users who reported having "a little" experience rated much worse the usability of the systems. Figure 6.9b shows that the predicted values of usability are the highest when participants rate the spatial presence perceived as the worst possible. Similarly, figure 6.11a shows that the predicted values of effectiveness are suspiciously higher when participants rate the spatial presence perceived as worst possible. figure 6.13a shows that the predicted values of spatial presence peak in the range of 36-45 of the usability scores.

From the other graphs, we can see that there are no particular outliers or strange patterns, and we might suppose that there is a positive correlation between the two variables depicted, as their values change together and in the same direction[1]. However, correlation does not imply causation, and it is up to a future study to verify whether effectiveness and usability are actually positively correlated.

[1]https://www.investopedia.com/ask/answers/040915/what-difference-between-positive-correlation-and-inverse-correlation.asp

Regarding the extraneous variables, only the difference in the level of VR experience of participants could be statistically significant in explaining the variations of the usability scores. All the other extraneous variables did not appear to play a role in the answers given by participants in the questionnaires. From the five extraneous variables, only the "meeting progression and order" was mentioned by the participants in the interview sessions. Out of the 32 participants, ten of them stated that they had a better feeling during the second experience in the Metaverse. However, we did not find in the exploratory analysis any meaningful change in the data between one Metaverse experience and the other. If, on one side, this result might seem to be underwhelming, as it appears that users did not get used enough to the new platform and their general QoE did not improve, this data also shows that we avoided the "first timer effect" and "habituation effect" thanks to the counterbalancing given by the latin square design.

Regarding the interaction effects of the variables, the differences in the usability scores are best described by the ratings of effectiveness, social and spatial presence, and VR experience; the differences in the effectiveness scores are best described by the ratings of usability; the differences in the spatial presence scores are best described by the ratings of social presence; the differences in the social presence scores are best described by the ratings of effectiveness. These relationships should be further investigated in future studies.

7.5 RQ5: Other Explanatory Insights

We discuss in this section the results of the thematic analysis conducted to answer the research question "Which other subjective factors influence the meeting QoE?" from the qualitative data collected from the interviews. We did not employ any independent or dependent variable for this part of the study, as it is an exploratory analysis. Our goal is instead to find factors that negatively or positively affected the meeting experience, which could not be observed through a questionnaire.

During the interviews, not many factors emerged that were not related to one of the three QoE components; we could argue that it is because the three factors we chose initially describe sufficiently the meeting experience, without the need to involve other factors. On the other hand, we suppose that the main cause is that the initial set of questions focused on usability, effectiveness, and social communication; this might have biased the participants, who also commented almost exclusively on the three factors. However, we still collected insights from the interview sessions; although we did not discover any other subjective factor that might influence the meeting QoE, we understood better the reason why usability, effectiveness, and social communication influence (or not) the meeting experience. The qualitative data from the interviews provided another point of view to help us understand the results of the quantitative data from the questionnaires.

Since technical issues were the usability problem mentioned the most, it is a high priority to have a reliable and robust system, starting from the faulty Pointcloud technology; other improvements should cover the slides not working properly, and the frequent audio issues. Then, the TNO ICS should be more fast and intuitive to set up; participants complained because they needed to access the Metaverse with the computer (to share their Pointcloud image, captured by the depth camera), and with the HMD, to see the other meeting participants. The lengthy setup times also influenced the meeting's effectiveness according to some participants, as they commented that had less time to work on the task. The technology itself is also blamed, as many participants complained that the HMD was too heavy. Furthermore, the majority of employees who wear glasses complained that it was not feasible to keep them, thus making the video blurry and out of focus. On a posi-

tive note, participants felt more comfortable in the virtual environment during the second Metaverse session, as mentioned before. We initially supposed that the more the participants would have meetings in the Metaverse, the more they would adapt and appreciate the system, thus rating the QoE higher in the later meetings; however, this phenomenon was not supported by the quantitative data, as the expected mean score of the SUS questionnaire for the second meetings is only 3.29 points higher than the first, with a p-value $> .05$; instead, we now suppose that participants might also get used to the novelty of the application, thus losing the first interest and excitement they initially had, leading to rate the system lower. When asked if they would have participated in a third Metaverse meeting, many users answered negatively, or at least not in the current system; as participant 1C1Z said, "I feel that I'm now at the point that I would ask for improvements to have the same enthusiasm as I had before the second time [in the Metaverse]".

The instructions are yet another element that weighs in when it comes to setup issues, as none of the participants could independently start the system even for the second Metaverse meeting. Written instructions would be a simple answer, but users disliked them and preferred to call for help instead of carefully reading the guide. As a result, we opted for always providing assistance via MS Teams for the setup of the ICS. This solution was feasible during an experiment, but it might not be possible to do so in a real-life setting, as it would require someone to always be available for help. As a user suggested, employees could instead be instructed to watch a video tutorial on how to start the system, which would be easier to follow than written instructions and would not require assistance in most cases.

Then, the theme of effectiveness covered mostly requests and ideas for improvements from participants. For example, users would like to have the possibility to take notes, sketch or attach post-its on a whiteboard, and share files on their laptop screens. Although these comments were shared after brainstorm sessions only (considering that broadcast sessions felt like brainstorm), and not from all kinds of meetings possible at a company, we still believe that these features could be used in other meeting types as well, as they serve for a general purpose, and not something specific to brainstorm meetings. However, we need to keep in mind that participants might like the idea of having these tools in the Metaverse, but it is not guaranteed that they will use them or appreciate them once implemented in the system; further study on this matter should be completed before developing these functionalities. Having a similar amount of comments indicating that the Metaverse is both more distracting and less distracting than a Teams meeting is another indicator that participants' opinions are not always coherent and that it might not be possible to design a system that all participants like entirely.

Lastly, the factor of social communication received the most comments. First of all, the majority of participants complained that the HMD was covering the eyes of their meeting partner(s), which had several implications. Some participants, for example, 2B1X felt "bothered in a weird manner" and "distracted" by it; participant 1C1Z, among others, felt that the nonverbal communication was worse, stating "I did not see such recognizable nonverbal communication. [The HMD] hindered my way of communicating". The low-quality Pointcloud received also many negative comments; many participants, such as 1C1Z, felt that the faulty Pointcloud hindered the communication, making it "not efficient enough"; other participants commented similarly to 2B2X that the low-quality Pointcloud lowers the feeling of immersion, saying "your partner is kind of floating at a weird angle and it doesn't look very real. It sort of takes away from the immersion of the whole thing". However, social communication received more positive comments than the other QoE factors. For example, some participants instead appreciated the Pointcloud; for example, 2C1Y said:

"it really felt like I was in a different room and [participant 2C1X] was in front of me; sometimes the image was glitching and whatever, but I think the rendering was pretty nice. I really felt in another meeting room with [participant 2C1X] in front of me". Other participants felt more involved than in a Tems meeting because they could see each other; for example, participant 1D1Z said: "[in the Metaverse I] felt more involved ... being able to look around you and seeing the people, especially with this [brainstorm] task".

The chapter on the discussion is now concluded. Table 7.1 summarizes the results of each RQ. The tables in chapter D of the Appendix, instead, expand table 7.1 to provide a holistic overview of each research question by including their most relevant data. In this chapter, we showed that the results rejected only the null hypothesis of RQ1, RQ2.2, and RQ3.3, and did not reject the null hypothesis of RQ2.1, RQ3.1, RQ3.2, and RQ3.4, for the following reasons: for RQ2.1 the tasks were too similar, for RQ3.1 the virtual environment was not of sufficient quality, and for RQ3.2 and RQ3.4 a group of three meeting participants is not large enough. In the next chapter, we will discuss the limitations that prevented us from conducting an experiment of better quality and our suggestions for future works on the topics of ICS evaluation.

RQ	Null Hypothesis	Result	Null Hypothesis Rejected?
RQ1	There is no difference in usability scores between the TNO ICS and MS Teams	Usability of MS Teams > Usability of TNO ICS	Yes
RQ2.1	There is no relationship between perceived effectiveness and communication media.	TNO ICS and MS Teams: Eff. Brainstorm = Eff. Broadcast	No
RQ2.2	There is no relationship between perceived effectiveness and meeting type	Brainstorm and broadcast: Eff MS Teams > Eff. TNO ICS	Yes
RQ3.1.1	We will find no effect of the number of participants on the spatial presence section of the H-MSC-Q scores	3 and 2 participants: Spa.Pres. of ICS = Spa.Pres. of MS Teams	No
RQ3.1.2	We will find no effect of communication media on the spatial presence section of the H-MSC-Q scores	TNO ICS and MS Teams: Spa.Pres. with 3 = Spa.Pres. with 2	No
RQ3.2.1	We will find no effect of the number of participants on the social presence section of the H-MSC-Q scores	3 and 2 participants: Soc.Pres. of MS Teams > Soc.Pres. of TNO	Yes
RQ3.2.2	We will find no effect of communication media on the social presence section of the H-MSC-Q scores	TNO ICS and MS Teams: Soc.Pres. with 3 = Soc.Pres. with 2	No

Table 7.1: The results summary.

Chapter 8

Future Work and Limitations

First, repeating the study with a better, improved system would surely result in better scores for all three subjective factors, as the difference in communication media was statistically significant in each case. We also think that the majority of the feedback and comments we got on the system's usability were actually about the hardware issues, not the system's usability itself, even though the latter undoubtedly had an impact on the former. Similarly, the experiment should be repeated, but instead of using Microsoft Teams, two versions of the TNO ICS should be compared: one with the changes described in this thesis, and one without. This comparison could help the team determine whether the suggested adjustments were successful. The most urgent and relevant improvements to be made are discussed in the sections related to the interview results.

Then, several limitations of this study are related to time issues, as we did not have as much time as we would have liked to plan the experiment, coordinate the meetings, run the interviews, and analyze the data. However, we still consider the results of our study to be valid and reliable.

If we had more time, we would have also tested more models with the Akaike Information Criterion, as the array of combinations possible, when using the three independent variables and five extraneous variables, is quite large and time-consuming to analyze in its entirety. We believe that spending more time on it would have led us to find another model that would fit better the data; nevertheless, we are positive that the models we used are sufficiently good for our analysis.

Then, a follow-up study should make sure that the tasks can be carried out as expected, and that participants perceive them as different. For example, we expect to receive different results from this experiment if the slides had worked. Additionally, having participants behave and communicate differently in the tasks will help in understanding whether the Metaverse is truly suited for a task better than another or not. To increase the validity of the findings, we would also recommend repeating the experiment by having users participate in proper work meetings. This is because several participants reported that the tasks we suggested did not give them the impression that they were in a formal business meeting but rather something more casual and relaxed, thus deviating from the actual use intended for this platform.

It is also recommended to test the system with more than three participants at a time, to assess the extent to which participants are positively drawn by the social presence aspect of the Metaverse, to engage more during meetings compared to their experiences with MS Teams. We also suggest repeating the experiment with a better virtual room, as we expect it to provide participants with a higher sense of spatial presence.

Another possibility for future works would be to implement objective data in the study;

for example, we suggest counting the ideas generated during a brainstorm meeting; if the experiment implements the "lost at sea" and "desert survival" tasks, it is recommended to consult the provided interpretation by the task authors to determine the correct sequence of the items of these tasks, to verify whether certain meeting conditions allow users to perform better. Furthermore, the study of [3] showed the feasibility of measuring aspects such as turn duration, turn frequency, and gaze to objectively analyze the conversations. Collecting objective data would also avoid relying solely on the users' perception, as we believe that the participants' answers might not be always correct or unbiased. Furthermore, we noticed that it is hard to measure those instances where users had positive and negative experiences together in the same meeting. Since questionnaires do not allow for different measures over time, we believe that participants may remember a small technical issue that broke the feeling of presence more vividly than the rest of the meeting experience, which may have been more than satisfactory, leading them to rate the experience lower than they should have. However, we argue that gathering only objective data may not be the optimal approach when assessing the feasibility of deployment based on the user's QoE; employees may favor one system over another because they believe it to be more usable and efficient, even though the objective data may indicate the opposite.

We also suggest repeating the experiment with more participants; we noticed that not all levels of the extraneous variables had a meaningful number of participants associated with them. For example, only two participants had meetings in the Leiden location; then, users were mostly males under the age of 35; all participants defined themselves as confident or extremely confident in using technological devices; most employees had little to no experience in VR meetings. We expect that having a more heterogeneous and diverse population for the experiment would allow us to have more reliable answers on whether the extraneous variables have a strong influence on the three QoE factors. In fact, we suppose that many effects exist in the population, but our tests did not detect them because the sample size might have been too small to detect the effect. Only the variable "knowing each other" had an even distribution of participants across its levels, but we discovered that it is not a meaningful parameter in the QoE factors.

On the matter of RQ4, and RQ5, we only considered them for an exploratory analysis due to the lack of time. We hope that the results of those chapters, and the related recommendations, will be the first step of a more in-depth, follow-up study. Focusing on RQ4 in particular, we suggest in a future study to study the extraneous variables and interaction effect only for the meetings held in the TNO ICS, as we think it might yield more interesting results.

Regarding the interview sessions, we did not take into consideration the population demographic and extraneous variables when coding the interviews. For example, in a future study, we suggest comparing whether the number of negative comments about the TNO ICS is higher in the second meeting, following a decrease in excitement and novelty, or if the number is lower in the second meeting, as the participants get used to the system and its features.

We recommend selecting participants for a future study based on their demographics, to have a sufficient number of users to accurately reflect, for example, each range of age and technology experience.

We believe that following these suggestions will bring future studies to reject those null hypotheses that were not rejected in our experiment.

Regarding the communication media used, even though we discussed in-depth the State of the Art as well, i.e. Meta Horizon Workrooms, we did not think it was feasible to compare and evaluate it due to a lack of time. We leave the comparison between Workrooms and

the TNO ICS for future research.

The chapter on future work suggestions and limitations of this study is now concluded. In summary, we realize that the limited amount of time, the technological immaturity of the TNO ICS, and the sample size not big enough were the main limitations of this study, which we believe prevented the null hypothesis from being rejected. We hope that this study will be followed by another research that follows our suggestions to find better results on the topic of evaluating ICSs for business meetings. The next chapter is the conclusions of this study, where we summarize the main findings from the research.

Chapter 9

Conclusion

This chapter draws the conclusions of the experiment, which is one-of-a-kind in the field of VR platforms for business meetings, as no study was conducted with this many participants. In this chapter, we will report the answers to our RQs, the key findings, and the thesis contributions to the scientific field.

RQ1 is "How does usability change in different communication media?". As we expected, the answer to RQ1 is that usability is significantly higher in MS Teams, and lower in the TNO ICS. The TNO ICS did not address the usability issues of non-immersive platforms, and did not receive the same positive responses as the ICS from the study of Gunkel et al. [19] that we were expecting.

RQ2 states "How does the perceived effectiveness of meeting types change in different communication media?" and "How does the perceived effectiveness of communication media change in different meeting types?". Contrary to what we expected, the answer to RQ2.1 is that the perceived effectiveness of communication media does not change in different meeting types; at most, there is a tendency in the data where brainstorm meetings are perceived as more effective than broadcast meetings in the TNO ICS and MS Teams. The answer to RQ2.2 is that the perceived effectiveness of meeting types changes in different communication media; although we expected the TNO ICS to perform better in brainstorm meetings, it is actually MS Teams that received the higher grades. Our expectation of the TNO ICS performing worse in broadcast meetings is instead confirmed. The TNO ICS did not address the issues of lower engagement and effectiveness provided by non-immersive platforms and did not receive better ratings for brainstorm meetings as we expected from the studies of Standaert et al. [39, 38, 37].

RQ3.1 states "How does the spatial presence of communication media change with different numbers of meeting participants?" and "How does the spatial presence of meeting participants change in different communication media?" Contrary to what we expected, the answer to RQ3.1.1 is that the perceived spatial presence of communication media does not change with different communication media; at most, there is a tendency in the data where meetings in the TNO ICS provide more spatial presence than meetings in MS Teams, in both meetings with two and three participants. Contrary to what we expected, the answer to RQ3.1.2 is that the perceived spatial presence of communication media does not change with different number of participants; at most, there is a tendency in the data where meetings with three participants provide more spatial presence than meetings with two participants, in both meetings in the TNO ICS and MS Teams.

RQ3.2 states "How does the social presence of communication media change with different numbers of meeting participants?" and "How does the social presence of meeting participants change in different communication media?" The answer to RQ3.2.1 is that

only the perceived social presence of meetings with two participants changes depending on the communication media; more specifically, meetings in MS Teams provide significantly more social presence than meetings in the TNO ICS with two participants, and a similar level of social presence in meeting with three participants. These results do not match our expectations, as we expected the TNO ICS to provide a better feeling of social presence. Contrary to our expectations, the answer to RQ3.2.2 is that the perceived social presence of communication media does not change with different numbers of participants; at most, there is a tendency in the data where meetings with three participants provide more social presence than meetings with two participants, in meetings in the TNO ICS, but not in MS Teams. The TNO ICS does not address the issue of non-immersive platforms related to providing fewer social cues and worse non-verbal communication; its immersive space also failed to provide a significantly higher quality of social communication, as we expected from the studies discussed in 2.2.3.

We wrote RQ4 as "How do extraneous variables and further interaction effects influence the subjective QoE factors?". The response to RQ4 is that participants' level of VR experience appears to influence the usability scores in a way that those with little experience rate it much worse than the rest. The other extraneous variables do not seem to influence the meeting QoE. In addition, the usability scores are best explained by communication media, meeting type, number of participants, and the level of VR experience; the effectiveness scores are best explained by the usability scores; the spatial presence scores are best explained by the social presence scores; the social presence scores are best explained by the effectiveness scores.

RQ5 states "Which other subjective factors influence the meeting QoE?". From the interview session, we did not discover any other major subjective factor, but we instead collected enough answers to give an explanation for the unexpectedly lower ratings of usability, effectiveness, and social communication. The lower ratings for the TNO ICS usability are related to technical issues and long and complex setup times; the lower ratings for the effectiveness are related to the missing tools and functionalities; the lower ratings for the effectiveness are related to the HMD covering the participants' face, and the low-quality Pointcloud.

The overarching research question states "how does the perceived QoE of meetings change for users in different conditions?". The answer is that the meeting QoE changes mainly based on the communication media used. Our study did not find any other independent variable that had a statistically significant effect on the three subjective factors.

Regarding the initial problem statement, we supposed that ICSs might be a valid solution to the drawbacks of MS Teams, such as technical issues, less effectiveness, lower engagement, and struggling to read social cues. However, despite what the literature on immersive systems suggested, using the TNO ICS to substitute traditional meeting platforms does not appear to be a viable alternative in the current state of technology.

The high-level goal of this thesis is to understand the feasibility of deploying the TNO ICS for business meetings, by analyzing the meeting QoE. As of today, ICSs provide a better meeting QoE thanks to their immersive space, according to the literature review; however, our experiment results indicate that the TNO ICS is not yet suitable for hosting meetings due to the low QoE it provides; instead, as participants suggested, having meetings in the Metaverse can be considered a fun and new experience to try. Although the SocialXR team greatly improved the TNO ICS following Singh et al. [35] evaluation and their suggestions, both this thesis and their study concluded that the ICS is not yet ready for deployment. Additionally, the findings of our study differed from those of the scientific literature cited in the problem statement: participants in our study did not perceive any critical issue during

meetings in MS Teams, despite the fact that our research on non-immersive platforms came to the opposite conclusion. Nevertheless, we see the potential in this technology. Even with its technical limitations, there is a tendency in the data where the TNO ICS provided its users slightly more spatial presence than MS Teams. We believe that thanks to this study, which highlights the areas that needs to be improved, and the technological advancement that will take place in the next few years, ICSs will be used in the future for business meetings, starting from brainstorm meeting in a large group of people.

Figure 9.1 graphically depicts the results of this experiment. The green arrows indicate a significant difference found in the results.

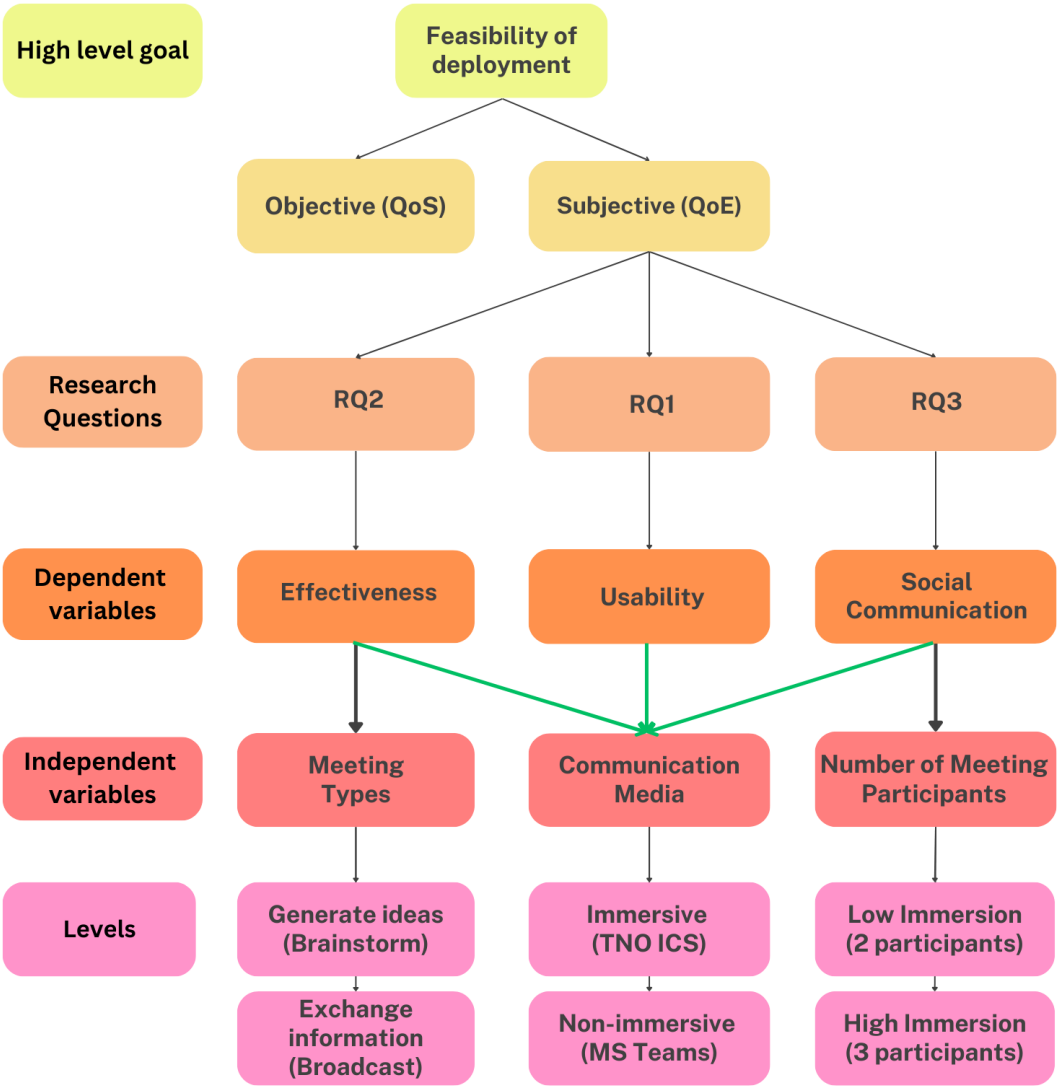


Figure 9.1: The results of the experiment.

Table 9.1 shows the statistically significant effects that dependent and independent variables have on each other. In our case, MS Teams is the communication media that presented a statistically significant difference from the TNO ICS of $<.001$ (indicated with "***"), in Social Presence provided, Perceived Effectiveness, and Usability. The four cells with a "-" indicate a p-value $> .05$, which does not imply a statistical significance. The remaining five empty cells represent those combinations that, although tested, do not have a p-value, because they were part of an exploratory analysis and we did not define a

	Number of participants	Meeting type	Communication media
Social presence	-		MS Teams (***)
Spatial presence	-		-
Effectiveness		-	MS Teams (***)
Usability			MS Teams (***)

Table 9.1: A summary of which levels for each independent variable received higher ratings.

hypothesis for them.

Figure 9.2 graphically depicts the overview of the literature review and the experiment.

This study is an important contribution to the scientific field of immersive communication because, to our knowledge, there is no other study that employed such a number of participants in a real-life setting, and not in a controlled environment.

We also want to highlight that the IEEE MetroXRAINE 2023 conference published the paper, written by this thesis author, covering the literature review presented in chapter 2. Furthermore, a second paper was accepted for publication at the EuroXR 2023 conference, covering the experiment and its results.

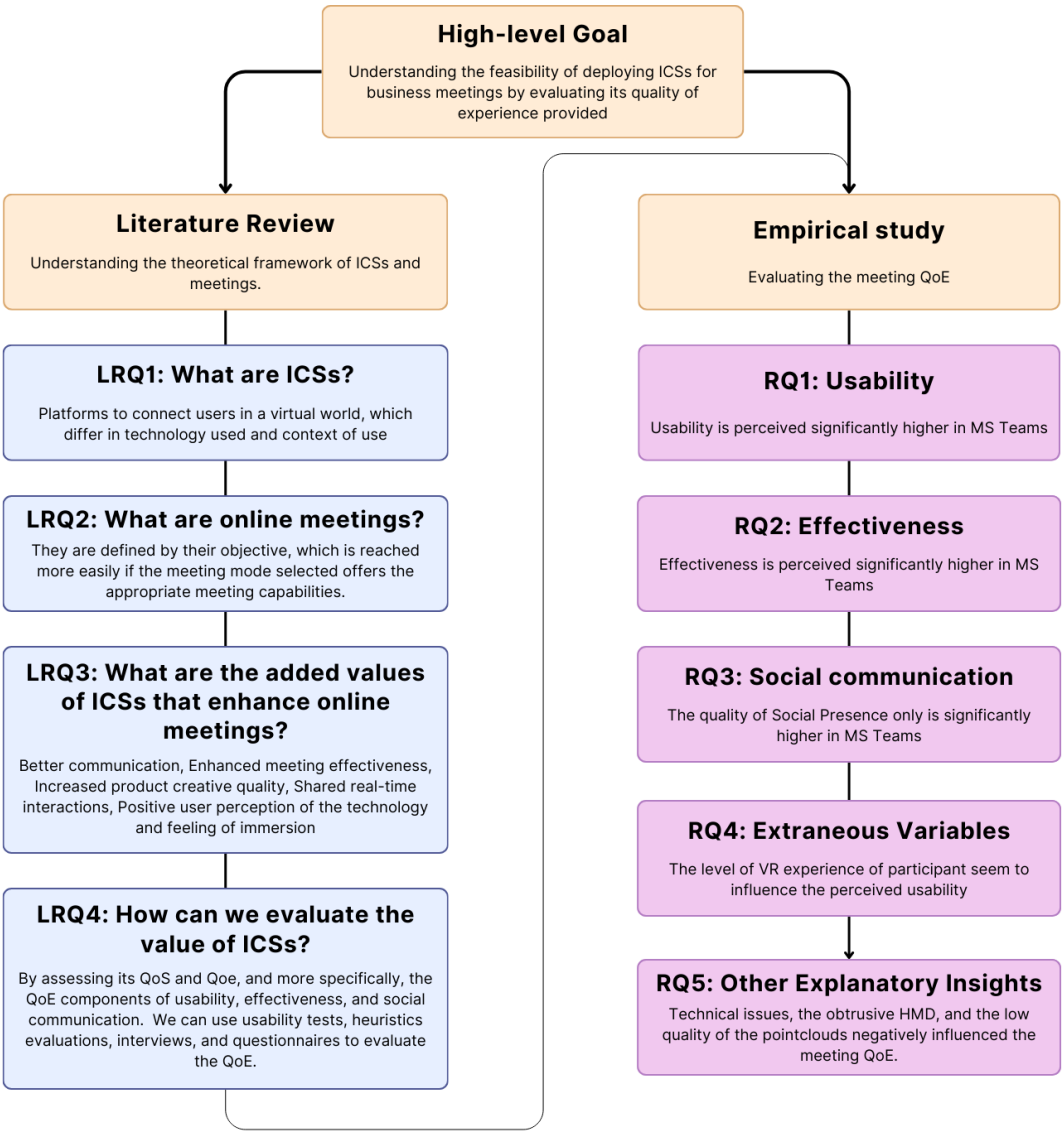


Figure 9.2: The overview of the literature review and the empirical study.

Bibliography

[1] Meta horizon workrooms: Virtual workroom: Work with meta. URL: https://www.meta.com/nl/en/work/workrooms/.

[2] Sep 2015. URL: https://www.iso.org/standard/45481.html.

[3] Ahsan Abdullah, Jan Kolkmeier, Vivian Lo, and Michael Neff. Videoconference and embodied vr: Communication patterns across task and medium. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2), oct 2021. doi:10.1145/3479597.

[4] Michelle Aebersold. Simulation-based learning: No longer a novelty in undergraduate education. *OJIN: The Online Journal of Issues in Nursing*, 23(2), 2018. doi:10.3912/ojin.vol23no02ppt39.

[5] Wadee Alhalabi. Virtual reality systems enhance students' achievements in engineering education. *Behaviour & Information Technology*, 35(11):919–925, 2016. arXiv:https://doi.org/10.1080/0144929X.2016.1212931, doi:10.1080/0144929X.2016.1212931.

[6] Mohammed Alreshoodi and John Woods. Survey on qoecorrelation models formulti-media services. *International Journal of Distributed and Parallel systems*, 4(3):53–72, 2013. doi:10.5121/ijdps.2013.4305.

[7] Aaron Bangor, Philip Kortum, and James Miller. Determining what individual sus scores mean: Adding an adjective rating scale. *Journal of usability studies*, 4(3):114–123, 2009.

[8] J.M. Christian Bastien. Usability testing: A review of some methodological and technical aspects of the method. *International Journal of Medical Informatics*, 79(4), 2010. doi:10.1016/j.ijmedinf.2008.12.004.

[9] Kathy Baxter, Catherine Courage, and Kelly Caine. *Understanding your users: A practical guide to user research methods*. Morgan Kaufmann, Elsevier Ltd, 2015.

[10] Pritha Bhandari. Extraneous variables | examples, types controls. *Scribbr*, Jun 2023. URL: https://www.scribbr.com/methodology/extraneous-variables/.

[11] Doug A. Bowman, Joseph L. Gabbard, and Deborah Hix. A survey of usability evaluation in virtual environments: Classification and comparison of methods. *Presence: Teleoperators and Virtual Environments*, 11(4):404–424, 2002. doi:10.1162/105474602760204309.

[12] John Brooke. Sus: a retrospective. *Journal of usability studies*, 8(2):29–40, 2013.

[13] Melanie S. Brucks and Jonathan Levav. Virtual communication curbs creative idea generation. *Nature*, 605(7908):108–112, 2022. doi:10.1038/s41586-022-04643-y.

[14] Fred D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319, 1989. doi:10.2307/249008.

[15] G. Fauville, M. Luo, A.C.M. Queiroz, J.N. Bailenson, and J. Hancock. Zoom exhaustion amp; fatigue scale. *Computers in Human Behavior Reports*, 4:100119, 2021. doi:10.1016/j.chbr.2021.100119.

[16] Adrian Fernandez, Emilio Insfran, and Silvia Abrahão. Usability evaluation methods for the web: A systematic mapping study. *Information and Software Technology*, 53(8):789–817, 2011. doi:10.1016/j.infsof.2011.02.007.

[17] J.L. Gabbard, D. Hix, and J.E. Swan. User-centered design and evaluation of virtual environments. *IEEE Computer Graphics and Applications*, 19(6):51–59, 1999. doi:10.1109/38.799740.

[18] Simon N.B Gunkel, Marleen D.W. Dohmen, Hans Stokking, and Omar Niamut. 360-degree photo-realistic vr conferencing. *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 2019. doi:10.1109/vr.2019.8797971.

[19] Simon N.B. Gunkel, Martin Prins, Hans Stokking, and Omar Niamut. Social vr platform. *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, 2017. doi:10.1145/3084289.3089914.

[20] Scott G Isaksen et al. *A review of brainstorming research: Six critical issues for inquiry*. Creative Research Unit, Creative Problem Solving Group-Buffalo Buffalo, NY, 1998.

[21] ISO. Iso 9241-11:1998 ergonomic requirements for office work with visual display terminals (vdts) — part 11: Guidance on usability, Apr 2018. URL: https://www.iso.org/standard/16883.html.

[22] Katherine A. Karl, Joy V. Peluchette, and Navid Aghakhani. Virtual work meetings during the covid-19 pandemic: The good, bad, and ugly. *Small Group Research*, 53(3):343–365, 2021. doi:10.1177/10464964211015286.

[23] Anastasia Kuzminykh and Sean Rintel. Low engagement as a deliberate practice of remote participants in video meetings. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. doi:10.1145/3334480.3383080.

[24] James Clayton Lafferty and Alonzo William Pond. *Desert survival situation*. Human Synergistics International, 1987.

[25] Sergi Fernandez Langa, Mario Montagud, Gianluca Cernigliaro, and David Rincon Rivera. Multiparty holomeetings: Toward a new era of low-cost volumetric holographic meetings in virtual reality. *IEEE Access*, 10:81856–81876, Aug 2022. doi:10.1109/access.2022.3196285.

[26] Desmond J. Leach, Steven G. Rogelberg, Peter B. Warr, and Jennifer L. Burnfield. Perceived meeting effectiveness: The role of design characteristics. *Journal of Business and Psychology*, 24(1):65–76, 2009. doi:10.1007/s10869-009-9092-6.

[27] Glenn Littlepage, William Robison, and Kelly Reddington. Effects of task experience and group experience on group performance, member ability, and recognition of expertise. *Organizational Behavior and Human Decision Processes*, 69(2):133–147, 1997. doi:10.1006/obhd.1997.2677.

[28] Divine Maloney, Guo Freeman, and Donghee Yvette Wohn. "talking without a voice". *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25, 2020. doi:10.1145/3415246.

[29] Paul Milgram and Fumio Kishino. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329, 1994.

[30] Paul M Nemiroff and William A Pasmore. Lost at sea: A consensus-seeking task. *The Pfeiffer book of successful team-building tools: Best of the annuals*, pages 165–172, 2001.

[31] Hadar Nesher Shoshan and Wilken Wehrt. Understanding "zoom fatigue": A mixed-method approach. *Applied Psychology*, 71(3):827–852, 2021. doi:10.1111/apps.12360.

[32] Catherine S. Oh, Jeremy N. Bailenson, and Gregory F. Welch. A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5, 2018. doi:10.3389/frobt.2018.00114.

[33] Freddy Paz and Jose Antonio Pow-Sang. Current trends in usability evaluation methods: A systematic review. *2014 7th International Conference on Advanced Software Engineering and Its Applications*, 2014. doi:10.1109/asea.2014.10.

[34] Amir H Sadeghi, Ali R Wahadat, Adem Dereci, Ricardo P Budde, Wilco Tanis, Jolien W Roos-Hesselink, Hanneke Takkenberg, Yannick J Taverne, Edris A Mahtab, Ad J Bogers, and et al. Remote multidisciplinary heart team meetings in immersive virtual reality: A first experience during the covid-19 pandemic. *BMJ Innovations*, 7(2):311–315, 2021. doi:10.1136/bmjinnov-2021-000662.

[35] Simardeep Singh, Sylvie Dijkstra-Soudarissanane, and Simon Gunkel. Engagement and quality of experience in remote business meetings: A social vr study. In *Proceedings of the 1st Workshop on Interactive EXtended Reality*, IXR '22, page 77–82, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3552483.3556457.

[36] Mel Slater. A note on presence terminology. *Presence connect*, 3(3):1–5, 2003.

[37] Willem Standaert, Steve Muylle, and Amit Basu. An empirical study of the effectiveness of telepresence as a business meeting mode. *Information Technology and Management*, 17(4):323–339, 2015. doi:10.1007/s10799-015-0221-9.

[38] Willem Standaert, Steve Muylle, and Amit Basu. How shall we meet? understanding the importance of meeting mode capabilities for different meeting objectives. *Information amp; Management*, 58(1):103393, 2021. doi:10.1016/j.im.2020.103393.

[39] Willem Standaert, Steve Muylle, and Amit Basu. Business meetings in a postpandemic world: When and how to meet virtually. *Business Horizons*, 65(3):267–275, 2022. doi:10.1016/j.bushor.2021.02.047.

[40] Alexander Toet, Tina Mioch, Simon Gunkel, Omar Niamut, and Jan Erp. Towards a multiscale qoe assessment of mediated social communication. *Quality and User Experience*, 7:article 4, 06 2022. doi:10.1007/s41233-022-00051-2.

[41] Thomas Tullis and Jacqueline Stetson. A comparison of questionnaires for assessing website usability. 06 2006.

[42] Xiaozhe Yang, Lin Lin, Pei-Yu Cheng, Xue Yang, Youqun Ren, and Yueh-Min Huang. Examining creativity through a virtual reality support system. *Educational Technology Research and Development*, 66(5):1231–1254, 2018. doi:10.1007/s11423-018-9604-z.

[43] Nicole Yankelovich, William Walker, Patricia Roberts, Mike Wessler, Jonathan Kaplan, and Joe Provino. Meeting central. *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, 2004. doi:10.1145/1031607.1031678.

[44] Ryan Yung, Truc H. Le, Brent Moyle, and Charles Arcodia. Towards a typology of virtual events. *Tourism Management*, 92:104560, 2022. doi:10.1016/j.tourman.2022.104560.

Appendices

During the preparation of this work, I used Grammarly, ChatGPT, and Google Bard to improve the style and readability of a few sentences, and edited the LaTeX code for tables. After using this tool/service, we thoroughly reviewed and edited the content as needed, taking full responsibility for the final outcome.

Appendix A

Description and Items of Tasks T5 and T6

A.1 Task T5: Lost at Sea

A.1.1 Task Description

You have chartered a yacht with three friends, for the holiday trip of a lifetime across the Atlantic Ocean. Because none of you have any previous sailing experience, you have hired an experienced skipper and two-person crew. Unfortunately in mid Atlantic a fierce fire breaks out in the ships galley and the skipper and crew have been lost whilst trying to fight the blaze. Much of the yacht is destroyed and is slowly sinking. Your location is unclear because vital navigational and radio equipment have been damaged in the fire. Your best estimate is that you are many hundreds of miles from the nearest landfall. You and your friends have managed to save 15 items, undamaged and intact after the fire. In addition, you have salvaged a four man rubber life craft and a box of matches. Your task is to rank the 15 items in terms of their importance for you, as you wait to be rescued. Place the number 1 by the most important item, the number 2 by the second most important and so forth until you have ranked all 15 items.

A.1.2 Task Items

A sextant
A shaving mirror
A quantity of mosquito netting
A 25 liter container of water
A case of army rations
Maps of the Atlantic Ocean
A floating seat cushion
A 10 liter can of oil/petrol mixture
A small transistor radio
20 square feet of opaque plastic sheeting
A can of shark repellent
One bottle of 160 proof rum
15 feet of nylon rope
2 boxes of chocolate bars
An ocean fishing kit and pole

A.2 Task T6

A.2.1 Task Description

It is approximately 10:00 am in mid-July and you have just crash landed in the Atacama Desert in South America. Your light twin-engined plane containing the bodies of the pilot and co-pilot has completely burned out with only the frame remaining. None of you have been injured. The pilot was unable to notify anyone of your position before the crash. However, he had indicated before impact that you were 50 miles from a mining camp, which is the nearest known settlement, and approximately 65 miles off the course that was filed in your Flight Plan. The immediate area is quite flat, except for occasional cacti, and appears to be rather barren. The last weather report indicated that the temperature would reach 110 F today, which means that the temperature at ground level will be 130 F. You are dressed in lightweight clothing-short-sleeved shirts, pants, socks, and street shoes. Everyone has a handkerchief and collectively, you have 3 packs of cigarettes and a ballpoint pen. Before your plane caught fire, your group was able to salvage the 15 items listed on the "Salvaged Items" page. Your task is to rank these items according to their importance to your survival, starting with a "1" for the most important, to a "15" for the least important

A.2.2 Task Items

Torch with 4 battery-cells
Folding knife
Air map of the area
Plastic raincoat (large size)
Magnetic compass
First-aid kit
45 calibre pistol (loaded)
Parachute (red and white)
Bottle of 1000 salt tablets
2 litres of water per person
A book entitled 'Desert Animals That Can Be Eaten'
Sunglasses (for everyone)
2 litres of 180 proof liquor
Overcoat (for everyone)
A cosmetic mirror

Appendix B

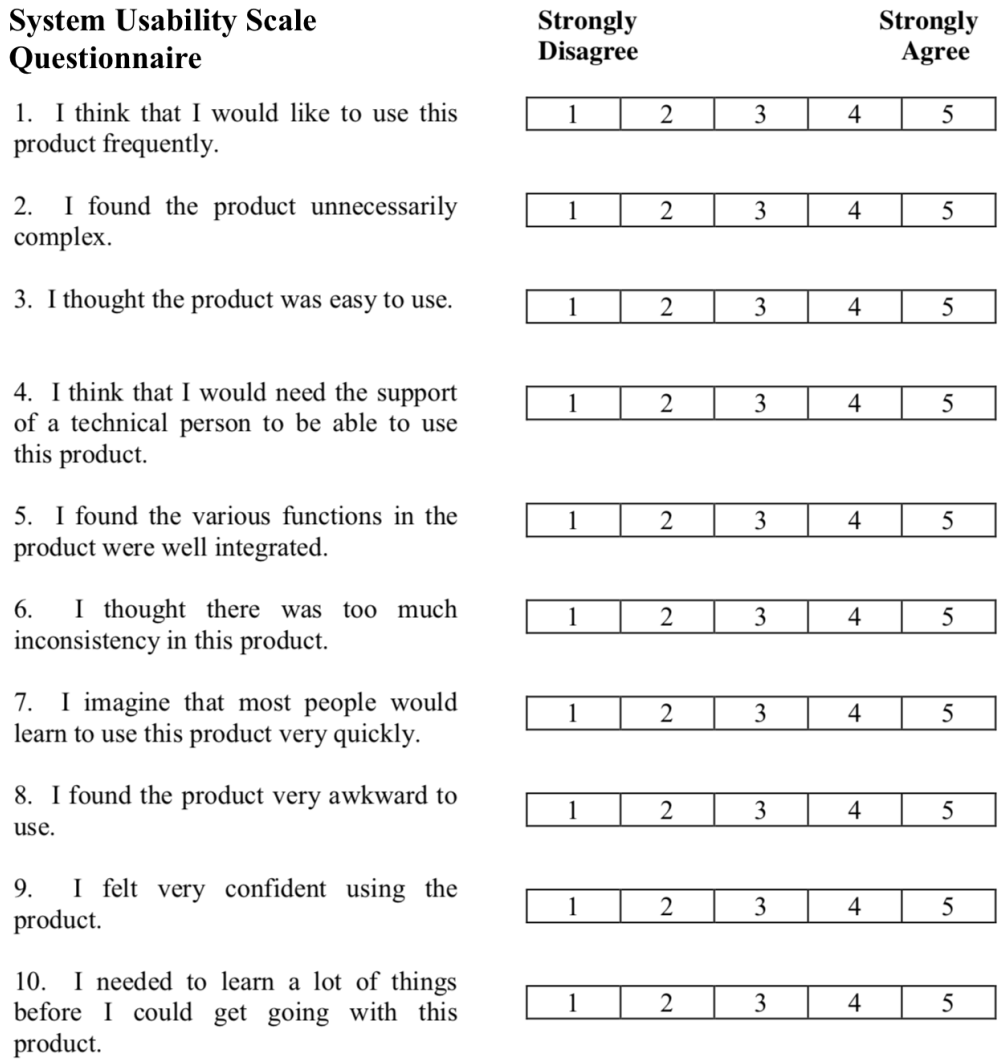
Questionnaires

B.1 Questionnaire Used for the User Testing

In which TNO location were you? [Den Haag - New Babylon; Leiden - Sylviusweg]
What is your ID code? You can find this in my email titled "Start Booking!". (e.g. 1A2X) [Short answer]
What was the type of meeting? [MS Teams - Brainstorming (ID: 1) - Tasks T1 or T2 or T3 or T4; Metaverse - Brainstorming (ID: 2) - Tasks T1 or T2 or T3 or T4; Metaverse - Broadcast (ID: 4) - Tasks T5 or T6; MS Teams - Broadcast (ID: 3) - Tasks T5 or T6]
This meeting was your... [First; Second; Third; Fourth]
What is your age? [Short answer]
How confident are you in your ability to use computers, smartphones, and other digital devices? [Not confident at all; Not confident; Neutral; Confident; Extremely confident]
How much experience do you have with Virtual Reality (VR) technology before participating in this experiment? [None; A little; Moderate; Quite a bit; Very much]
Did you know your meeting partner(s) before this meeting? [No; Partially; Yes]

The rest of the questionnaire contained the following three questionnaires.

B.2 System Usability Scale



System Usability Scale Questionnaire	Strongly Disagree				Strongly Agree
1. I think that I would like to use this product frequently.	1	2	3	4	5
2. I found the product unnecessarily complex.	1	2	3	4	5
3. I thought the product was easy to use.	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this product.	1	2	3	4	5
5. I found the various functions in the product were well integrated.	1	2	3	4	5
6. I thought there was too much inconsistency in this product.	1	2	3	4	5
7. I imagine that most people would learn to use this product very quickly.	1	2	3	4	5
8. I found the product very awkward to use.	1	2	3	4	5
9. I felt very confident using the product.	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this product.	1	2	3	4	5

Figure B.1: The System Usability Scale Questionnaire.

B.3 Perceived Usefulness

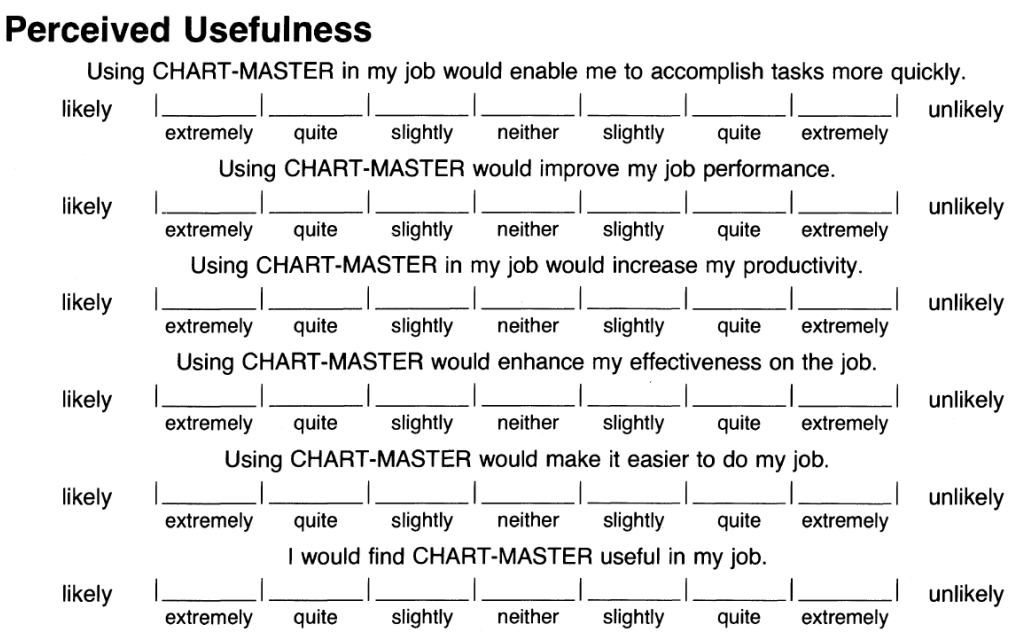


Figure B.2: The Perceived Usefulness Questionnaire.

B.4 Holistic Mediated Social Communication Questionnaire

Question	Level of agreement						
	Strongly disagree	Disagree	Some-what disagree	Neither agree or disagree	Some-what agree	Agree	Strongly agree
1. *I felt in direct contact with the environment*	☐	☐	☐	☐	☐	☐	☐
2. *My sensations were consistent and agreed with the environment*	☐	☐	☐	☐	☐	☐	☐
3. *The environment appeared real*	☐	☐	☐	☐	☐	☐	☐
4. *The environment affected my thoughts just as its real counterpart would*	☐	☐	☐	☐	☐	☐	☐
5. *My interaction with the environment felt realistic*	☐	☐	☐	☐	☐	☐	☐
6. *I felt the presence of the other person(s)*	☐	☐	☐	☐	☐	☐	☐
7. *The other person(s) appeared to feel my presence*	☐	☐	☐	☐	☐	☐	☐
8. *I felt an emotional and intellectual connection with the other person(s)*	☐	☐	☐	☐	☐	☐	☐
9. *The other person(s) appeared to feel an emotional and intellectual connection with me*	☐	☐	☐	☐	☐	☐	☐
10. *The appearance of the other person(s) felt normal*	☐	☐	☐	☐	☐	☐	☐
11. *My appearance seemed normal to the other person(s)*	☐	☐	☐	☐	☐	☐	☐
12. *While communicating, my reasoning felt normal*	☐	☐	☐	☐	☐	☐	☐
13. *While communicating, the reasoning of the other person(s) felt normal*	☐	☐	☐	☐	☐	☐	☐
14. *While communicating, my behavior felt normal*	☐	☐	☐	☐	☐	☐	☐
15. *While communicating, the behavior of the other person(s) felt normal*	☐	☐	☐	☐	☐	☐	☐

Figure B.3: The H-MSC-Q Questionnaire.

Appendix C

Interview Questions

Initial Set of Questions	Additional Questions Added Later
How was your meeting experience?	How was the image quality of your partner(s)?
	Can you make a comparison with the previous meeting(s)?
	How would you rate your [. . .] in the meeting? [engagement, focus, productivity, ideas, flow, social interactions, social connection, communication, distractions]
	How much time (approximately) would you say you spent on the tasks?
Which aspects of this meeting did you like the most?	What expectation do/did you have for the Metaverse?
	To what extent did the metaverse match your expectation?
Which aspects of this meeting did you dislike the most?	
Which aspects, if any, caused you frustration?	

Initial Set of Questions	Additional Questions Added Later
Which kind of issues did you encounter, if any?	To what extent could you accurately identify which meeting partner was speaking?
What do you think was not good enough in the system or in the meeting?	To what extent did you feel the need of more social presence for your meeting (in Teams)?
	How did the added value of presence in the Metaverse help you in the meeting?
What did you feel was missing from the system or in the meeting?	Was there anything missing from the Metaverse that would have made it feel more like a f2f meeting?
To what extent would you use this system for future meetings?	What stops you from using the system?
	What elicits you into using it?
For what kind of tasks/meeting types would you consider using the system in the future?	Thinking about the most important meetings you have at TNO, to what extent would you consider having them in the Metaverse?
	Thinking about the most frequent meetings you have at TNO, to what extent would you consider having them in the Metaverse?
	For what kind of tasks/meeting types would you not consider using the system in the future?
	To what extent does this meeting compare to a real work meeting?

Initial Set of Questions	Additional Questions Added Later
With how many other colleagues would you see yourself using the system?	How does knowing your meeting partner(s) influence you?
	How do you perceive the potential for getting to know a colleague when interacting on Teams versus in the Metaverse?
To what extent would you suggest this system to a colleague that has never used it?	What would make you suggest it to a colleague (or a family member)?
	How would you describe the system to your colleagues?
What was something unexpected, positively or negatively, that happened during the meeting?	
How did you feel after using the system?	How would you describe the feeling of being in the Metaverse?
Were you feeling fine as before, or did you experience headaches or felt dizzy?	
Based on this meeting, what is your opinion on the system?	How much are you looking forward to the next metaverse meetings?

Appendix D

Summary of Research

RQ1	
RQ definition	How does usability change in different communication media?
Dependent Variable	Usability
Independent Variable(s)	Communication media
Evaluation method used	SUS
Null Hypothesis	There is no difference in usability scores between the TNO ICS and MS Teams.
Alternative Hypothesis	There is a difference in the usability of communication media
Expectations	Usability of MS Teams >Usability of TNO ICS
Results of AIC	Communication Media
LMER Random effects (variance)	Subject = 50.87
LMER Fixed effects (p-value)	Communication media: <.001 (***)
ANOVA (p-value)	Communication media: <.001 (***)
Predicted Mean	MS Teams = 89.39 ± 1.85 (SE) TNO ICS = 52.75 ± 2.71 (SE)
Tukey HSD	-
Bonferroni Correction (p-value)	MS Teams-ICS <.001 (***)
Cohen's d	d = 2.97 (large effect size)
Null hypothesis rejected?	yes
Result	Usability of MS Teams >Usability of TNO ICS
Followed expectations?	No
Why?	Technical immaturity of TNO ICS

RQ2.1	
RQ definition	How does the perceived effectiveness of meeting types change in different communication media?
Dependent Variable	Effectiveness
Independent Variable(s)	Communication media, meeting types
Evaluation method used	Perceived Usefulness
Null Hypothesis	There is no relationship between perceived effectiveness and communication media.
Alternative Hypothesis	Communication media has an effect on the meeting effectiveness
Expectations	TNO ICS: Eff. Brainstorm >Eff. Broadcast. MS Teams: Eff. Broadcast >Eff. Brainstorm
Results of AIC	Communication Media
LMER Random effects (variance)	Subject = 0.4
LMER Fixed effects (p-value)	Meeting type: >.05
ANOVA (p-value)	Meeting type: >.05
Predicted Mean	Brainstorm = 4.76 ± 0.17 (SE) Broadcast = 4.59 ± 0.18 (SE)
Tukey HSD	-
Bonferroni Correction (p-value)	Broadcast-Brainstorm: >0.5
Cohen's d	Broadcast-Brainstorm: d = 0.09 (small effect size)
Null hypothesis rejected?	no
Result	TNO ICS and MS Teams: Eff. Brainstorm > Eff. Broadcast (not significantly)
Followed expectations?	Partially, and not significantly
Why?	The tasks were too similar

RQ2.2	
RQ definition	How does the perceived effectiveness of communication media change in different meeting types?
Dependent Variable	Effectiveness
Independent Variable(s)	Communication media, meeting types
Evaluation method used	Perceived Usefulness
Null Hypothesis	There is no relationship between perceived effectiveness and meeting type
Alternative Hypothesis	Meeting types have an effect on the meeting effectiveness
Expectations	Brainstorm: Eff. TNO ICS >Eff. MS Teams Broadcast: Eff MS Teams >Eff. TNO ICS
Results of AIC	Communication Media
LMER Random effects (variance)	Subject = 0.4
LMER Fixed effects (p-value)	Communication media: <.001 (***)
ANOVA (p-value)	Communication media: <.001 (***)
Predicted Mean	MS Teams = 5.97 ± 0.16 (SE) TNO ICS = 3.37 ± 0.19 (SE)
Tukey HSD	MS Teams-ICS: <.001 (***)
Bonferroni Correction (p-value)	MS Teams-ICS: <.001 (***)
Cohen's d	MS Teams-ICS: d = 2.42 (large effect size)
Null hypothesis rejected?	yes
Result	Brainstorm and broadcast: Eff MS Teams > Eff. TNO ICS
Followed expectations?	Partially
Why?	Technical immaturity of TNO ICS

RQ3.1.1	
RQ definition	How does the spatial presence of communication media change with different numbers of meeting participants?
Dependent Variable	Social communication (spatial presence)
Independent Variable(s)	Communication media, number of participants
Evaluation method used	H-MSC-Q
Null Hypothesis	We will find no effect of the number of participants on the spatial presence section of the H-MSC-Q scores
Alternative Hypothesis	The different numbers of participants have an effect on the spatial presence
Expectations	3 and 2 participants: Spa.Pres. of ICS > Spa.Pres. of MS Teams
Results of AIC	Communication Media
LMER Random effects (variance)	Subject = 0.25
LMER Fixed effects (p-value)	Communication media: >.05
ANOVA (p-value)	Communication media: >.05
Predicted Mean	MS Teams = 3.81 ± 0.18 (SE) TNO ICS = 4.07 ± 0.23 (SE)
Tukey HSD	
Bonferroni Correction (p-value)	MS Teams-ICS: >.05
Cohen's d	MS Teams-ICS: d = 0.21 (small effect size)
Null hypothesis rejected?	no
Result	3 and 2 participants: Spa.Pres. of ICS > Spa.Pres. of MS Teams (not significantly)
Followed expectations?	yes, but not significantly
Why?	Substandard virtual environment

RQ3.1.2	
RQ definition	How does the spatial presence of meeting participants change in different communication media?
Dependent Variable	Social communication (spatial presence)
Independent Variable(s)	Communication media, number of participants
Evaluation method used	H-MSC-Q
Null Hypothesis	We will find no effect of communication media on the spatial presence section of the H-MSC-Q scores
Alternative Hypothesis	The different communication media have an effect on the spatial presence
Expectations	TNO ICS and MS Teams: Spa.Pres. with 3 > Spa.Pres. with 2
Results of AIC	Communication Media
LMER Random effects (variance)	Subject = 0.26
LMER Fixed effects (p-value)	Number of participants: >.05
ANOVA (p-value)	Number of participants: >.05
Predicted Mean	Two participants = 3.87 ± 0.22 (SE) Three participants = 3.94 ± 0.23 (SE)
Tukey HSD	
Bonferroni Correction (p-value)	Number of participants: >.05
Cohen's d	Two-Three: d = 0.01 (small effect size)
Null hypothesis rejected?	no
Result	TNO ICS and MS Teams: Spa.Pres. with 3 > Spa.Pres. with 2 (not significantly)
Followed expectations?	yes, but not significantly
Why?	3 people are not enough

RQ3.2.1	
RQ definition	How does the social presence of communication media change with different numbers of meeting participants?
Dependent Variable	Social communication (social presence)
Independent Variable(s)	Communication media, number of participants
Evaluation method used	H-MSC-Q
Null Hypothesis	We will find no effect of the number of participants on the social presence section of the H-MSC-Q scores
Alternative Hypothesis	The different numbers of participants have an effect on the social presence
Expectations	3 and 2 participants: Soc.Pres. of ICS >Soc.Pres. of MS Teams
Results of AIC	Communication Media
LMER Random effects (variance)	Subject = 0.07
LMER Fixed effects (p-value)	Communication media: <.001 (***)
ANOVA (p-value)	Communication media: <.001 (***)
Predicted Mean	MS Teams = 5.35 ± 0.10 (SE) TNO ICS = 4.78 ± 0.12 (SE)
Tukey HSD	MS Teams-ICS: <.001 (***)
Bonferroni Correction (p-value)	MS Teams-ICS: <.001 (***)
Cohen's d	MS Teams-ICS: d = 0.81 (large effect size)
Null hypothesis rejected?	yes
Result	3 and 2 participants: Soc.Pres. of MS Teams > Soc.Pres. of TNO ICS
Followed expectations?	partially
Why?	Substandard Pointcloud

RQ3.2.2	
RQ definition	How does the social presence of meeting participants change in different communication media?
Dependent Variable	Social communication (social presence)
Independent Variable(s)	Communication media, number of participants
Evaluation method used	H-MSC-Q
Null Hypothesis	We will find no effect of communication media on the social presence section of the H-MSC-Q scores
Alternative Hypothesis	The different communication media have an effect on the social presence
Expectations	TNO ICS and MS Teams: Soc.Pres. with 3 > Soc.Pres. with 2
Results of AIC	Communication Media
LMER Random effects (variance)	Subject = 0.07
LMER Fixed effects (p-value)	Number of participants: >.05
ANOVA (p-value)	Number of participants: >.05
Predicted Mean	Two participants = 5.05 ± 0.11 (SE) Three participants = 5.07 ± 0.12 (SE)
Tukey HSD	
Bonferroni Correction (p-value)	Number of participants: >.05
Cohen's d	Two-Three: d = 0.23 (small effect size)
Null hypothesis rejected?	no
Result	TNO ICS and MS Teams: Soc.Pres. with 3 > Soc.Pres. with 2
Followed expectations?	yes, but not significantly
Why?	3 people are not enough

RQ4	
RQ definition	How do extraneous variables and further interaction effects influence the QoE?
Dependent Variable	
Independent Variable(s)	
Evaluation method used	All of the above
Null Hypothesis	
Alternative Hypothesis	
Expectations	The extraneous variables or further interaction effects might explain better the variations in the data
Results of AIC	
LMER Random effects (variance)	
LMER Fixed effects (p-value)	
ANOVA (p-value)	
Predicted Mean	
Tukey HSD	
Bonferroni Correction (p-value)	
Cohen's d	
Null hypothesis rejected?	
Result	The dependent variables influence each other. Age influences usability and effectiveness
Followed expectations?	yes
Why?	

RQ5	
RQ definition	Which other subjective factors influence the meeting QoE?
Dependent Variable	
Independent Variable(s)	
Evaluation method used	Interviews
Null Hypothesis	
Alternative Hypothesis	
Expectations	Other factors that could not be observed through a questionnaire might affect negatively or positively the meeting experience
Results of AIC	
LMER Random effects (variance)	
LMER Fixed effects (p-value)	
ANOVA (p-value)	
Predicted Mean	
Tukey HSD	
Bonferroni Correction (p-value)	
Cohen's d	
Null hypothesis rejected?	
Result	No other factors emerged
Followed expectations?	no
Why?	Participats explained their questionnaire answers instead

Table D.1: The summary of the results.

Appendix E

Recruitment Poster

Figure E.1: The poster used to recruit participants.