NTNU

Norwegian University of
Science and Technology

Impact of Conversation Task When Testing QoE in WebRTC

Marie Haga

Master of Science in Communication Technology
Submission date: June 2018
Supervisor: Poul Einar Heegaard, IIK
Co-supervisor: Katrien De Moor, IIK
Doreid Ammar, IIK

Norwegian University of Science and Technology
Department of Information Security and Communication Technology

NTNU – Trondheim
Norwegian University of
Science and Technology

Impact of Conversation Task When Testing QoE in WebRTC

Marie Haga

Submission date: June 2018
Responsible professor: Poul Einar Heegaard, IIK
Supervisors: Katrien De Moor and Doreid Ammar, IIK

Norwegian University of Science and Technology
The Department of Information Security and Communication Technology

Title: Impact of Conversation Task When Testing QoE in WebRTC

Student: Marie Haga

Problem description:

Video communication allows us to talk with people on the other side of the earth whenever we want to and via multiple devices. In video conferencing (also referred to as telemeeting or videochat) we are talking about "real-time" communication since it enables us to talk back and forth synchronously, as in a face-to-face conversation. Application and services enabling Real-Time Communication has become more popular the last years. Web Real-Time Communication (Web Real-Time Communication (WebRTC)) enables real-time audiovisual communication in the browser with multiple parties, without need for plug-ins or other requirements that you need to have in similar applications, such as Skype [1].

Quality of Experience or QoE (i.e., and the degree of delight or annoyance) of WebRTC-based telemeeting services is to date not yet fully understood. The evaluation of telemeeting experiences is a complex matter that is strongly related to voice, audio or audio-visual quality. As a result, users can experience different types and gradations of quality impairments during a conversation. Telemeeting Quality of Experience (QoE) can not only be influenced by system-related factors, but also by human and context influence factors [2]. This complexity and the huge variety of telemeetings (e.g., in terms of number of parties involved, devices used, type network connection, purpose of the meeting, ...) makes it difficult to use one common assessment method that is valid for all types of equipments and settings. It is important that application and service providers gain insight in this, and can address such issues to ensure an "as optimal as possible" QoE to prevent that users stop using the application [3]. If the quality of the WebRTC is good enough, audiovisual telemeetings have the possibility to reduce the need for traveling to perform face-to face meeting and replace the need of face-to-face communication.

Given this broader context, in this work we build on previous research that has indicated that system factors alone are not sufficient to understand the QoE of an individual in a specific situation. More concretely, we focus on the test task as a specific contextual factor that may influence QoE. When the participant is in the test lab and conduct the test, some tasks may be more engaging than intended [4] and may affect the test result and users' (in)tolerance towards certain impairments. The participants may be more committed to complete the task than to observe and notice what happens on the screen. A better understanding of how the performed task during a call may play a role in this respect can help service providers to adjust as good as possible to the circumstances of a telemeeting.

The main objective for this project will be to:

- Perform a study in a controlled lab setting, using the testbed located on Norwegian University of Science and Technology (NTNU). In the study - use different conversation tasks and network conditions.
- Analyze the data to get an understanding of whether and how the task influences the results of the study. Does the task have a big impact on how the user rates the QoE?
- Provide recommendations on the advantages and disadvantages of the investigated tasks, and on which task may be most appropriate to use when testing QoE in a given scenario.

Responsible professor: Poul Einar Heegaard
Supervisors: Katrien De Moor
Doreid Ammar

Abstract

Video conferencing was previously associated with dedicated hardware and professional conferencing systems. The last years applications and services enabling Real-Time Communication with no need for additional software support, called WebRTC, have become more popular. Variety in telemeetings is huge, and the increasing complexity makes it challenging to evaluate telemeeting Quality of Experience (QoE), by selecting one approach that is valid for all types of equipment and circumstances. Many studies have been conducted to achieve an understanding of which factors play a role/have an impact on telemeetings. Several of these have indicated that system factors alone are not sufficient to understand the QoE of an individual in a specific situation.

Previous studies have put participants in an experimental environment and given them a conversation task to simulate a natural conversation. The results from some of these earlier studies indicate that the chosen task have an impact when the test subject rates the QoE, and that the tolerance towards technical impairments may differ, depending on the task at hand. Given this context, the main motivation for this thesis is to contribute to literature in this respect, by investigating whether and how the conversation task influences QoE under different circumstances.

I have conducted an experiment by using a QoE-testbed. Participants were exposed to four different technical quality conditions (good quality, distorted audio and video, distorted audio and distorted video). Two conversation tasks were included in the experimental design, namely the LEGO-task (building blocks task) and the Free Conversation task. Data gathered in a previous study, using the Celebrity Name Guessing task, were also integrated into the dataset, in order to compare the findings.

The results of this study point in the direction that the task does matter and does influence QoE. When it comes to the traditional perceived quality ratings, there are significant differences when comparing the different conditions *within* each task, but not *between* the different tasks. Still, there are some tendencies. Overall, the quality ratings are slightly higher for the Free Conversation task than for the other two tasks (LEGO-task and Celebrity Name Guessing). In addition, it could be observed that in most cases, the Free Conversation yields more compact ratings with lower variability among the test subjects. When looking at affective state the test subjects clearly feel the least aroused during the Free Conversation task and more aroused in LEGO-task and Celebrity Name Guessing. A plausible explanation in this respect is the competition element that is inherently in these tasks.

To summarize, the choice of conversation task is complex, and many factors have to be taken in to consideration. There are advantages and disadvantages for both tasks and some differences are clear, while others are more implicit. In general, the Free Conversation task provides higher agreement of the quality ratings, but is clearly less engaging. The LEGO task and Celebrity Name Guessing task are more engaging and implicitly the tolerance towards impairments seems to be lower.

Sammendrag

Videokonferanser var tidligere knyttet til dedikert maskinvare og dedikerte applikasjoner. De siste årene har applikasjoner og tjenester som muliggjør sanntidskommunikasjon, uten behov for noe annet enn en nettleser, kalt WebRTC, blitt mer populære. Mangfold i videosamtaler er enorme, og den økende kompleksiteten gjør det utfordrende å evaluere brukeropplevelsen (QoE) på videosamtalene, særlig å finne en tilnærming som kan gjelde for alle applikasjoner og situasjoner. Mange studier har blitt gjennomført for å få en forståelse av hvilke faktorer som påvirker videosamtaler. Flere av disse har indikert at systemfaktorer alene ikke er tilstrekkelige til å forstå brukeropplevelsen til et individ i en bestemt situasjon.

Tidligere studier har satt deltakerne i et eksperimentelt miljø og gitt dem en samtaleoppgave, som skal simulere en naturlig samtale. Resultatene fra disse indikerer at den valgte oppgaven kan ha innvirkning når testpersonen vurderer brukeropplevelsen og at toleransen mot tekniske forstyrrelser kan variere, avhengig av oppgaven som er gitt. I denne konteksten er hovedmotivasjonen for denne oppgaven å bidra til litteratur ved å undersøke om og hvordan oppgaven påvirker brukeropplevelsen under ulike forhold.

Jeg har utført et eksperiment ved å bruke en QoE-testbed. Deltakerne ble utsatt for fire forskjellige tekniske kvalitetsforhold (god kvalitet, forstyrret lyd og video, forstyrret lyd og forstyrret video). To samtaleoppgaver ble inkludert i den eksperimentelle utformingen, en instruksjons preget LEGO-oppgave og naturlig samtale. Data samlet i en tidligere studie, som brukte oppgaven Gjett Hvilken Kjendis, ble også integrert i datasettet for å sammenligne mot funnene.

Resultatene av denne studien peker i retning av at den gitte oppgaven påvirker brukeropplevelsen. Når det gjelder de tradisjonelle oppfattede kvalitetsverdiene, er det betydelige forskjeller når man sammenligner de forskjellige forholdene *innen* hver oppgave, men ikke *mellom* de forskjellige oppgaver. Likevel er det noen tendenser. Samlet sett er kvalitetsvurderingene litt høyere for natrulig samtale enn for de to andre oppgavene (LEGO-oppgave og Gjett Hvilken Kjendis). I tillegg kan det observeres at i de fleste tilfeller gir naturlig samtale mer kompakte svar og lavere variasjoner blant testpersonene. Når du ser på affektiv tilstand, føler testpersonene seg minst opphisset under naturlig samtale og mer oppglødd i LEGO-oppgaven og Gjett Hvilken Kjendis. En trolig forklaring på dette er konkurranseelementet som er i disse to oppgavene.

For å oppsummere er valg av samtaleoppgave komplisert, og mange faktorer må tas i betraktning. Det er fordeler og ulemper med begge oppgaver, og enkelte forskjeller er klare, mens andre er mer implisitte. Generelt gir naturlig samtale høyere enighet i rangeringen av kvalitet, men er tydelig mindre engasjerende. LEGO-oppgaven og Gjett Hvilken Kjendis er mer engasjerende og implisitt synes toleransen mot nedsatt funksjonsevne å være lavere.

Preface

This thesis is submitted as the final part of my master's degree at the Department of Information Security and Communication Technology at the Norwegian University of Science and Technology.

This thesis could not have been completed without the help of several people.

Enormous gratitude is therefore given to my supervisor, Katrien De Moor, for her guidance, support, and valuble feedback, throughout this final semester. You have been an incredible resource. I would also like to thank my Professor Poul Einar Heegaard for his thoughts and advice which have helped form this thesis, and also Doreid Ammar for quick responses and guidance from France.

Thanks to all the people who did not dare say no, when I asked them if they wanted to contribute to my master thesis by testing a videoconferencing solution for one and a half hour.

Thanks to Mom, Julie Haga, and Ane Tronstad who have spent countless hours reading and improving this thesis. And also thanks, to Marie Buøen for staying with me to the very end.

Thanks to the professors at NTNU who lent me their offices, so I could conduct the experiment.

Finally, I would like to thank all my amazing friends in Trondheim, it would not have been the same without you.

Hopefully the Quality of Experience reading this thesis would be excellent.

Marie Haga

Trondheim, June 18th 2018

Contents

List of Figures

List of Tables

List of Acronyms

ACR Absolute Category Rating.

EEG Electroencephalogram.

EMG Electromyography.

fMRI functional Magnetic Resonance Imaging.

GSR Galvanic Skin Response.

iSCT interactive Short Conversation Test Scenarios.

ITU International Telecommunication Union.

MOS Mean Opinion Score.

NIRS Near-InfraRed Spectography.

NTNU Norwegian University of Science and Technology.

QoE Quality of Experience.

QoS Quality of Service.

SAM Self-Assessment Manikin.

SCT Short Conversation Test Scenarios.

SD Standard Deviation.

SSH Secure SHell Commands.

WebRTC Web Real-Time Communication.

Chapter 1
Introduction

Video communication allows us to talk to people on the other side of the world whenever we want to. In video conferencing I refer to "real-time" communication since it enables us to talk back and forth like in a face-to-face conversation. Over the last years, video conferencing has become more common, and the range of applications enabling video conferencing (also called video-chat or telemeeting) is increasing.

The International Telecommunication Union (ITU) defines telemeeting as: *A meeting in which participants are located at at least two locations and the communication takes place via a telecommunication system*[6]. The term telemeeting is used to emphasize that a meeting often is more flexible and interactive than a conventional business teleconference and could also be a private meeting. The telemeeting could be audio-only, audiovisual, text-based or a mix of these modes.

Whereas video conferencing previously was more associated with dedicated hardware and professional conferencing systems[1]: WebRTC, explained in section 2.1.2, enables real-time audiovisual communication in the browser with multiple parties, without the need for plug-ins or other requirements that you need to have in similar application, such as in Skype[1].

1.1 Motivation

This huge variety of telemeetings (e.g., in terms of technical circumstances, professional versus leisure context, and number of parties involved) implies a number of important challenges, and many studies have been conducted to achive an understanding of which factors play a role/have an impact on telemeetings Quality of Experience (QoE). QoE is a measure of the delight or annoyance of a customer's experiences with a service [2].

[1] https://www.skype.com/

One challenge is the degree to which teleconferencing systems can meet users' expectations in different settings, and provide the best possible experience given the circumstances of a call. This will strongly determine the success or failure of video conferencing applications and services. The quality provided to and, experienced by, the users is thus an essential element to consider. Users can experience different types and graduations of quality impairments during a conversation. It is important that application and service providers address such issues, to prevent that users stop using the application [3]. If the quality of the WebRTC is good enough, audiovisual telemeetings have the possibility to reduce the need for traveling to perform face-to-face meeting and replace the need for face-to-face communication.

Furthermore QoE in WebRTC are not fully understood. The increased complexity makes it difficult to use one common approach/method to evaluate telemeeting QoE, that is valid for all types of equipment and circumstances. QoE of a telemeeting is also more than one quality score and QoE of a telemeeting can be influenced by a number of factors; human influence, system influence, and context influence.

Several studies have indicated that system factors alone are not sufficient to understand the QoE of an individual in a specific situation. Previous studies have put participants in an experimental environment and given them a conversation task to simulate a natural conversation. The results of some of these earlier studies indicate that the given task may have an impact when the test subject rates the QoE and that the tolerance towards technical impairments may be very different, depending on the task at hand. Some tasks may also be more engaging than others [4], thus affecting the test results. The participants may be more committed to complete the task than to observe and notice what happens on the screen. The given task may also influence whether the participants notice delays in the network, or not. If the conversation consists of breaks and does not flow naturally, nor will the test subjects discover delay and disruptions in the network.

Given this context, the main motivation for this thesis is to contribute to the literature in this respect, by investigating whether and how the conversation task influences QoE under different circumstances.

1.2 Objectives

The project description has three primary objectives for this master thesis:

- Perform a study in a controlled lab setting, using the testbed located on NTNU. In the study - use different conversation tasks and network conditions.
- Analyze the data to get an understanding of whether and how the task influences the results of the study. Does the task have a big impact on how the user rates the QoE?
- Provide recommendations on the advantages and disadvantages of the investigated tasks, and on which task may be most appropriate to use when testing QoE in a given scenario.

1.3 Outline

The thesis has the following structure

Chapter 2: Background: Provides the reader with the theory behind QoE, WebRTC and conversation task.
Chapter 3: Methodology and Experimental Setup Provide a brief presentation of methodology as a term, and an explanation of how I conducted the experiment.
Chapter 4: Results Presents the results observed by analyzing the collected data.
Chapter 5: Discussion Discussing the results and the limitations of this research. The discussion will further be a basis for the conclusion.
Chapter 6: Conclusion and Future Work Presents the conclusion. The section for future work provides suggestions for further work based on the results from this research.

Chapter 2
Background

After a short introduction to areas of utilization's of video conferencing, I will in this chapter render an account of the concept of WebRTC, a definition of Quality of Experience (QoE) and Quality of Service (QoS), factors influencing WebRTC, how to measure Quality of Experience, and different conversation tasks.

2.1 Video Conferencing

The range of possibilities and affordances for videoconferansing users has thus strongly increased. Yet, as I will further discuss, it also introduces interesting challenges, which require a better understanding of users' QoE (and factors influencing it) with video conferencing.

2.1.1 Video Conferencing: At Work Versus at Home

In a professional context, video chat is most frequently used in teleconferencing and telepresence solutions. Teleconferencing can be used for both spontaneous telemeetings and larger conferences, which minimize the need for businesses traveling. Many available video chat solutions offer screen sharing, which enables the participants working on the same screen regardless the distance between them. As long as the technology is reliable, videoconferencing can save companies both time and traveling costs [7].

Video calls have also become increasingly common in family households. Users of video solutions explain that video communication brings them closer and is socially more involving compared to an audio call [2]. Therefore video communication is highly popular among families and friends living far apart from each other.

2.1.2 WebRTC

Traditionally, video conferencing required a particular application, plug-in, and complete multimedia stack for streaming. A newer solution in the market, Web Real-Time Communication (WebRTC), enables real-time video communication with no need for additional software support, registration, or expenses [1]. Now the website takes on both control point, source, and destination of the realized communication, providing implementation flexibility of interactive communication services in various topologies (point-to-point, many-to-many), offering developers the choice of using existing protocols or developing their own protocols [8].

appear.in

appear.in[1] is a WebRTC-based free browser-to-browser service from Telenor. It does not require any registration or downloading, only a browser [9]. appear.in supports up to eight participants and the users can enter an easy-to-remember URL with the specified room name, and wait for other users to access the same room. The users can also chat and use screen sharing.

The Department of Information Security and Communication Technology at NTNU hosts a research version of the appear.in server, which constitutes the QoE-testbed [10, 11]. This testbed, later explained in section 3.3.2, enables video conferencing in a controlled environment. Therefor appear.in will be used as the concrete application further in this study.

2.1.3 Implications and Challenges of Assuring High Video Conferencing QoE

Video conferencing is used and can be used in a wide range of settings. Ultimately, the degree to which teleconferencing systems can meet users' expectations and provide the best possible experience given the circumstances of a call, will strongly determine of the success or failure of video conferencing applications and services.

Given the variety of telephone- and video conference solutions, and given the fact that such systems are used for both professional and private life, assessing QoE of telemeetings is very difficult and calls for a high degree of variability regarding assessment methods.

Even though quite an amount work already has been conducted, the QoE of telemeetings is not yet thoroughly understood and investigated. Although there is a standardized recommendation on quality evaluation test for multiparty telemeetings

[1] appear.in

available [6], there are many detailed questions and potential influence factors that require further studies.

In the next section, I will introduce the concept of QoE more thoroughly. In a broad sense QoE refers to the degree of delight or annoyance of the user of an application or service. Thereupon, I introduce potential influence factors, first in general and then for video conferencing settings more specifically.

2.2 Perceived User Quality; Definition of QoS and QoE

There are several different perspectives on how to measure user quality in a service; but QoE and QoS are the most common. The aim of a service should be to enable good and positive experiences for users (QoE), (QoS) is the primary tool to get there. However, produced/delivered and perceived quality is not the same, and therefore it is therefore important to consider these two related, yet different concepts in more detail.

2.2.1 Definition of Quality of Service

Quality of Service (QoS) is defined as the ability of a network to provide a service at an assured service level. QoS is the most frequently used method to measure the performance of a service. There are several different definitions of QoS, but the definition stated by ITU, International Telecommunication Union is:

> Quality of Service (QoS): Totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service [12].

QoS weighs the actual service delivered, and the three main network QoS-parameters are delay, jitter, and packet loss. However, how services are perceived and experienced by a user, refers to a much broader range of aspects. QoE refers to the user-perspective on the experience of the service, which is influenced by factors such as user characteristics and context of the user [13].

2.2.2 Definition of Quality of Experience

While QoS measures the performance of the network, a good QoS does not necessarily imply that a user will be happy and satisfied with the service. Quality of Experience, QoE, focuses on the entire service experience and is expressed in human feelings; good, excellent, and poor. ITU defines QoE as followed:

> Quality of Experience (QoE): The overall acceptability of an application or service, as perceived subjectively by the end-user [12].

This definition is rather vague and narrow because it does not define what "overall acceptability" means. The European Network of Excellence Qualinet proposed a new definition, which to a larger extent takes human-related factors into account:

> Quality of Experience (QoE): Is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations concerning the utility and/or enjoyment of the application or service in the light of the user's personality and current state [13].

This definition refers to QoE as an affective state and avoids the vague concept of "overall acceptability" as a measure of QoE. In addition, it refers to not only utility, but also to enjoyment as a desired outcome. This implies that the goal of QoE is no longer only about satisfying expectations related to the utility of a service or an application. It is also about how users feel, and about how experiences with technology involve and move people emotionally[14].

In 2017 ITU released a recommendation, ITU-T P.10/G.100, which contains terms and definitions associated with network performance, Quality of Service, and Quality of Experience [15]. ITU's updated definition of QoE is now:

> Quality of Experience (QoE): The degree of delight or annoyance of the user of an application or service [15].

QoE is related to (and may in some cases entirely depend on) QoS; If the QoS-parameters are low, the user experience may also decrease. Low QoS-parameters are most likely a result of poor connection. While if the QoS-parameters are acceptable, the individual backgrounds and expectations of users may lead to different QoE-values, trade-offs and tolerance levels.

QoE is user-dependent. It is influenced by many parameters, such as importance of content, the user's terminal device, and by environment. When evaluating QoE, it is important to consider all factors which contribute to the overall user value; such as suitableness, cost, reliability, efficiency, privacy, security, and user confidence [16].

2.3 Factors Influencing QoE

The Qualinet White Paper on Definitions of Quality of Experience defines the factors influencing QoE as follows[13]:

> Influence Factor: Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user [13].

QoE is strictly subjective and refers to the end users feelings, motivation, and background. In this sense, the influence factors discussed here are the independent variables, whereas the resulting QoE as perceived by the end user is the dependent variable [2]. Factors influencing QoE can be grouped into three categories; human, system, and context influence factors.

Human Influence Factors: are any variant or invariant properties or characteristics of a human user. The characteristics can describe the demographic and socio-economic background, the physical and mental constitution, or the user's emotional state [13].

System Influence Factors: refer to properties and characteristics that determine the technically produced quality of an application or service [13].

Context Influence Factors: are factors that embrace any situational property to describe the user's environment [13].

In the next sections, I readdress these influence factors from the telemeeting point of view. There is a broad range of telemeetings, that can be very different in their character. The situation can be complex, with several different types of equipment and a different number of interlocutors at the site. Skoweonek, Schoeneberg, and Berdtsson describe in "Quality of Experience" [2] three different types of influence factors and how they can influence the perceived QoE of a telemeeting. The different types will be elaborated in the following subsections.

2.3.1 Human Influence Factor in Telemeetings

Different personalities of the participants or combination of different personalities can have an impact. If the pattern of the conversation is that one person talks a lot, while the other person mainly listens; the two persons could have a very different perception of the conversation. Combination of personalities also determines the conversational structure concerning turn-taking behavior[2]. It is a good solution to chose and allocate the participants randomly. Hopefully, the related effects and personalities are distributed equally.

Another essential aspect is the familiarity of the participants; participants that know each other tend to have more natural and fluent conversations and may detect abnormalities like longer response times or differences in voice characteristics faster [2]. They are also most likely more sensitive to impairments. In the study "Subjective quality assessment of video conferences and telemeetings" [17, 18], Berndtsson, Folkesson, and Kulyk addresses that delays seemed to be more noticed at the end of the test, when the test subjects got to know each other more. However, in real-life

usage, systems are not always used by participants who are familiar with each other.

2.3.2 System Influence Factor in Telemeetings

System influence factors refer to all technical aspects of the system that contribute to the quality judgment. This includes network aspects like packet delay, jitter, and bandwidth. The users of the system only judge the results of video artifacts and audio quality [19].

In order to ensure the optimal usage of the network resources to get the best possible audio and video quality in WebRTC, different studies have been conducted: They have looked at which sacrifice is preferred by the user, and which qualities the user prefers in a telemeeting:

- It is more important to synchronize audio and video than to minimize audio delays (for delays less than 600 ms) [17].
- Distortion of both audio and video leads to lowest quality ratings and highest annoyance ratings [4].
- Studies indicate that a 800 ms end-to-end delay is considered unacceptable and affects the experienced interaction quality in a negative way [20]

2.3.3 Context Influence Factors in Telemeetings

The experience quality of a call can differ for participants in the same room, even though the technical setup is the same [2]. This is a result of context factors, which include everything not earlier mentioned: How far from the screen the participant is located, how many people being present in the same room, how many parties are involved in the call, which devices the other parties in a call use, and which type of network they call from, etc.

A particularly context influenced factor for telemeetings is the different potential areas of utilization, e.g., business meeting versus private and more personal conversations. It is assumed that business telemeetings are mainly motivated by a particular agenda, or to accomplish specific tasks while private sessions are mostly driven by desire of social connectivity [2]. Possibilities of video calls are endless: catching up with family living far away, students having left home needing help with the washing machine, or meeting with a department of your company that is located in another city. In this respect, the setting of a call is of importance as it may have implications on the tolerance towards technical impairments. This also poses challenges to the

measurement and evaluation of QoE in controlled lab settings: Which tasks are most appropriate and what are the consequences of using one task instead of another? The potential impact of the task on QoE and tolerance towards different types of impairments is currently still poorly understood. Before going into detail about the setting and task for a telemeeting in section 2.5, I first share some general considerations related to the measurement of QoE.

2.4 Measuring QoE

When measuring QoE, you get a numerical value you can use to objectively assess the user satisfaction. Measurements of QoE are usually categorized into subjective and objective methods [21]. The different methods of measuring will be explained in the following subsections.

2.4.1 Subjective Methods

In a subjective experiment, the subjects are asked to provide their opinions using a "rating scale". The purpose of the scale is to translate a subject's quality assessment into a numerical value that can be averaged across subjects and other experimental factors [5]. Mean Opinion Score (MOS), gives a numerical indication of the perceived QoE. MOS is expressed in a number, from 1 to 5, 1 being the poorest and 5 the best score. MOS is subjective, as it is based on figures from test subjects during the test[22]. The most commonly used scale is the 5-point Absolute category rating (ACR) scale:

Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 2.1: 5-point ACR scale [5]

Subjective methods are based on recommendations from ITU[23]. These recommendations regard quality in different application domains, conduction, scales and environment etc. Studies involving users typically take place in controlled lab settings and provide a high level of control. If conducted properly, they give a high level of internal validity and they are typically based on the manipulation of one or more independent variables and of exposing test participants to different conditions [19].

A tool that can be used to measure emotions is the Self-Assessment Manikin, commonly referred to as the SAM-scale. The SAM provides a simple, fast and non-linguistic way of assessing emotional state along the principal emotional dimensions of valence, arousal, and dominance [24]. Figure 2.1 shows three different SAM-scales. From the top; valence, arousal, and dominance.

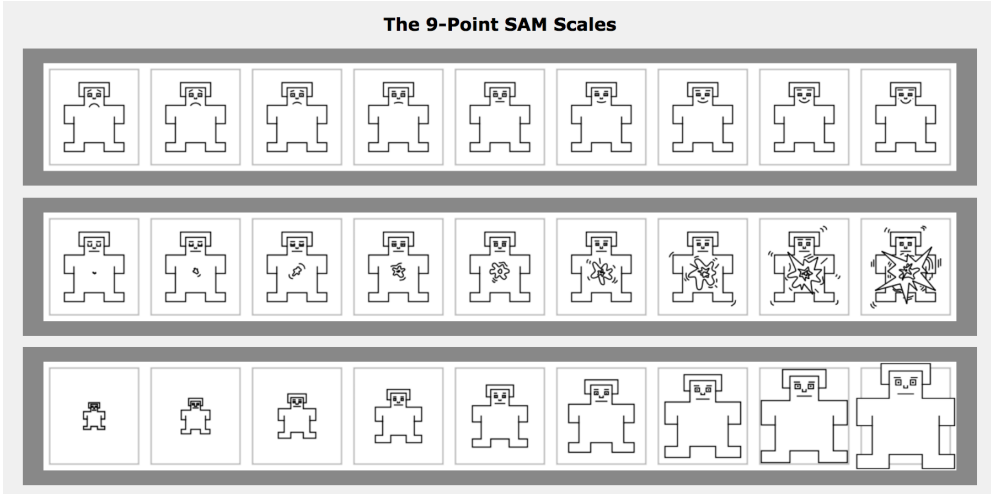


Figure 2.1: The 9-Point SAM Scales for Valence, Arousal and Dominance

These and other scales can be integrated into a **questionnaire** and data are typically analyzed using statistical methods. Experiments can also be complemented with more qualitative methods such as **in-depth interview**. Using such methods, participants get the possibility to express in detail how they experience the system, what they find most important, and why they rated the stimuli as they did, etc.

2.4.2 Objective Methods

Objective methods are based on externally observable and measurable changes in behavioural, and do not involve explicit user feedback. Objective methods provide data which is expected to be linked to experience, and which potentially allows to estimate QoE [23].

- **Task Scores:** uses as a metric to see whether the user manages to complete a task during the experiment. Common metrics are task completion times or successful vs. unsuccessful attempts.
- **Speech Patterns:** looks directly at the ongoing interaction. Counts speaking times, length of turns and pauses, simultaneous speech, etc.
- **Physiological measurements:** covers the quantitative measurement and visualization of physiological structure: Galvanic Skin Response (GSR), heart rate variability, Electroencephalogram (EEG), Near-InfraRed Spectography (NIRS), Electromyography (EMG), functional Magnetic Resonance Imaging (fMRI) ect.

2.5 Conversation Task

As mentioned earlier, subjective testing in the context of video conferencing QoE entails several challenges. In this work, I focus more explicitly on the role of the conversation task.

When testing QoE in WebRTC (or another video conferencing application) you put together two, or more, participants in a WebRTC-application. Then you need the participants to start talking to each other. This is the purpose of the conversation task; to simulate a conversation between friends, colleagues, or strangers.

This section consists of justification of why the test task is essential.

- The test task is part of the broader category of context influence factors, and the perceived conversation quality will probably depend on the task for the conversation. Having attention to the **test modes** is of essential importance [2]. Which test to choose depends on the technical conditions we should test on, and the conclusions concerning validity that shall be drawn.
- Further, we can have **formal or casual conversations**; The conversation form may influence the result[2]. If the task is formal, like booking a train or asking for information, it is not an advantage that the participants know each other. If the purpose of the conversation is fluency, it is recommended that

the test subjects know each other from before or have the possibility to get to know each other before the test.

- Especially when analyzing tolerance towards video impairments, it is crucial that the test participants pay **attention to the screen** during the whole session, avoiding focusing on written material[2]. A consequence of this is that aspects, such as instructions necessary during the test, should be limited. Moreover, if the conversation is not scripted, speakers can be interrupted spontaneously, and end up in natural double-talk situations.
- In general, it is highly recommended **to forbid the participants to talk about the system and its properties**, then this could influence their assessment of the system
- The task must have sufficient **face value**: It must resemble real-life audiovisual communication to an adequate degree[2]. In particular, it is preferable that the task is performed by two subjects and not by one subject and an experimental leader.
- It is preferred that the task is, in itself, **sufficiently rewarding for the subjects.** This has several advantages: The subjects learn the task faster, and they are less susceptible to fatigue and loss of motivation.
- To simulate a more **interactive conversation,** the participants could be challenged to debate, take opposite standpoints, and neglect common courtesy.

The goal of the task given, is to encourage free conversation. I have chosen to look into three conversation tasks from the Recommendation ITU-T P.1301 [6] and Recommendation ITU-T P.920 [18]:

2.5.1 Celebrity Name Guessing Task

In each round each player is given the identity of a celebrity and they try to disclose which celebrity by guessing. The players are only allowed to ask questions that can be answered with "Yes" or "No", and it is your turn until you get the answer "No", then it will be the other player's turn. If one player guesses the correct celebrity, he gets one point and may continue guessing a new celebrity. The test supervisor signals the end of the game and stops the call after 4-5 minutes, declaring the player that has guessed most celebrities the winner of the game [6].

2.5.2 Free Conversation

The intention of the task is to make the participants chat, and have a conversation as natural as possible[6]. To facilitate the discussion topics, a paper with topic

suggestions can be handed out. Nevertheless, the participants are free to choose any topic. To strive for the conversation to be as natural as possible, it is especially advantage if the participants know each other from before.

Berdtsson, Folkesson, and Kulyk have performed a study to investigate how perceived video quality depends on video resolution and bit rate, at viewing distances frequently used during tests [17]. Two different tasks were used; Free Conversation and a quiz game where one participant was given a word that the other participant cooperated to guess. In this study they used 14 different combinations of ear-to-ear audio and video delays. The findings were that participants responded more critical to long delays during the Free Conversation than during the quiz task, in both the audio and the audiovisual test. Berdtsson, Folkesson, and Kulyk adresses that the Free Conversation to be an appropriate test task for both audio and audiovisual conversational tests.

2.5.3 Building Block Task

One participant is given a bag of multicolor, interlocking construction blocks. The other subject gets a completed figure, made from an identical set of blocks. The first participant's goal is to build the same figure with help from the other subject and verify its correctness.

Bräuer, Ehsan, and Kubin ,[25], tested the Building Block scenario from [18], and customized it for the purpose of assessing conversational audiovisual communication. They called the new scenario "LEGO Model scenario". In this version both of the test persons have a construction plan and they are both giving building instructions. The construction plans are designed with an information gap, with every second step asynchronously excluded. Bräuer, Ehsan, and Kubin experienced that most of the subjects were not bored by the task, rather enthusiastic about finishing the models: This being an advantage because bored users tend to lose motivation and get distracted. During the task it was natural for the participants to make use of both the audio and the video channel for explaining the individual steps. The design of the building plans forces turn-taking in the active speaker role. The dialogue structure turned out to be very interactive since exploitation was often interrupted by interposed question of the listening side. The most important result of the assessment is that test persons seem to tolerate delay to a large extent. Bräuer, Ehsan, and Kubin assume the test subjects were deeply concentrated and involved in the task. They suggest that, to accomplish a particular task successfully, users seem to be more willing to tolerate delay or are not that sensitive to it.

2.5.4 Comparison Different Conversation Scenarios

It is well-known that the QoE may differ/depend on the specified task performed in the study. From the users' point of view, it is natural that the perceived quality varies according to how crucial or demanding the conversation task is [26]. In this section I will present several papers which have tested different scenarios/tasks and their findings.

In the paper "Audio and Video Channel Impact on Perceived Audio-visual Quality in Different Interactive Contexts" [27], Belmudez, Moeller, Lewcio, Raake, and Mehmood use two different scenarios; one emphasizing more on the video aspect and one focusing more on the audio part. They used the "building block scenario" from [18] and Short Conversation Test Scenarios (SCT) [28]. Their result indicates that test participants were more sensitive to the video degradation's, and they were able to distinguish the differences between the predefined levels of quality. The type of task showed to be a factor which influence significantly the perceived audio quality depending on how much the subjects dedicate their attention to either the audio or the video channel. The audio quality got better ratings when the subject's attention was more focused on the task and video. The results of this experiment showed that in an interactive context, where the people have to focus on a task and interact, the audio quality is rated in a less differentiated way. This might be an indication of audio quality being mainly judged on an overall acceptability criteria.

Wang, Yang, Xie, and Wan used six different conversation tasks with predetermined content, referred to as "non-free conversations", to evaluate the delay perception [26]. When the delay is minimal, no subjects discovered the delay, independent of the task given. When the delay is increasing, the differences between tasks become more and more obvious. The different tasks show distinctly different declining trends and the task with the most interaction detected the delay easiest.

Berdtsson, Folkesson, and Kulyk used SCT and interactive Short Conversation Test Scenarios (iSCT) in their study [17]. Informal test showed that the iSCT were more sensitive to delay than the SCT scenarios, but the difference was not large. Even if the iSCT scenarios are more interactive, the fact that one person is supposed to reply made the other wait politely for the answer. Two people who know each other participated in each conversation. Their findings illustrates that it is more important to synchronize the audio and video than to have short delay for audiovisual Free Conversations. Also, there are tasks that make it easier to notice delays, but it is important to provide a natural situation.

In the short paper "Exploring diverse measures for evaluating QoE in the context of WebRTC"[4] they addresses that a distortion of both audio and video leads to lowest quality and highest annoyance rating. The result also shows a low tolerance for

audio distortions and suggest an impact of quality impairments on experienced affect, however, this can only be observed for annoyance. The self-reported valence is rather similar across the experimental conditions. However, the authors have a hypothesis that the test task was more engaging than intended and may have confounded the intended effect.

Given this broader context and some of the particular challenges introduced in this chapter, this work aims to contribute to the literature by focusing more explicitly on the conversation test task. More specifically, in this thesis I will set up an experiment to investigate the impact of the conversation task when evaluating QoE in WebRTC under four different technical conditions that represent common quality problems in video conferencing/telemeeting settings. The experimental setup is explained in chapter 3.

Chapter 3
Methodology and Experimental Setup

The main objective of this thesis is to perform an experimental study using the QoE-testbed (section 3.3.2) in a controlled lab setting, in order to further investigate the impact of the conversation task on the tolerance towards impairments. Different conversation tasks were used and evaluated under different technical conditions. The methodology chapter describes the procedure used to answer the problem description, with the intention of enabling continuation or testing of the work implemented.

The chapter will initially provide a brief presentation of methodology as a term as well as different research methods. Choice for research method and research design will be described in detail. Further, this chapter will explain how I conducted the experiment, give a description of the testbed and a review of how the data have been collected and processed.

3.1 Quantitative and Qualitative Methods

In the social sciences, a distinction is commonly made between quantitative and qualitative methods. The difference between these two categories of methods lies amongst others in how data are recorded and analyzed. Qualitative methods aim at capturing meaning and experience that cannot easily be quantified or measured. Quantitative approaches, on the other hand, have the advantage that they focus on shaping the information into measurable devices (numerical data), which in turn allows the researcher to perform mathematical operations of a more substantial amount of data [29]. The main disadvantage of quantitative research is that the context of the study or experiment is ignored, and it does not study effects in a natural setting. A large sample of the population must be studied to get accurate and generalizable results; the larger the sample of people researched, the more accurate the result will be [30]. Quantitative methods are characterized by precision; they are broad-based, can help to reveal common features, and often use fixed response options and closed questions, e.g., in questionnaires. Furthermore, collected data are

linked to distinct phenomena, and the design aims to convey explanations based on the gathering of numerical data [29].

In this project, I will use questionnaires to gather quantitative self-reported data. The research is based on an experimental study design, as will be discussed in more detail.

There are two main types of quantitative research design: Experimental designs and non-experimental designs. The basis of the experimental method is the experiment, which is defined by Brown and Melamed as "a test under controlled conditions that is made to demonstrate a known truth or examine the validity of a hypothesis" [31]. When we do experimental research, we want to control the environment, making it possible to isolate the variables that we want to study. When you control the environment, you can claim to have determined causality more than in any other type of research.

However, experimental research can create artificial situations that do not always represent real-life situations. This is largely due to the fact that all other variables are tightly controlled, which may not create a fully realistic situation. Experimental research design helps to ensure internal validity, but this tends to be at the expense of external validity. If the study is conducted in a natural environment, such as in a hospital, at home, or in a office it is as a rule not possible to control the extraneous variables [32]. Non-experimental research, on the other hand, is when a researcher cannot control, manipulate or alter the predictor variable or subjects, but relies on interpretation, observation or interactions to come to a conclusion [33].

In the work presented in this thesis, the QoE-testbed helps me control the environment and permits me to only concentrate on those variables that I want to examine. This makes it possible to choose experimental research for this study.

3.2 Research Methodology

To ensure both repeatability and validity, this research is prepared and conducted in a systematic way. The research consists of three parts. The first part is a literature study focusing on QoE and factors influencing QoE-ratings, and different conversation tasks used when testing WebRTC. The second part is planning and conducting of the study using the QoE-testbed. The third part of my research is an analysis of the data gathered in the experiment and then a discussion to get an understanding of whether and how the task influences the results from the study (i.e., the included self-report measures of QoE).

3.2.1 Literature Study

The literature study was performed to place the work in this dissertation in a context of research that has already been published. Firstly I received relevant papers and studies from my supervisor and later a snowball-search was conducted using Goolge Scholar[1] and IEEE [2].

Keywords listed in table 3.1 were put together with logical operators like OR and AND to build a query for narrowing the search for literature. Whenever literature satisfying a high level of quality was found, the reference list was further utilized as a source for exploring new and relevant research.

WebRTC	QoE
Conversation Task	Free Conversation
LEGO	Building Block Scenario
Video Conferencing	telemeeting
measuring QoE	QoE Influencing Factors

Table 3.1: Keywords used to narrow the search for literature.

3.2.2 Designing and Running the Experiment

The experiment was performed at NTNU, in a QoE-testlab that uses a research version of appear.in. Here, a short general overview is provided. In section 3.3.1, the set-up is described in more detail.

The experiment consisted of two people talking with each other in a controlled environment. The experiment was divided into two different conversation tasks, a LEGO-task and a Free Conversation. Each task was again divided into four different segments - or four different network conditions. After each condition, every test subject was asked to fill out a short questionnaire and rate the perceived QoE (in terms of different self-reported QoE measures) for each network condition. Every test participant performed both tasks under the different conditions. This allows to investigate how the participants rated QoE under the same four network conditions, but when conducting different conversation tasks.

3.2.3 Data Analysis

To analyze the data, the software package for statistical data analysis SPSS Statistics was used [34]. Depending on the type of analysis and the measurement level of

[1] https://scholar.google.no/
[2] https://ieeexplore.ieee.org/Xplore/home.jsp

the included variables, different statistical tests were performed. These are shortly introduced in chapter 4, which also presents the results.

3.3 Experimental Set-Up

First, the test procedure is briefly described. Next, the QoE testbed, the technical conditions, and the conversation task are discussed. Lastly, the included questionnaires are introduced, and a number of characteristics of the test participants are presented.

3.3.1 Test Procedure

Before the experiment started, each of the two participants, read the "Instruks til deltakere"(appendix A.3), filled out the pre-questionnaires (appendix A.2), and signed the consent form (appendix A.4). I performed the two-party video conversations in two separate offices. The WebRTC video conversation was established using the testbed, section 3.3.2. The testbed was remotely controlled from a separate location, and I was only present with the participants during the setup and collection of data. One of the participants had a chat window open throughout the whole experiment. Through the chat window, the test subject could notify when they were ready for another condition.

When the experiment started, the participants first performed a test condition. The purpose of the test condition was to let the participants get used to the environment, become familiar with the questionnaires, and clarify any misunderstandings. Afterward followed eight short sessions, either first four with LEGO-task and then four with Free Conversation, or the opposite. Whether the participants got the LEGO-task or the Free Conversation task first, was randomized in order to account for potential order effects. After each sessions, the participants filled out a short questionnaire, as described in section 3.3.5, to report their experienced affect, annoyance and perceived quality etc.

3.3.2 The QoE Testbed

For running the experiment, an existing testbed was used. The testbed topology is illustrated in fig. 3.1. It consists of a number of clients, a switch and a testbed controller. In my experiment, I used two clients. All clients are connected to the controller, the clients are not directly linked together, and all the traffic is redirected to the switch. All traffic from the clients will pass by the controller, and over the internet and eventually reach the appear.in test server. The controller is used to

remotely run the scripts by using Secure SHell Commands (SSH), SSH-commands [11].

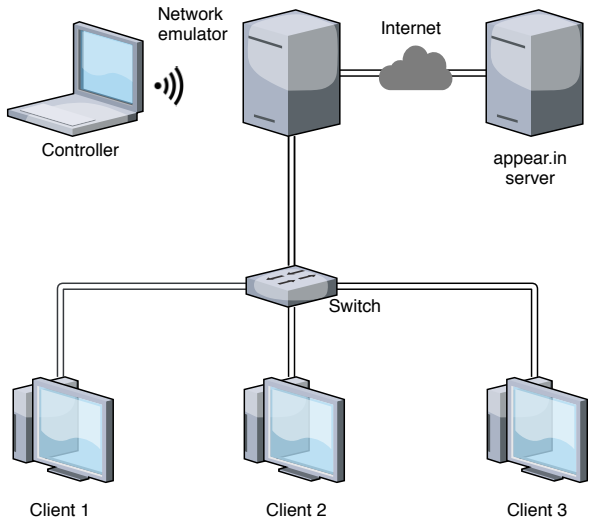


Figure 3.1: Topology of the QoE-testbed [4]

The network emulator controls the network conditions and provides network emulation functionality by emulating network impairments. It allows bandwidth throttling, adding delay and its variations; adding packet loss and emulate packet burst losses. The controller remotely controls the network conditions in real-time. The testbed uses a research version of appear.in [4].

All the desktop computers used in the testbed have the same specifications[11]:

- **Computer:** HP Compaq Elite 8100 SFF
- **Processor:** Intel® Core™i7 CPU 860 @ 2.80GHz x 4
- **Headphones:** Koss SB45
- **Web camera:** Microsoft LifeCam Studio

Screen Recordings

A screen recorder was used during the tests to record the video and audio during the experiment. The tool used for the screen recordings was *SimpleScreenRecorder*[3], a Linux program that is created to record programs and games.

[3]http://www.maartenbaert.be/simplescreenrecorder/

3.3.3 Conditions

As mentioned earlier, I used four different network conditions for the experiment. As this work aims to extend previous research conducted at NTNU, the conditions and general settings followed the conditions used in the paper "Exploring diverse measures for evaluating QoE in the context of WebRTC" [4]. These four conditions, being G: good quality, AV: distorted audio and video, A: distorted audio and V: distorted video, had been carefully pre-tested and the settings were selected to represent common technical distortions in telemeetings/video conferencing. The pre-testing assured that there were differences between the conditions and that the conditions never were so poor that the session could get disconnected.

The same conditions as in the study [4], were used in this project. As the length of an experiment is a non-negligible variable, it was not possible to include both the Free Conversation and LEGO-task, as well as the Celebrity Name Guessing Task in my experiment. The data I gathered can be compared with those gathered in the previous experiment, and makes it possible to investigate potential differences in the perceived QoE considering three different conversation tasks, namely the Celebrity Name Guessing task versus Free Conversation and LEGO-task.

Table 3.2 present an overview over the conditions used [4]:

Conditions	**Comments**
G: Good quality	No distortions
AV: Distorted audio and video	Packet delay of 500 ms and jitter of 300 ms
textbfA: Distorted audio	Limiting the CPU usage in the client side - of the WebRTC application to 70 %
textbfV: Distorted video	Packet loss ratio of 20%

Table 3.2: Overview over the Network Condition Used in the Study.

Each of the conditions lasted for 5 minutes, and the distortion was introduced after 30 seconds and continued through the whole conversation. The order of the conditions was randomized in order to avoid that the conditions gradually got better or worse. This to prevent the participants to assume that the condition of the current conversation is better or worse than the previous one. Also if the worst scenario (audio and video disorientation) comes right after the session with no network alterations, the user is more likely to give a worse rating because he will compare the two most recent conditions.

3.3.4 Conversation Task

Observations from the literature study was taken into consideration when picking suitable conversation task. The choice ended on two different types of scenarios, one focusing more on the video part and the other one on the audio part; LEGO-task and Free Conversation.

LEGO-task

The LEGO-task is described in ITU-T Rec. P.920 [18] where it is called "building blocks task". I have modified the building block task slightly, and the task looks more like the task used in "Towards context-aware interactive Quality of Experience evaluation for audiovisual multiparty conferencing" [35]. In the original "building block task" one of the participants had the instruction for building the LEGO, the other participant had all the LEGO bricks. The modified version encourage too more interaction and conversation between the test subjects, when both participant are more involved in the activity.

Each participant had their own LEGO building set, and half of the instructions for building the LEGO bricks. The goal of the task was to produce the same figure with help from the other subject and the instructions. Interaction with the other participants was, therefore, necessary to be able to build the whole model. The task also requires visual communication, so the participant can verify that they have build correctly.

Free Conversation

The Free Conversation task is described in ITU-T P.1301 [6]. The goal of the task is to get participants to chat and have a natural conversation. To facilitate the discussion, a paper with topic suggestions was handed out (appendix A.1). Although, the participants were still free to choose their own topic, the only limitation for the conversation was: do not talk about the environment and the conditions in the testlab.

3.3.5 Questionnaires

As mentioned, participants were asked to fill in a pre-session questionnaire and also a short questionnaire after every test condition.

Pre-session survey

The pre-session survey (appendix A.2) was handed out to the participant before the experiment started. The reason for this survey was to get an insight in which participants attended in the experiment, and which background they had.

Post-conversation survey

After each session participant was asked to rate the perceived quality using the ACR-scale. They were also asked to express their emotions on three dimensions *Valence, Arousal and Dominance* through the 9-point pictorial Self-Assessment Manikin or SAM-scale.

The post-session survey(appendix A.5) was inspired by the questionnaires given in the study [4] and included amongst others also the following dependent measure: *effort* that one had to put into the conversation (5-point scale ranging from "No special effort required" to "Very large effort required") and *talktime*: how they distributed the talk time between them (5-point scale ranging from "Only I talked" to "Only the other participant talked"). In the last questionnaire, for each task, the test subjects were asked to evaluate the conversation task (e.g., in terms of how engaging the task was). The questionnaire was translated into Norwegian, because all the participants had Norwegian as a native language.

3.3.6 Participants

To gather enough participants I asked people in my network to contribute to the study. The study consisted of 18 participants, both male (N=4) and female (N=14) aged from 20 to 27 years old (mean age 24,1). Most of the subjects were students at NTNU, and no one had previous experience with WebRTC-testing. All of the participants had normal, or close to normal, vision and hearing. How often they used services for online video chat varied from once a week to two-three times a week. Most of the test subjects had not been using appear.in lately, but a lot of them regularly used Facetime and Facebook-messenger.

It is an advantage that the conversation is as fluent and natural as possible. If uncomfortable silence occurs the participants may miss disturbances in the conditions[2]. To avoid this problem, I chose to put together people who knew each other, despite the recommendations to randomize the compositions of the personalities.

Extensive testing before officially starting the experiment, did not prevent that I ran into some technical problems. After completing 14 of the tests, the picture froze for one of the participants, even in the condition without any distortion. Hence only 16 of the 18 recruited persons conducted the experiment.

Chapter 4
Results

In this section, the results of the experiment are presented. A large number of comparisons were conducted; and I have chosen to present only graphs and data which represent a statistical difference. Although it is interesting with respect to the aims of this work; I also present results where there are *no* clear differences between conditions. The test condition is left out of the results, since this condition was only included to familiarize the participants with the the system and setting of the experiment. The results will be further discussed in chapter 5.

4.1 Short description of the statistical analyses

Statistical hypothesis testing is used to find out whether there is a connection between the variables of interest or not, and to rule out that the connection is due to "coincidence". We formulate a null hypothesis, H0; There is no significant difference (e.g., between two or more groups), and an alternative hypothesis, H1; There is a significant difference. Then we have to investigate whether the results are significant or not from a statistical point of view, using the appropriate statistical tests. We calculate a p-value; which refers to the probability that we observe an effect, given that the null hypothesis is true. The result are significant when the p-value $< \alpha$. If the calculated p-value is lower than the chosen significance level, often 0.05, we can conclude that the effect reflects the characteristics of the population. We can then reject the null hypothesis and gain confidence that the alternative hypothesis is true. If the result $p > 0.05$, the effect is not big enough to be significant. What we find may be a coincidence or the effect is too small to be detected. It does not necessarily mean that the null hypothesis is true, but implies that in the dataset, no evidence was found to support the claim that there is a significant difference [34].

The approach and statistical tests used to find eventual significant differences:

Perceived QoE in LEGO-task and Free Conversation, and the comparison between the two tasks:

- Friedman's ANOVA [34] is a non-parametric test for comparing several groups where the same participants have been used in all conditions.
- If the p-value is less than 0.05 the groups are significantly different, but the Friedman's ANOVA test does not say anything about in which groups the differences are situated.
- To detect between which groups the differences are situated, I perform post-hoc tests, in this case seperate Wilcoxon signed-rank tests [34], for the different comparisons. The Wilcoxon signed-rank test is a non-parametric test to compare two related groups (meaning that the same participants were involved in both conditions).
- In the Wilcoxon signed-rank test the comparison are significantly different if the p-value are below 0.05/number of comparisons.

Perceived QoE in Free Conversation versus Celebrity Name Guessing and LEGO-task versus Celebrity Name Guessing:

- In this case there are different participants for each condition, and we have to use the Kruskal Wallis test [34], which is a non-parametric test to compare two unrelated groups.
- If the p-values are less than 0.05, the groups are significantly different
- Since I have only two comparisons (e.g.: G condition for LEGO-task and G condition in Celebrity Name Guessing Task) there are no need for post-hoc tests, and p-values from Kruskal Wallis are used to determine significant differences between conditions.

4.1.1 Structure of the Chapter

First, I investigate how the test subjects rate the perceived QoE in the four different conditions within the same task. Then I will compare the two tasks in the experiment with each other. And finally, I compare the findings with the data from "Exploring diverse measures for evaluating QoE in the context of WebRTC" [4], who used Celebrity Name Guessing Task.

In the following, each subsection presents the results per QoE-measure (e.g., Overall Audiovisual Quality, Annoyance, and Arousal). These QoE measures refer to the dependent variables that were included in the Post-conversation survey, as explained in chapter 3. The QoE-measures are visualized in box plots and/or bar charts (displaying the mean ratings and 95% Confidence Intervals) and are briefly commented on in the text. Significant differences as identified with the statistical analyses, are represented

in tables. In addition, the mean value and standard deviation for selected QoE-measures in LEGO-task are represented in table 4.1. The mean value and standard deviation for selected QoE-measures in Free Conversation are given in table 4.5.

The different conditions are referred to as the G condition (good condition, no distortions), AV condition (audiovisual distortion), A condition (audio distortion) and V condition (video distortion).

4.2 Perceived QoE in LEGO-task

In this section, the QoE-ratings within the LEGO-task are presented. Mean values for QoE-ratings are presented in table 4.1.

Table 4.1: Representation of Mean Value and SD for QoE-measuers in LEGO-task. The scale is from 1-5 for the quality ratings(overall, audio and video) 1-9 for valence and arousal

	LEGO-Task							
	G		**AV**		**A**		**V**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Valence	6.94	1.181	5.13	1.857	6.88	1.708	6.19	1.223
Arousal	4.19	2.509	5.13	1.455	4.25	2.145	4.75	1.693
Overall Audiovisual Quality	4.06	.998	1.94	.929	3.56	1.263	2.75	1.000
Video Quality	4.13	.957	1.94	.929	3.63	1.408	2.31	1.195
Audio Quality	4.25	1.000	1.94	.574	3.94	1.340	3.00	1.095
Annoyance	1.63	.957	2.75	.931	1.56	.814	2.13	.806

4.2.1 Rated Overall Audiovisual Quality in LEGO-Task

Technical quality condition is an influencing factor when test participants rate overall audiovisual quality for the conversation. You can find p-values from post-hoc tests table 4.2

Conditions	p-value
AV - A	0.004
AV - G	0.001

Table 4.2: Conditions That Showed Significant Differences in Rated Overall Audiovisual Quality

Both the A and G condition have higher ratings than the AV condition: fig. 4.1. The lowest ratings for overall audiovisual quality were thus given in the condition in which both audio and video were distorted, as expected. The condition without any distortions have the best ratings. It can also be observed that the box plot for the A condition has a large spread, which indicates that there are disagreements between the test subjects about how good the overall audiovisual quality was in this condition.

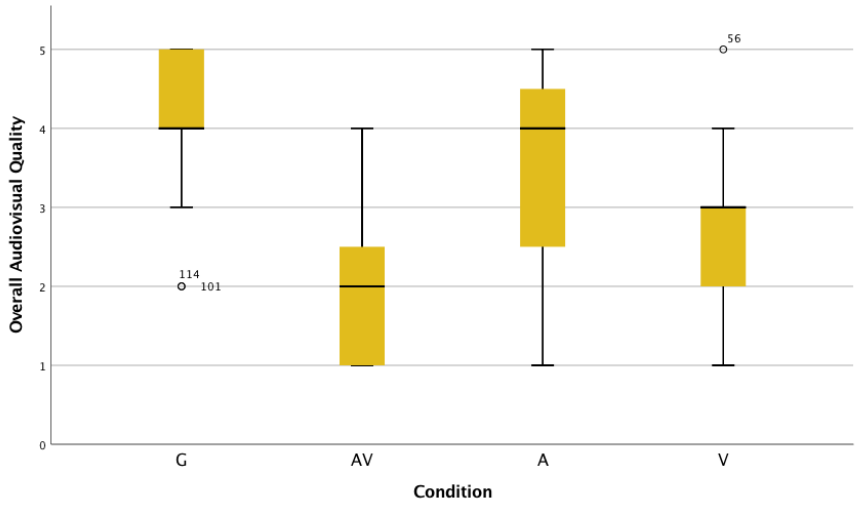


Figure 4.1: Box Plot of Rated Overall Audiovisual Quality in LEGO-task

4.2.2 Rated Audio Quality in LEGO-Task

Again, the technical quality condition is a influencing factor in QoE when test subjects rate audio quality in the LEGO-task. The Friedman ANOVA gives p= 0.000. The post-hoc tests showed significant differences between some of the conditions, represented in table 4.3.

Conditions	p-value
G - AV	0.000
V - AV	0.004
A - AV	0.001

Table 4.3: Conditions That Showed Significant Differences in Rated Audio Quality

A visualization of the rated audio quality in the LEGO-task can be seen in fig. 4.2. The ratings were lowest in the AV-condition (mean=1.94). The V-condition has the second lowest mean, 3.0. As expected, the condition without any distortion corresponds to the highest ratings for perceived audio quality (mean=4.25). The condition with distorted audio has a mean of 3.94. This is surprisingly high, better than the condition with only video distortion.

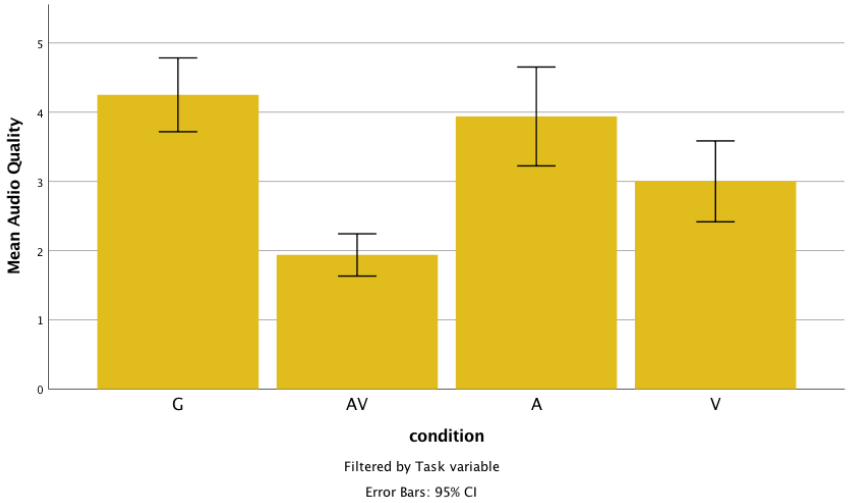


Figure 4.2: Rated Audio Quality in LEGO-task

4.2.3 Rated Video Quality in LEGO-Task

Both Friedman ANOVA (p= 0.000) and Wilcoxon signed-rank test (p-values in table 4.4) show significant differences for rated video quality.

Conditions	p-value
AV - G	0.001
V - G	0.005
A - AV	0.003

Table 4.4: Conditions That Showed Significant Differences in Rated Video Quality

Here we can observe that the same pairs that significantly differed from each other in terms of the audio quality ratings, also yield significant differences in terms of perceived video quality. A visualization of rated video quality in LEGO-task is displayed in fig. 4.3. Again, the AV condition and V condition have the lowest ratings. The best rating do we find in the G condition (mean=4.13 and median=4). Again the box plot for the A condition is rather tall, which is strange and unexpected since there are no video distortions included in this condition. Still, the mean value (3.63) and median (4) are high for A condition. This will be discussed further in chapter 5.

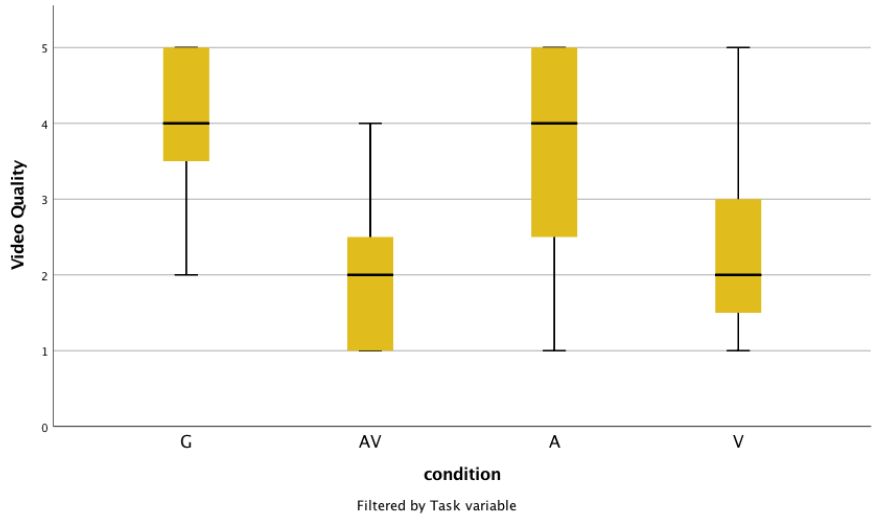


Figure 4.3: Box plot of Rated Video Quality in LEGO-task

4.2.4 Arousal and Valence in LEGO-Task

The Friedman ANOVA test did not yield any significant differences in arousal between the different conditions. The median is lowest for the condition without any distortion, indicating that participants felt most calm and relaxed when no distortions occurred. However, as we can see in figure 4.4, the box plots for especially good quality and the condition in which the audio was distorted are rather tall, indicating that the participants had rather varying opinions here. In the condition in which both audio and video were distorted, participants felt relatively aroused (and the short box plot indicates a rather high level of agreement between the participants).

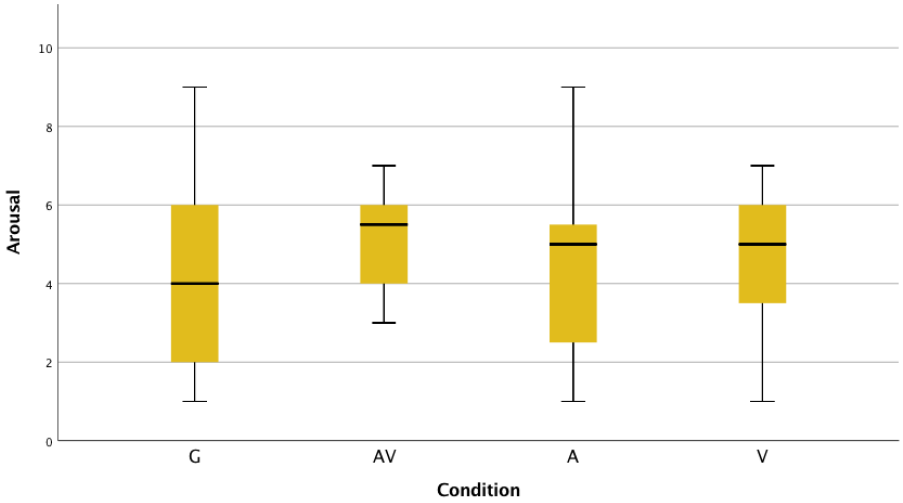


Figure 4.4: Box Plot of Rated Arousal in LEGO-task

Friedman Test shows a significant difference for valence in the LEGO-task, p=0.004. However, the Wilcoxon signed-rank tests shows no significant differences between the four conditions. Overall, valence is rated rather high in all the four conditions, shown in figure 4.5. The condition that is rated lowest is the AV condition, and the G, A and V conditions have almost the same ratings. These result indicates that overall, the test subjects were rather happy and pleased throughout the whole experiment.

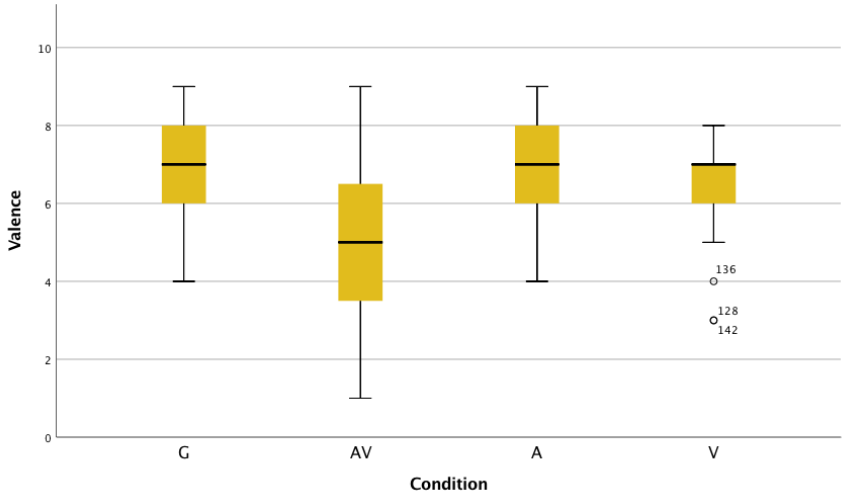


Figure 4.5: Box Plot of Rated Valence in LEGO-task

4.2.5 Rated Annoyance in LEGO-Task

Technical quality conditions clearly has an impact on the QoE-ratings for felt annoyance in the LEGO-task as well. Significant differences were identified through the Friedman ANOVA test (p=0.009). The Wilcoxon signed-rank test showed a significant difference in annoyance between the AV and the A condition (p=0.006), where the reported annoyance was significantly higher in the condition in which both audio and video were distorted. A visualization of felt annoyance in LEGO-task is shown in figure 4.6. Overall, annoyance is relatively low: the highest average annoyance is in the AV condition, with median=3 and mean value=2.75 on a scale to 5.

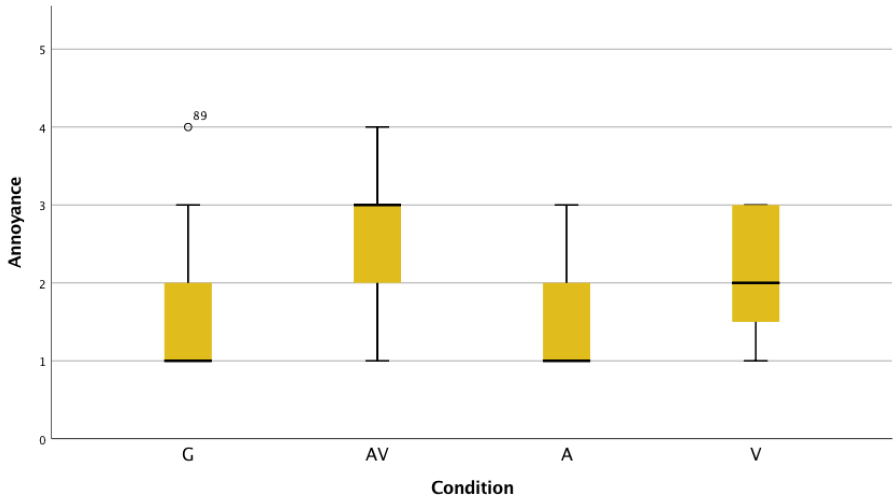


Figure 4.6: Box Plot of Rated Annoyance in LEGO-task

4.3 Perceived QoE in Free Conversation

In this section, the results of the different QoE-ratings from the Free Conversation task will be presented. Mean values for QoE-ratings are presented in table 4.5.

Table 4.5: Representation of Mean Value and SD for QoE-scores in Free Conversation. The scale is from 1-5 for the quality ratings(overall, audio and video) 1-9 for valence and arousal

	Free Conversation							
	G		**AV**		**A**		**V**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
Valence	7.00	1.512	4.88	1.996	7.19	1.167	6.50	1.414
Arousal	2.67	1.113	3.69	1.580	2.31	1.014	2.88	1.258
Overall Audiovisual Quality	4.33	.617	1.69	.602	4.31	.602	2.81	.911
Video Quality	4.40	.632	1.94	.772	4.31	.479	2.44	1.315
Audio Quality	4.47	.640	1.62	.619	4.56	.512	3.50	.894
Annoyance	1.00	.000	2.75	1.83	1.13	.342	1.50	.632

4.3.1 Rated Overall Audiovisual Quality in Free Conversation

There are significant differences between the different conditions in the QoE-ratings for audiovisual quality in the Free Conversation scenario; Friedman ANOVA shows p=0.000. The post-hoc tests show significant differences between some of the conditions; p-values are given in table 4.6.

Conditions	p-value
AV - G	0.000
AV - A	0.000
AV- V	0.000
V - G	0.001
V - A	0.000

Table 4.6: Conditions That Showed Significant Differences in Rated Overall Audiovisual Quality for Free Conversation

As expected, we can observe that the non-distorted condition was rated significantly better in terms of overall audiovisual quality than respectively in the AV and V conditions, see figure 4.7. The differences between the condition in which only audio was distorted (A) and the conditions in which either only video (V) or both audio and video were distorted (AV) are also significant, with the latter receiving significantly lower ratings for overall AV quality.

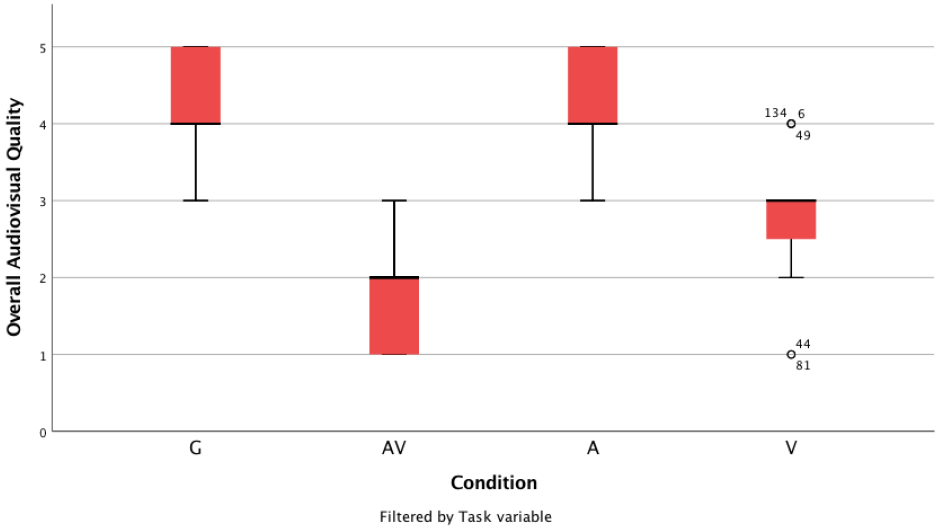


Figure 4.7: Box Plot of Rated Overall Audiovisual Quality in Free Conversation

4.3.2 Rated Overall Audio Quality in Free Conversation

Significant differences were found between conditions for rated audio quality in Free Conversation; Friedman ANOVA shows p=0.000. Wilcoxon signed-rank tests showed significant differences between almost all conditions; p-values are represented in table 4.7.

Conditions	p-value
AV - G	0.000
V - G	0.002
A - AV	0.000
AV - V	0.001
A - V	0.001

Table 4.7: Conditions That Showed Significant Differences in Rated Audio Quality for Free Conversation

The condition in which both audio and video were distorted is associated with the lowest ratings for audio quality and the audio quality is perceived as significantly worse than in the V and A conditions. The best ratings are found in the condition without any distortions and, unexpectedly in the condition where only the audio was distorted, figure 4.8.

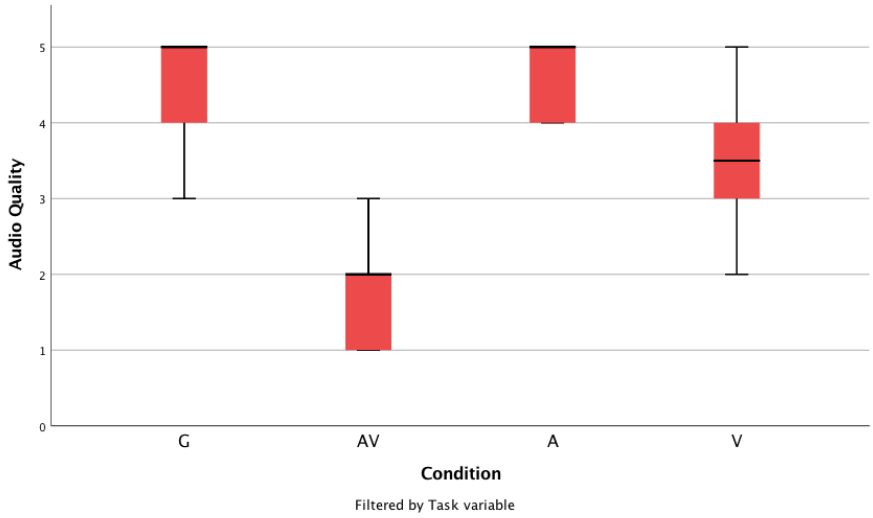


Figure 4.8: Box Plot of Rated Audio Quality in Free Conversation

4.3.3 Rated Overall Video Quality in Free Conversation

Friedman ANOVA (p= 0.000) and Wilcoxon signed-rank tests (p-values are presented in table 4.8) indicated significant differences between the conditions in terms of the perceived video quality in the Free Conversation task.

Conditions	p-value
AV - G	0.001
V - G	0.001
A - AV	0.000
A - V	0.002

Table 4.8: Conditions That Showed Significant Differences in Rated Video Quality for Free Conversation

Video quality is rated high in the G and A condition, conditions without video distortions. The V condition has a tall box plot, indicating a disagreement amongst the participants (in terms of perceived video quality).

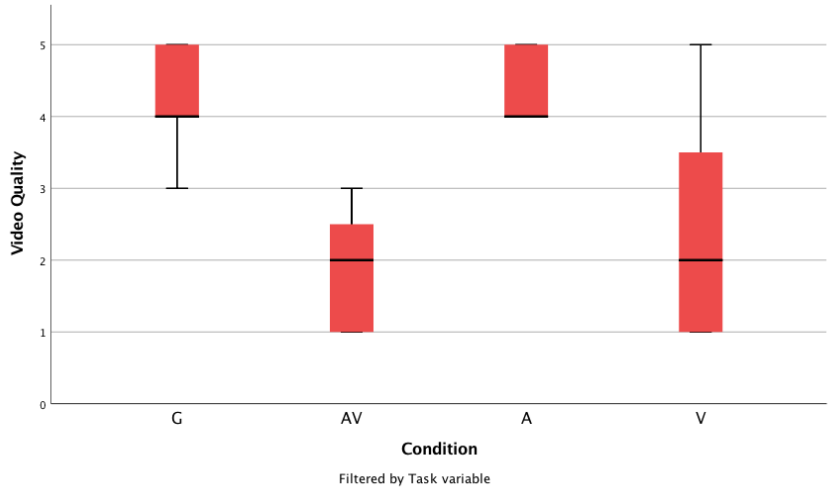


Figure 4.9: Box Plot of Rated Video Quality in Free Conversation

4.3.4 Arousal and Valence in Free Conversation

The self-reported arousal significantly differs depending on the test condition (Friedman ANOVA yields p= 0.007). But the post-hoc tests showed no significant differences between the conditions.

Overall, it can be observed that participants felt most aroused in the condition containing both audio and video distortion, figure 4.10. The box plots for the good condition, and respectively the conditions in which only audio or video were distorted, are rather compact and situated around the lower part of the scale, indicating that most participants felt rather calm/relaxed and only to a minor extent aroused in these conditions.

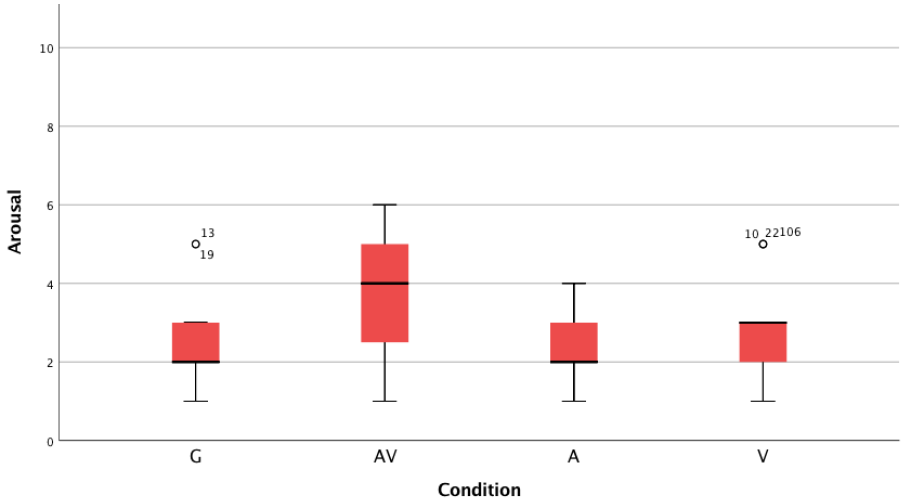


Figure 4.10: Box Plot of Rated Arousal in Free Conversation

For valence, significant differences can also be observed when comparing the different technical quality conditions (p=0.000). Wilcoxon signed-rank test shows significant differences between the conditions in table 4.9.

Conditions	**p-value**
AV - G	0.001
V - G	0.000
A - AV	0.002
V - AV	0.001

Table 4.9: Conditions That Showed Significant Differences in Felt valence for Free Conversation

The AV condition corresponds to the lowest ratings for valence, indicating that the participants felt more displeased when exposed to both audio and video distortions. The other three conditions have evenly good valence rating, fig. 4.11. This indicates

that the test subject were generally rather happy and pleased when conducting the Free Conversations under the different test conditions.

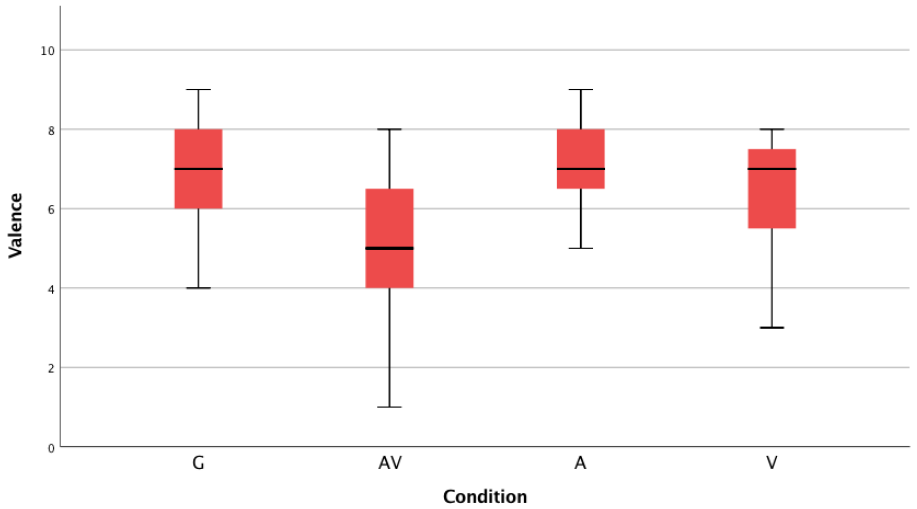


Figure 4.11: Box Plot of Rated Valence in Free Conversation

4.3.5 Rated Annoyance in Free Conversation

Finally, also the self-reported annoyance differs signficantly, depending on the technical quality conditions; p=0.000 in the Friedman ANOVA test. The results from the post-hoc tests are presented in table 4.10.

Conditions	p-value
AV - G	0.001
A - AV	0.002
V - AV	0.002

Table 4.10: Conditions That Showed Significant Differences in Rated Annoyance Quality for Free Conversation

We can clearly observe from the fig. 4.12 that the participants felt mostly annoyed in the AV condition. However, the annoyance level is not extremely high (mean value never higher than 3; in a scale to 5). Not surprisingly, lowest annoyance levels were reported in the condition without any distortions.

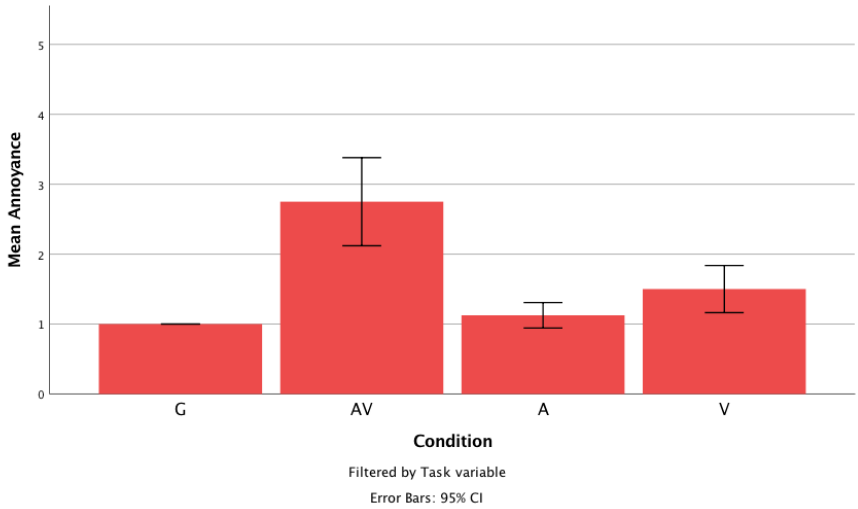


Figure 4.12: Rated Annoyance in Free Conversation

4.4 Perceived QoE in Free Conversation versus LEGO-task

In this section I am comparing the QoE-rating for the Free Conversation and the LEGO-task to look for differences in the rated QoE. As the same technical quality conditions were used in both parts of the experiment, such an analysis can allow us to investigate whether and how the conversation task itself may influence the quality ratings for the conversation.

4.4.1 Rated Overall Audiovisual Quality in LEGO-Task versus Free Conversation

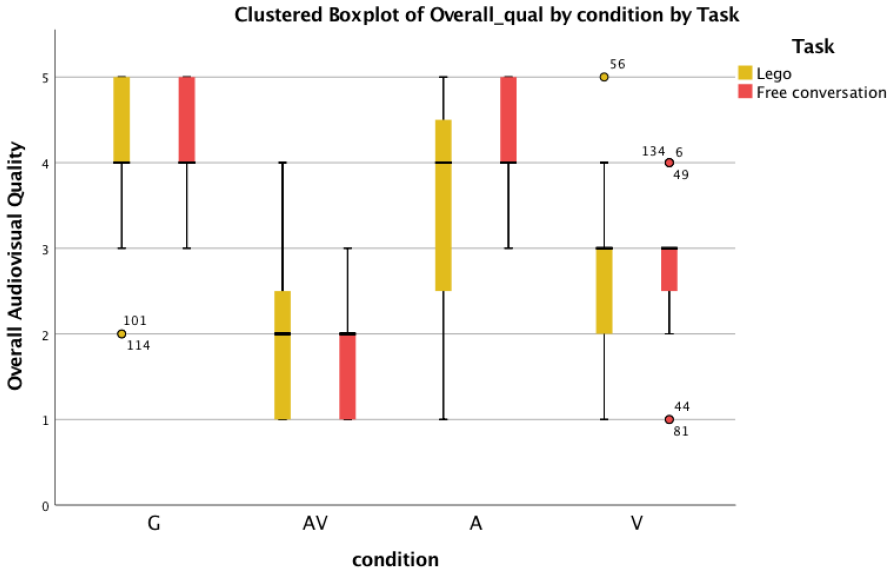


Figure 4.13: Box Plot of Rated Overall Audiovisual Quality in LEGO-task and Free Conversation

As can be observed in fig. 4.13, the overall audiovisual quality in LEGO-task and Free Conversation is rated quite similarly. The median is the same for every condition in the two tasks. For the AV condition, the overall quality ratings are slightly better for the LEGO-task. Interestingly, the LEGO-task has in all conditions a taller box plot with associated tail, and the box plots for Free Conversation are compact and short. No significant differences were found when running the Wilcoxon signed-rank tests on the respective pairs to compare. The p-values can be found in Appendix B.1.

4.4.2 Rated Audio Quality in LEGO-Task versus Free Conversation

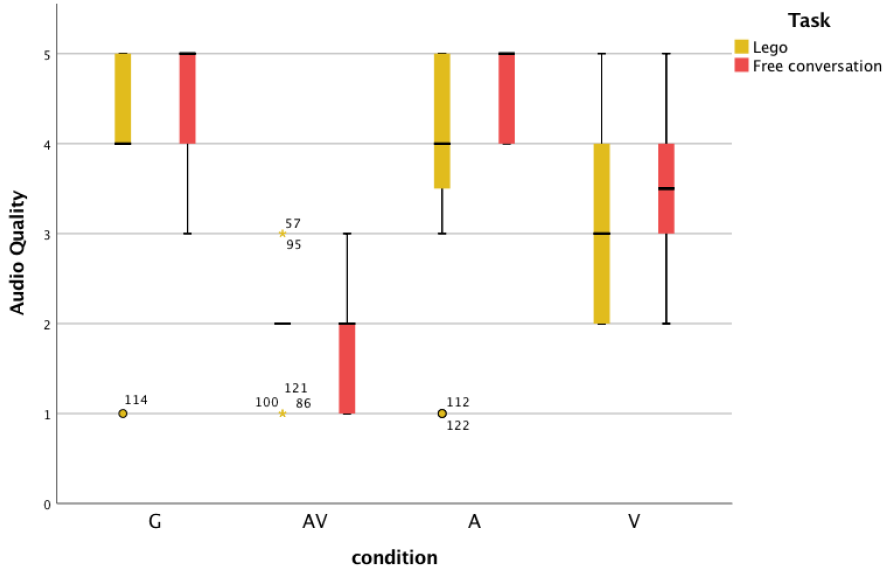


Figure 4.14: Box Plot of Rated Audio Quality in LEGO-task and Free Conversation

The rated audio quality for the Free Conversation and LEGO-task does not show any significant differences; p-values from the Wilcoxon signed-ranks tests are attached in Appendix B.2. From the box plots in 4.14 we can observe that the differences are rather small. However, the median for audio quality is different in G, A, and V condition. For all these conditions, there is a clear tendency that audio quality is rated higher in the Free Conversation task than in the LEGO-task.

4.4.3 Rated Video Quality in LEGO-Task versus Free Conversation

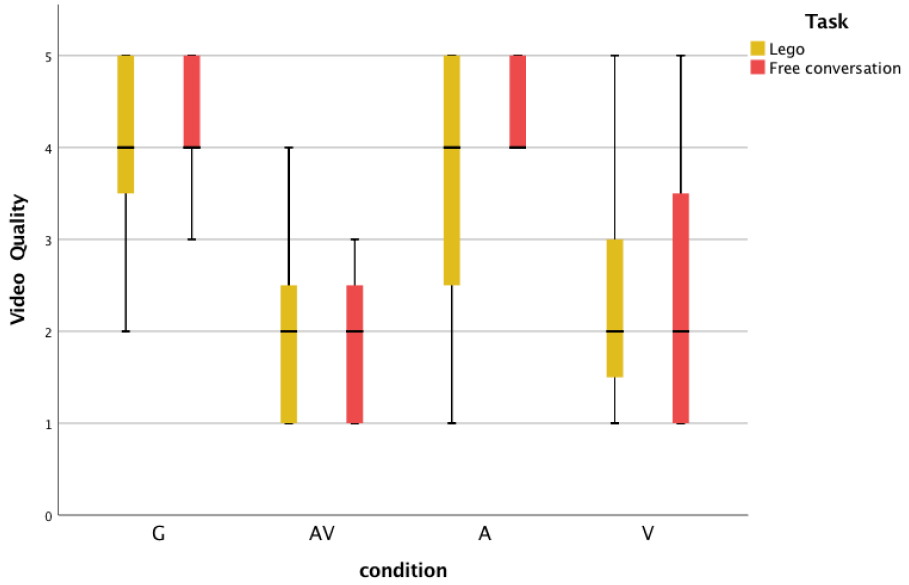


Figure 4.15: Box Plot of Rated Video Quality in LEGO-task and Free Conversation

The Wilcoxon Signed Ranks test did not yield any significant differences in rated video quality when comparing the different conditions for the Free Conversation and LEGO-task; p-values are added in Appendix B.3. There either no differences in the median for the conditions.

Despite the fact that no significantly differences were found in the statistical test, it seems that the participants rated the video quality of A condition generally lower for LEGO-task than Free Conversation. The box plot for this condition is also quite tall (there are disagreements among the participant of the video quality). For the good condition, the participants rated the video quality a bit lower than for Free Conversation. In addition, the video quality is rated better for Free Conversation than for LEGO-task when only the video is disrupted.

4.4.4 Felt Arousal and Valence in LEGO-Task versus Free Conversation

When comparing the reported arousal levels in the LEGO-task and Free Conversation, the Wilcoxon signed-rank shows significant differences; p-values are represented in table 4.11.

	FC-G/LEGO-G	FC-AV/LEGO-AV	FC-A/LEGO-A	FC-V/LEGO-V
p-value	0.004	0.003	0.003	0.002

Table 4.11: Conditions That Showed Significant Differences in Rated Annoyance for Free Conversation versus LEGO-task

Based on the ratings, it seems that the test participants felt more aroused in the LEGO-task conversation than in the Free Conversation. Especially for V and A condition the arousal is rated significantly higher for LEGO-task than for Free Conversation. Test subjects reported relatively high arousal in the AV condition, in relation to the other conditions for Free Conversation. Additionally, we can observe that the box plots for arousal under the LEGO-task are rather tall and have relatively long tails, indicating a larger variability in the ratings than for the Free Conversation task.

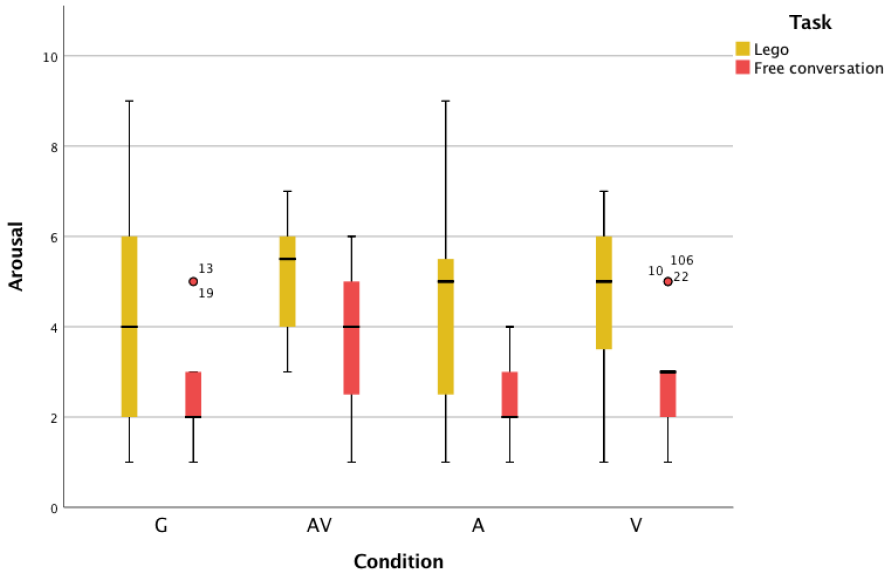


Figure 4.16: Box Plot of Rated Arousal in LEGO-task and Free Conversation

For rated valence, there were no significant differences between the LEGO-task and Free Conversation. The rated valence values follow each other in the different conditions, fig. 4.17. The AV condition is the condition with the lowest rating for both tasks. Valence is rated high for all the conditions, they all have a median over 5 on a scale to 10, meaning that the participants felt pleased throughout the whole experiment.

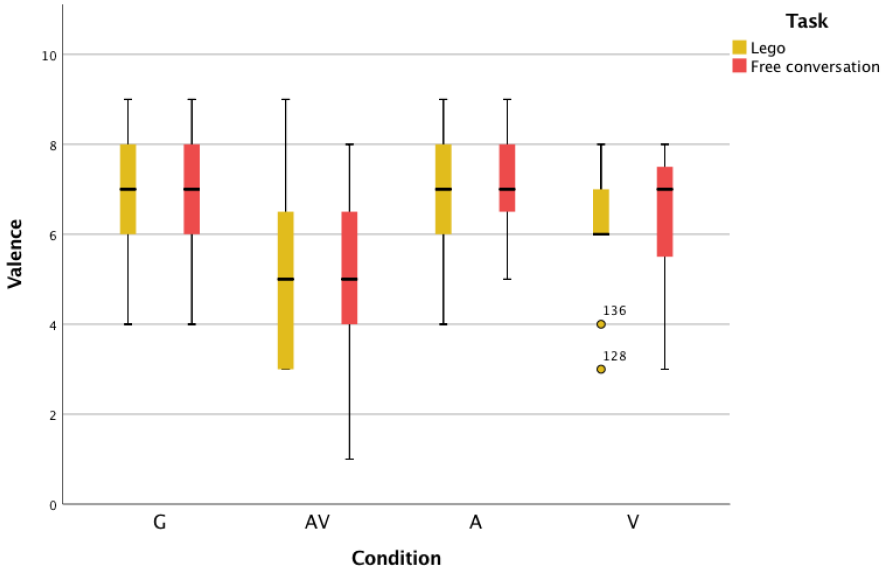


Figure 4.17: Box Plot of Rated Valance in LEGO-task and Free Conversation

4.4.5 Rated Annoyance in LEGO-Task versus Free Conversation

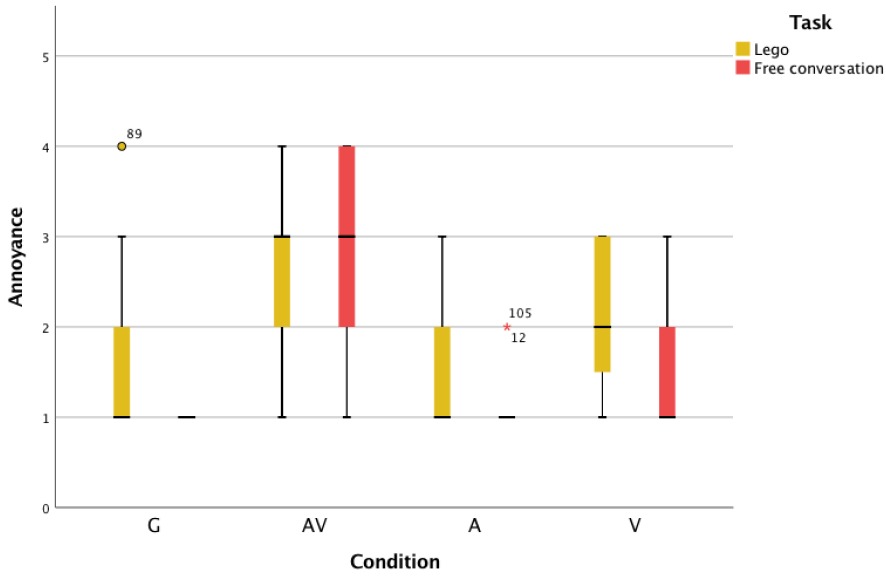


Figure 4.18: Box Plot of Rated Annoyance in LEGO-task and Free Conversation

There are no significant differences in rated annoyance for LEGO-task versus Free Conversation; p-values from the Wilcoxon signed-ranks test are attached in Appendix B.5. However, looking at fig. 4.18, we can still observe a number of differences in rated annoyance. Overall, the annoyance is rather low in all conditions. The LEGO-task yielded slightly higher annoyance ratings than the Free Conversation in all conditions, except for the AV condition.

4.4.6 Rated Focus on Screen in LEGO-Task versus Free Conversation

In order to discover the actual distortions as manipulated by the testbed, at least for the video impairments, it is important that test subject focus on the screen. For each of the two tasks I asked the participants: "To which extent did you focused on what happened on the screen?" and response options were:

- **5:** I saw everything that happened on the screen
- **4:** I saw most of what happened on the screen
- **3:** I saw neither much or little
- **2:** I saw almost nothing that happened on the screen
- **1:** I never saw on the screen

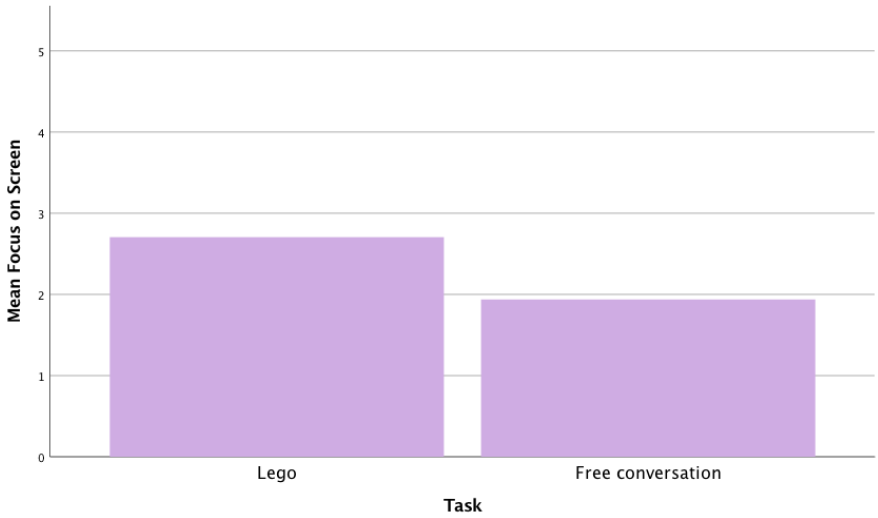


Figure 4.19: Focus on Screen in LEGO-task and Free Conversation

The results show that the test subjects focused more on the screen in the LEGO-task than in the Free Conversation. The Wilcoxon signed-ranks tests shows a significant difference in the self-reported focus on screen in LEGO-task versus Free Conversation; p=0.027. Figure 4.19 displays a plot of the ratings. When looking through the screen recordings (section 3.3.2) from the experiment, it is noticeably that several test subjects mostly look at the table in front of them and not at the other participant.

4.4.7 Rated Engagement of Task in LEGO-Task versus Free Conversation

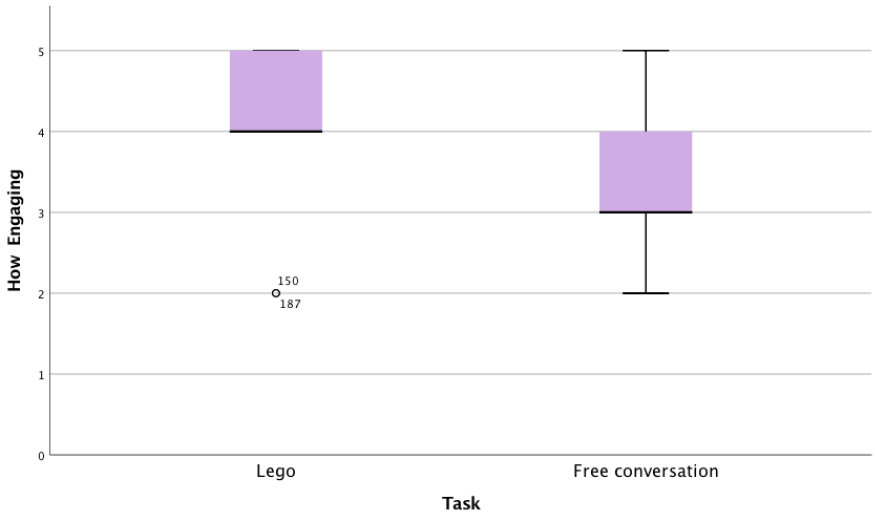


Figure 4.20: Rated Engagement of Task in LEGO-task and Free Conversation

As we can see from fig. 4.20 the LEGO-task is rated more engaging than Free Conversation. This observation is also supported in the statistical analysis, indicating a significant difference between both tasks in terms of engagement (p=0.004).

4.4.8 Task Order

Even though the task order was alternated in order to avoid potential order effects, I conducted a repeated-measures ANOVA for the most relevant dependent variables, using the four conditions and two tasks as within subject factors and the task order as co-variate. The results indicate that there is no evidence in the dataset that the task order had any impact on the ratings.

4.5 Correlations between dependent measures

I also conducted a correlation analysis to investigate the potential relation between the most relevant dependent variables. To this end, we used Spearman's rho [36], which is the appropriate test for evaluating the correlation between dependent variables at the ordinal level.

	Valence	Arousal	Overall Audiovisual Quality	Video Quality	Audio Quality	Effort	Annoyance
Valence	1	-1.88**	.498**	.396**	.475**	.449**	-.493**
Arousal	-1.88**	1	-.191**	-.135*	-.282**	.261**	.237**
Overall Audiovisual Quality	.498**	-.191**	1	.841**	.834**	.739**	-.733**
Video Quality	.396**	-.135*	.841**	1	.678**	.616**	-.591**
Audio Quality	.475**	-.282**	.834**	.678**	1	.775**	-.728**
Effort	-.449**	-.261**	-.739**	-.616**	-.775**	1	.698**
Annoyance	-.493**	.237**	-.733**	-.591**	-.728**	.698**	1

Table 4.12: Significant correlations between the most relevant dependent measures. (** means that p< 0.01 and * means that p< 0.05). The correlation coefficient can be interpreted as follows: r < .40: very low to low correlation (light grey); .40 < r < .70: moderate correlation (light blue) and r > .70 high correlation (violet)

Significant correlations are marked with * (**. means that p< 0.01 and *. means that p< 0.05). We use the guidelines by Guildford [36] to interpret the correlation coefficients.

The results indicate that perceived overall audiovisual quality correlates strongly and positively with the separate evaluations of respectively audio and video quality. There is also a strong negative correlation between on the one hand the perceived overall audiovisual, and audio quality and on the other hand, the effort put into the conversation and annoyance level. Put differently: lower quality ratings go hand in hand with higher annoyance ratings and higher efforts required to understand the other party during the conversation.

We also found a moderate positive correlation between the self-reported valence and the perceived overall audiovisual and audio quality: better perceived quality is thus associated with a more pleasant affective state. Not surprisingly, when more effort is needed to understand the other party and when a respondent feels more annoyed, the reported valence is significantly lower (negative and significant, yet moderate correlation).

4.6 Perceived QoE in Free Conversation, LEGO-task and Celebrity Name Guessing

As I have mentioned earlier the study reported in "Exploring diverse measures for evaluating QoE in the context of WebRTC" [4] used the same testbed and the same conditions as this study. I merged part of this original dataset with the data gathered in the experiments conducted as part of this master thesis in order to be able to compare some of the results. More concretely, I focused on a number of QoE-ratings (from the dependent variables overall audiovisual quality, audio and video quality, arousal and annoyance) from the test subjects, and used statistical hypothesis testing to find out whether there are significant differences in the ratings, depending on the conversation task used in the subjective test. Output from Kruskal-Wallis tests can be found in appendix C, and appendix D.

4.6.1 Rated Overall Audiovisual Quality in LEGO-Task, Free Conversation and Celebrity Name Guessing

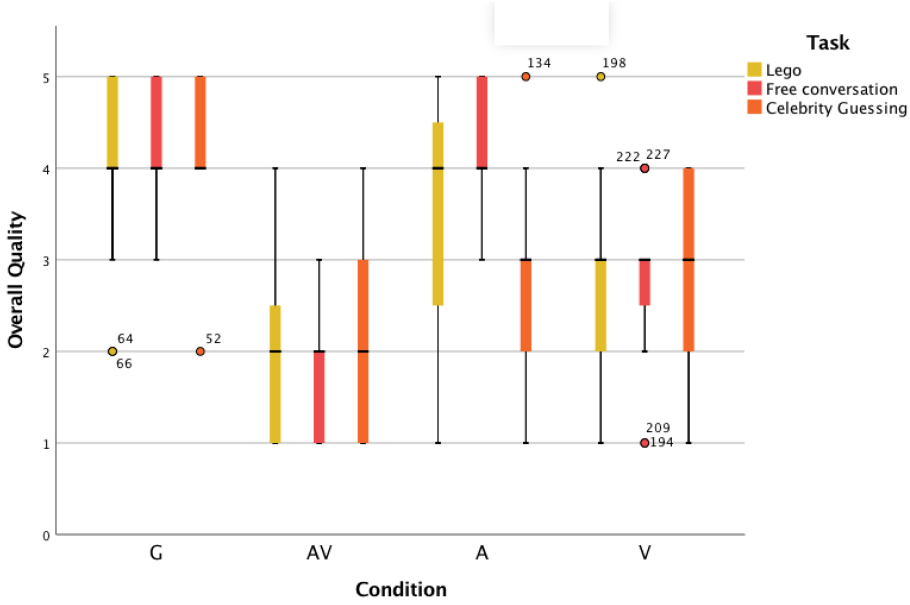


Figure 4.21: Box Plot of Rated Overall Audiovisual Quality in LEGO-Task, Free Conversation and Celebrity Name Guessing

The Kruskal-Wallis test shows significant differences in the rated overall audiovisual quality when comparing the Celebrity Name Guessing task to the Free Conversation

(p=0.000), but this only applies to the condition in which only the audio was distorted. Here, the overall audiovisual quality is rated lower in the Celebrity Name Guessing task, than in both LEGO-task and Free Conversation. In the rest of the condition the ratings are quite similar; all medians are the same. In all the conditions the LEGO-task and Celebrity Name Guessing have a taller box plot than Free Conversation.

4.6.2 Rated Audio and Video Quality in LEGO-Task, Free Conversation and Celebrity Name Guessing

The Kruskal-Wallis test shows significant differences in rated audio quality when comparing the Free Conversation task and the Celebrity Name Guessing task in the condition containing audio distortions. Also, the A condition for LEGO-task show significant differences (p-values in table 4.13)

	Free Conversation /Celebrity Name Guessing	LEGO-task /Celebrity Name Guessing
Condition	Q condition	A condition
p-value	0.000	0.002

Table 4.13: Conditions Who Showed Significant Differences in Rated Audio Quality for Free Conversation and Celebrity Name Guessing

In the AV condition the audio is rated all over quite low for all three tasks, the lowest ratings can be linked to the Celebrity Name Guessing task, fig. 4.22. In the A condition the ratings are much lower for the Celebrity Name Guessing task; the other two tasks have quite high ratings. Concerning V conditions, the Free Conversation and Celebrity Name Guessing task have about the same rating, Free Conversation on the other hand has widely spread answers in the ratings of audio quality, mainly from 2-4.

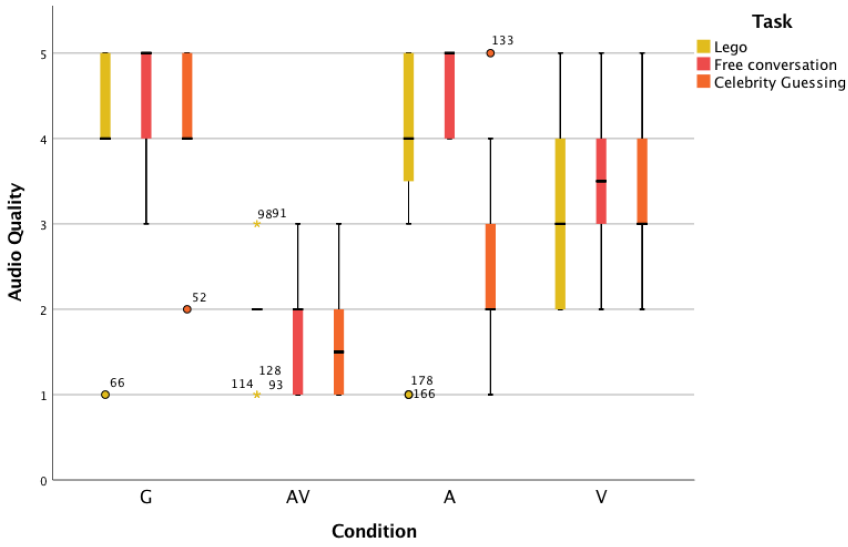


Figure 4.22: Box Plot of Rated Audio Quality in LEGO-Task, Free Conversation and Celebrity Name Guessing

When considering the perceived video quality, the only significant difference lies between the Celebrity-guessing task and the Free Conversation task, when considering the A condition (p=0.001). The video is rated much better in the Free Conversation than in the Celebrity Name Guessing task, fig. 4.23. Video is rated good in the G condition for all the three tasks. For the AV condition, the median for video quality rating is the same for the LEGO-task and Free Conversation, and slightly better for the Celebrity Name Guessing. Video quality ratings in the V conditions have a wide range, especially for Free Conversation, but the median is the same for all three tasks.

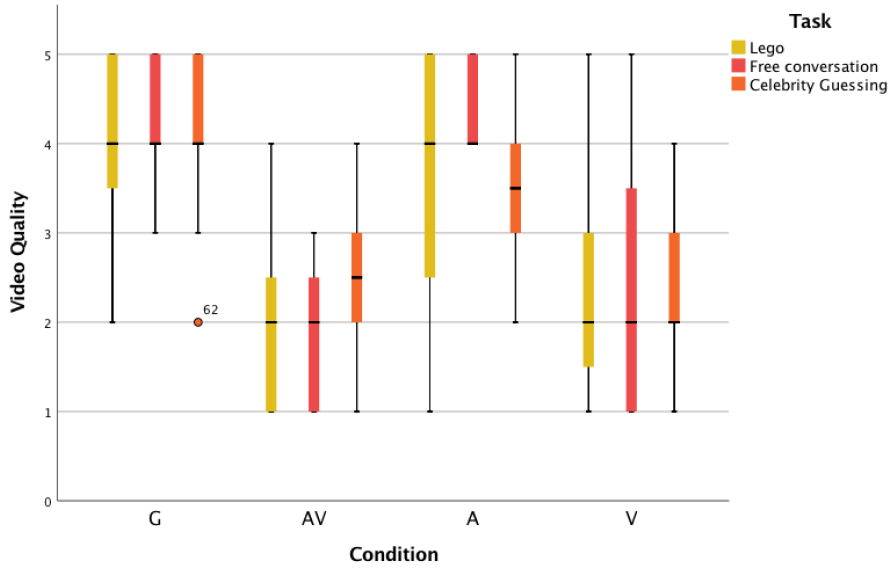


Figure 4.23: Box Plot of Rated Video Quality in LEGO-Task, Free Conversation and Celebrity Name Guessing

4.6.3 Felt Arousal and Valence in LEGO-Task, Free Conversation and Celebrity Name Guessing

Again, the Kruskal-Wallis test was used to investigate potential differences between the rated arousal in LEGO-Task, Free Conversation and Celebrity Name Guessing. Table 4.14 displays the conditions with the significant differences. There were no significant differences between LEGO-task and Celebrity Name Guessing. Figure 4.24 is a visualization of rated arousal for all three tasks.

	Free Conversation /Celebrity Name Guessing	Free Conversation /Celebrity Name Guessing	Free Conversation /Celebrity Name Guessing
Condition	G condition	A condition	V condition
p-value	0.002	0.000	0.0041

Table 4.14: Conditions Who Showed Significant Differences in Felt Arousal for Free Conversation and Celebrity Name Guessing

The test subjects generally felt minimal arousal in Free Conversation, also compared to the other tasks. The small range in the answers indicates that the participants overall are on the same line here. The Celebrity Name Guessing task is associated

with significantly higher levels of arousal in AV and in G as well. This was an unexpected result, and can probably be explained by the competitive character of the Celebrity Name Guessing task. The high ratings for arousal/excitement may be caused by the enthusiasm for solving the task, not the bad conditions. In addition this theory is supported by the high ratings in the G condition.

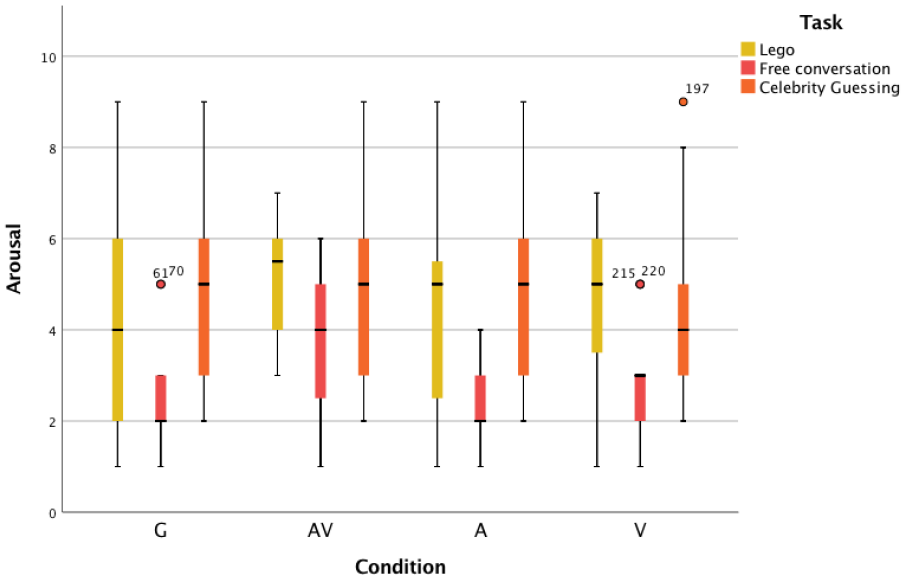


Figure 4.24: Box Plot of Rated Arousal for LEGO-Task, Free Conversation and Celebrity Name Guessing

Kruskal-Wallis tests showed neither significant differences for rated valence between LEGO-task and Celebrity Name Guessing, nor between Free Conversation and Celebrity Name Guessing task. Valence is rated relatively similar for all three tasks, and the ratings are high. The participants felt rather happy under all four conditions, even though there are some individual differences (for example in the AV condition, where the boxplot almost covers the whole spectrum of the scale).

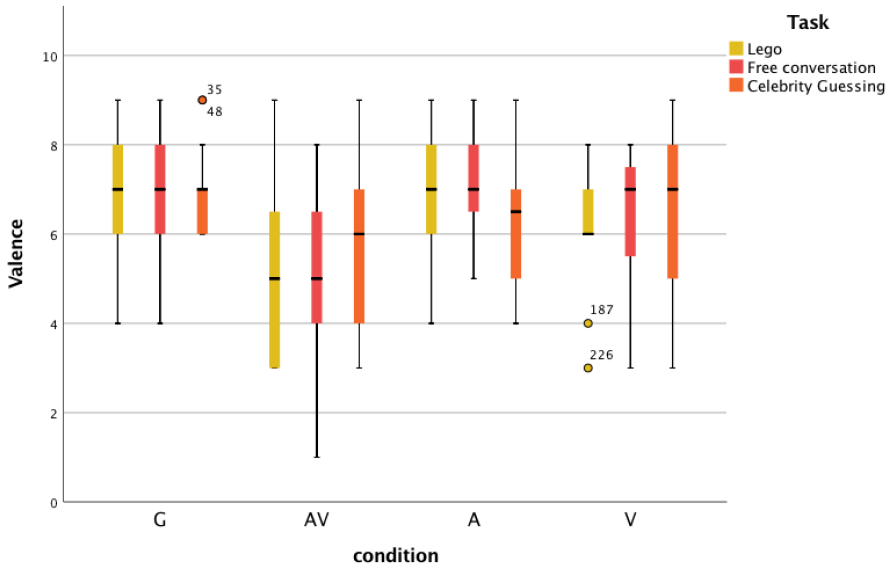


Figure 4.25: Box Plot of Rated Valence for LEGO-Task, Free Conversation and Celebrity Name Guessing

4.6.4 Rated Annoyance in LEGO-Task, Free Conversation and Celebrity Name Guessing

Kruskal-Wallis test shows a significant difference in rated annoyance between the Free Conversation and Celebrity Name Guessing task in the test condition without any distortions (p=0.08) and in the condition containing audio distortions (p=0.00). The LEGO-task shows a significant difference in the A condition, (p=0.004). In both cases, the annoyance level is significantly lower for the Free Conversation scenario.

It is clear that annoyance is highest in the condition with both audio and video are distorted, and this applies to all the three tasks. For AV and A condition the Celebrity Name Guessing task have the highest mean value. For G and V condition the LEGO-task has the highest ratings. Free Conversation has the lowest ratings for all conditions.

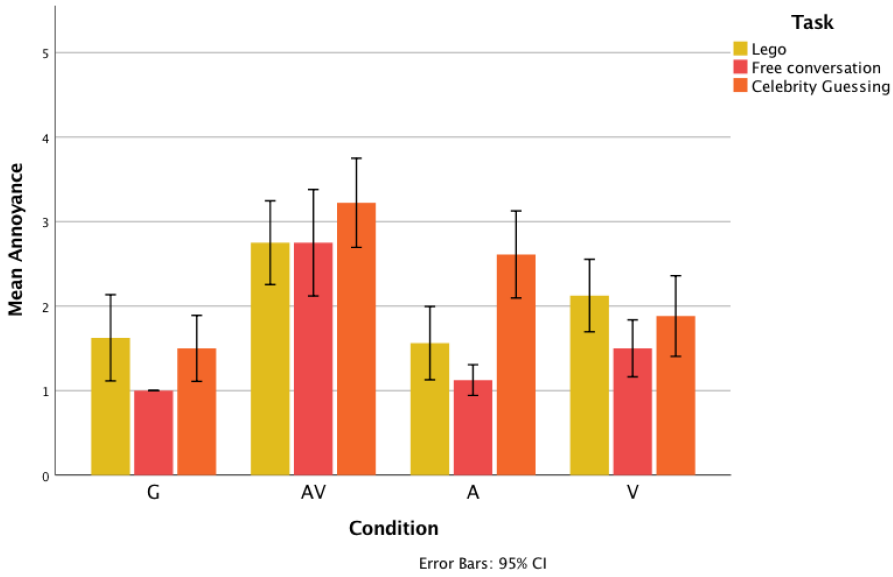


Figure 4.26: Rated Annoyance for LEGO-Task, Free Conversation and Celebrity Name Guessing

Chapter 5
Discussion

Chapter 3 described how I performed the experiment and chapter 4 presented the results from the experiment. This chapter further discusses the results and their implications, and puts them in a broader context of theory and previous relevant studies as presented in chapter 2. The last section is a discussion of the limitations of this research.

5.1 Impact of the technical quality condition on QoE within Free Conversation and LEGO-task

When considering the quality-related QoE-measures there are significant differences between the conditions in both Free Conversation and LEGO-task. As expected the condition without any distortions,G , was evaluated best in terms of audiovisual, audio and video quality in both tasks. On the other hand, the condition containing both audio and video distortions is, also as expected, associated with the lowest ratings for the quality-related measures, and this is the case in both the Free Conversation and LEGO-task sessions.

What is surprising with the results, is that the condition with respectively only audio distortions and only video distortions have such different quality ratings. The A condition yielded good ratings for perceived quality, almost as good as the condition without any distortions. This is surprising, as the relative importance of audio over video is generally underlined in research on telemeetings and video conferencing. Moreover, the high ratings for the A condition are more explicit and clear in Free Conversation than in LEGO-task, indicating that the audio distortions were to a lesser degree perceived when the test participants could just talk freely, without a concrete task at hand. When considering the LEGO-task especially, the box plots in the A condition are tall and have relatively long tails, for the quality-related QoE-measures. This indicates a larger variability in the ratings. Regardless, the differences between A-V and A-AV in terms of perceived quality were significant.

The unexpected high ratings and complications in the A conditions will be discussed in section 5.4.

When looking at the condition in which only video was distorted, we can conclude that this condition was perceived as slightly better than the AV-condition in terms of overall audiovisual quality, audio quality and video quality, and this applies to both tasks.

In [37, 25] increasing delay were introduced, although the mean MOS for rated overall quality were quite high. In the study presented in this thesis, the LEGO-task and Free Conversation clearly show a degradation in the average MOS for both AV and V condition, in terms of the quality-related QoE-measures. Mean values for LEGO-task and Free Conversation can be found in table 4.1 and table 4.5. However, the distortions introduced in the different studies vary (for instance, in the AV condition in this study, jitter was inserted), which makes it difficult to compare the numbers directly with each other.

When considering the more affective-state related QoE-measures, the results indicate that valence is generally high for all conditions in both the Free Conversation and LEGO-task, indicating that although respondents were exposed to different technical quality conditions, they felt rather happy and pleased. For both tasks, valence in the condition containing audiovisual distortions is lower than for the rest of the conditions.

Finally, in terms of the level of annoyance, small differences can be observed between the conditions in LEGO-task and the respondents felt mostly annoyed in the AV condition. In the Free Conversation task, a significant difference was identified: The rated annoyance in the AV condition is significantly higher than in the other conditions. However, it should be observed that the annoyance level is relatively low.

5.2 Impact of the task on QoE? Comparing Free Conversation and LEGO-task

In this section, the focus is more explicitly on differences between both conversation tasks. Such a comparison helps to gain a better insight into potential differences in terms of user tolerance towards certain impairments, influenced by the nature of the task. As mentioned in the previous section, a general observation worth noticing is that there is more variability in the ratings for the quality-related QoE measures in the LEGO-task, as indicated by the more extended box plots.

This is not the case of the ratings of the Free Conversation scenario, which are more compact. The conditions are the same, the test subjects are the same, and the only difference is the given task. The difference can indicate that additional, more implicit influence factors here affected the quality ratings less than in the LEGO-task.

The tables 5.1 and 5.2 are visualizations of in which QoE-measures I found significant differences between the conditions, respectively for LEGO-task and Free Conversation. Free Conversation shows clearer differences in the statistical analyses between the conditions, within the task.

Table 5.1: Overview over in which different conditions in the LEGO-task there is a significant difference

	LEGO-task					
	G-AV	G-A	G-V	AV-A	AV-V	A-V
Overall Audiovisual Quality	■			■		
Audio Quality	■			■	■	
Video Quality	■		■	■		
Arousal						
Valance						
Annoyance				■		

Table 5.2: Overview over in which different conditions in the Free Conversation there is a significant difference

	Free Conversation					
	G-AV	G-A	G-V	AV-A	AV-V	A-V
Overall Audiovisual Quality	■		■	■	■	■
Audio Quality	■		■	■	■	■
Video Quality	■		■	■		■
Arousal						
Valance	■		■	■	■	
Annoyance	■			■	■	

The study reported in the paper "Audio and Video Channel Impact on Perceived Audio-visual Quality in Different Interactive Contexts" [27], compared two different conversation tasks; a Short Conversation Task (SCT) and the building block scenario. Their findings address that the participants rated the audio quality better in the SCT than in the building block scenario. Although I do not have significant differences in rated audio quality for Free Conversation and LEGO-task, I can see the same tendencies: The median is lower for all conditions in LEGO-task, except in the AV condition. This indicates that other factors have affected the LEGO-task ratings. In the Free Conversation, the test subjects talk and listen to each other, but in the LEGO-task many other aspects and actions may influence the outcome: The engaging character of the task and engagement level of the participants, the experienced arousal, and the written materials.

My statistical analysis shows significant differences in terms of the reported engagement between the Free Conversation and the LEGO-task. In this case, the box plot for the LEGO-task is very compact and high, while the box plot for the Free Conversation is long with a tail, which indicates disagreement between the participants. Feedback from some of the participants further supports this result; Many of the test subjects expressed after the experiment was completed that "The LEGO-task was really fun". On one side, tasks such as this are valuable as they can make subjective tests more engaging and enjoyable for the test subjects, opposed to using very unnatural or boring tasks, for instance scripted conversations. In the literature [38], it has been pointed out that tasks should preferably have a design that keeps the attention level high throughout the test, and that there is a need for more immersive and engaging test paradigms. On the other hand, it is clear that such tasks may introduce additional challenges (especially when they are compared to "traditional" perceived quality measures) and require additional measures. Put differently, it is essential that not only feedback on quality-related issues but also on

engagement- and affective state related aspects are collected.

As mentioned in section 2.5 and in the studies [4, 25] engagement of a task might also affect the test result and the users' (in)tolerance towards specific impairments. The findings in this thesis indicate no clear confirmation on whether this is correct or not. However, in the overall quality ratings, the ratings in LEGO-task for the AV and the V conditions are slightly higher than in the Free Conversation, which points in the direction that the choice of task matters and participants have a higher tolerance for delay in engaging tasks.

In terms of felt arousal, significant differences between the two tasks could be observed in all the conditions. The participants indicated that they felt more aroused in the LEGO-task. This could explain why the quality ratings in the LEGO-task have a wider range; their overall judgment may (unconsciously) be affected by other influence factors and have less focus on the quality degradation.

For valence, it can be noted that in all conditions and for both Free Conversation and LEGO-task, the ratings are altogether rather high, with a median of over 5 for all conditions. The correlation analysis indicated a positive, moderate correlation between valence and overall audiovisual quality. This means that test subjects who rate the quality higher also report a higher valence-level and vice versa. In addition higher rated annoyance go hand in hand with lower valence ratings. However, the overall rated annoyance level is low; almost no responses higher than 3 (moderate annoyance). Even though the more detailed analysis of the correlations amongst different subjective QoE-measures included in this thesis, goes beyond the scope of this work, it is clear that quality ratings go beyond the purely perceptual level and that the affective dimension in this respect cannot be ignored. This is also in line with the state of the art. As cited in section 2.2 ITU changed their definition of QoE in 2017, into a definition which includes the words delight and annoyance. Several studies have therefore begun to include also affective state-related QoE measures into subjective testing and user studies. The obtained findings from the correlation analysis also indicate that these affective dimensions do correspond to rated quality for the application.

The term "attention" was coined earlier when discussing the engagement level associated with both tasks. The findings related to participants' self-reported focus on the screen are also relevant in this respect. Rated focus on screen is significantly higher in LEGO-task than in Free Conversation. From this result, we can assume that when the test subjects have a specific task requiring information from the video stream in order to be able to complete it, they have more focus on the screen, compared to having a normal conversation. An expected consequence of the focus ratings is that the video quality ratings are lower in the LEGO-task when the video is disrupted. In

the Free Conversation task, users can rely on the audio, which is good. When both audio and video are distorted, the ratings are slightly lower for the Free Conversation than in the LEGO-task, as can be observed in figure 4.15. However, the difference is not substantial and not significant from a statistical point of view.

5.3 Comparison with previous study: Celebrity Name Guessing Task

Finally, the results from the comparison with the third task, Celebrity Name Guessing, used in the preceding study [4] are discussed. In general, there are no major differences between the three tasks in terms of QoE, when considering overall quality, audio quality, and video quality. The main exception is the condition containing audio distortions. This may be another confirmation that something something has affected the A condition, since the ratings are much lower in Celebrity Name Guessing than in the other two conditions and the only statistically significant differences between the tasks are situated in the A condition. Otherwise, the box plots for Celebrity Name Guessing, in the rating of overall audiovisual quality, are even wider than the box plots for LEGO-task, which suggests a potentially even bigger influence from other (e.g., affective or context-related) influence factors.

For rated audio quality in each respectively task, the Celebrity Name Guessing values are close to the values in LEGO-task, and the ratings of audio are lower than in Free Conversation. Again, such as in the previous mentioned study [27], audio quality got better ratings when the subject's attention were focused on something else than just having a conversation. This study states that the lower audio scores in the Free Conversation come as a consequence of the focus on the screen. Even though the Celebrity Name Guessing task has minimal focus on video, my findings can indicate that character of the task, level of engagement, and other context influence factors may have more impact on the audio ratings. Perhaps the test subjects do not notice audio distortions as easy in engaging tasks? Closer investigation of this can be addressed in future work.

When considering to which extent test participants felt aroused (ranging from very calm to very aroused, excited), the Celebrity Name Guessing task yields significantly higher ratings for arousal than the other tasks. When I presented the result in chapter 4, I argued that these high ratings may be interpreted as a result from a strong competitive spirit, and in a task typically organized like a kind of competition. Even though the test subjects clearly find the LEGO-task engaging, there was no competition element that could lead to increased involvement. The impact of a strong competitive element may also help to explain the higher ratings for annoyance in Celebrity Name Guessing (same for the AV and A condition, annoyance is higher for the Celebrity Name Guessing task than for the two other tasks). The findings in

[4], also indicated that a distortion of both audio and video leads to lowest quality ratings and highest annoyance.

Nevertheless the valence is overall relatively high in all three tasks, for all four conditions. Even though the test subjects felt annoyed and were aroused, they were also roughly happy. An explanation for these seemingly paradoxical ratings could be that the feelings of annoyance and frustration are primarily triggered by bad technical quality (which is also illustrated by the strong negative correlation between perceived quality and annoyance), while the technical quality is not necessarily detrimental for the test subjects' positive affect (as expressed by the valence measure). We could notice that there is a moderate positive correlation between valence and overall audiovisual quality, but it is not as strong as the correlation with annoyance. As discussed earlier, here, it seems that other aspects (e.g., engaging character of the task) to some extent "make up for" or "compensate" for the negative impact of the technical distortions on the overall experience. As was also stated in [4]: quality degradation's does not necessarily go hand in hand with low valence, even though to some extent, valence is affected. [4] found significant differences between G and AV condition; The valence was significantly higher in the experimental condition with no distortion than in the condition with both audio and video distortion. The same trends can also be seen in my findings. The valance ratings are overall rather similar across the experimental conditions for all three tasks, for all experimental conditions, and seem to be more affected by other factors [4].

Thus, the results of my study confirm earlier findings to a certain extent and above all indicate that more research is needed to better understand the role of the task and the relation between different types of technical impairments, technical quality as experienced by test users, and more affective state-, engagement-, and attention-related measures of QoE.

5.4 Limitations

In this section, the limitations of the experiment and analysis will be presented.

First of all, during the time I conducted the experiment, despite extensive pre-testing before the actual start of the experiment, some technical problems occurred. Some of these were most likely due to power outages in the building where the servers were located, but for others, it is unclear what the root causes were. As mentioned in chapter 3 the video for one of the participants froze, and two answers had to be taken out of the dataset. In addition, after analyzing the results, I have suspicions that something has affected the A condition. The script that was used, was the same as in the previous study and it was pre-tested before the experiment. Yet, the results show that generally, the quality ratings for the condition with audio distortions are much

higher than expected. In some cases, they look unaffected by technical distortions that should have been there. To try to understand what the problem may have been, I examined some of the screen recordings for the A condition and observed that the audiovisual distortion was not as noticeable as it should have been. However, the ratings are not only good, so seemingly not all sessions have been affected in the same way. In any case, this issue puts constraints on the data obtained for the condition containing audio distortions, and further investigation of what happened or potentially even a replication of the study are left for future work.

Secondly, the number of participants who conducted the experiment is relatively small, which leads to uncertainty in the results and makes it hard to draw strong conclusions based on the material. However, I followed the guidelines by ITU[5, 6] as firmly as possible and the results can therefore be used as an indication. The results are to a large extent in line with earlier work. Furthermore, some of the indications and results can help to point out the direction for future research efforts in this area.

In addition, it should be acknowledged that using questionnaires as data collection method is valuable and has many advantages, but also has its limitations. The accuracy and honesty of people's responses cannot be verified. Since the questionnaires were given on paper, it may also have occurred typos when digitized the numbers. By means of thorough data cleaning and random checks, I tried to minimize this risk as far as I could. Further, it is also a challenge that the answers will be subjective for each participant, since QoE is individual for a person in a specific situation.

Another limitation has to do with the lack of entirely suitable testing spaces for subjective experiments and with the manner the experiment was run. Even though the experiment was conducted in (offices turned into) testlabs and I controlled the technical qualities for the conversation, other system factors and context factors may have affected the result. The lab setting does not always create a natural setting [32], and some test subjects can respond with unnatural behavior in such settings.

It could also be argued that the LEGO-task, even though it clearly engaged the test participants, does not reflect a natural and more realistic task. Still, the use of video chat increases in popularity and is used for many different purposes and settings, e.g., in computer games, for homework. In this way a collaborative tasks as the LEGO-task may still contain characteristics of realness.

Finally, in section 2.3, I have listed many impact factors for quality ratings; the personality of the participants, how neutral the location was, the familiarity of the participant, setting of the call, etc. The focus in this thesis was primarily on the impact of the conversation task, which made the focus on the other influence factors less. But I tried to follow the recommendation from ITU [6] as far as possible and hopefully this work can be continued with follow-up studies in which the primary focus is extended to some of these potentially confounding factors.

Chapter 6
Conclusion and Future Work

6.1 Conclusion

The main objective underlying this work was to investigate whether and how the task which is used in subjective testing of video conferencing QoE influences QoE-measures under different technical circumstances.

Even though many studies have been conducted to get an understanding of which factors may play a role in this respect for QoE, there are still a lot of open questions. Previous studies have put participants in an experimental environment and given them a conversation task to simulate a natural conversation under different technical conditions. However, it is still poorly understood which task is most suitable and how the task in itself may actually be a confounding factor influencing the impact of technical distortions on QoE. This thesis aims to make a contribution to the literature in this respect. I conducted an experiment to investigate how the conversation task influences QoE under different circumstances. Using a QoE testbed, participants were exposed to four different technical quality conditions (good quality, distorted audio and video, distorted audio and distorted video). Two conversation tasks were included in the experimental design, namely the LEGO-task (building blocks task) and the Free Conversation task. Data gathered in a previous study, using the Celebrity Name Guessing task was also integrated into the dataset, in order to compare the findings.

The results point in the direction that the task *does matter* and *does influence* QoE. However, it is important to way the possibilities when it comes to what type of QoE-measure that is used (traditional perceived quality measure or more affective state-oriented QoE-measures). When it comes to the traditional perceived quality ratings, there are significant differences when comparing the different conditions *within* each task, but not *between* the different tasks (except for the audio-distorted condition when comparing with the Celebrity Name Guessing task). Still, I have found some interesting tendencies. Overall, the quality ratings are slightly higher for the Free Conversation task than for the other two tasks (LEGO-task and Celebrity Name

Guessing). In addition, I have observed that in most cases, the Free Conversation yields more compact ratings with lower variability among the test subjects. This could also be observed for most of the other subjective measures. Previous research [27] indicates that the LEGO-task is more suitable for testing video degradations' impact on QoE. Even though in this study there are some tendencies in that direction (e.g., annoyance is clearly higher in the condition in which video is distorted when comparing the LEGO-task to the Free Conversation task), I found no hard evidence in the dataset to support the this claim.

When it comes to the affective state, I also found some interesting differences. The test subjects clearly feel the least aroused under the Free Conversation task and more aroused in the other two tasks. The highest arousal levels are found in the Celebrity Name Guessing task (however, as mentioned, the variability in the ratings is larger). A plausible explanation in this respect is that the inherently competition element plays a role. The annoyance is mostly low, except in the AV conditions. Still there are some interesting tendencies that the annoyance is generally higher in the Celebrity Name Guessing task and the LEGO-task, which can in a way be interpreted as an indirect indication of lower tolerance towards impairments in these two tasks. The relatively lower annoyance level in the V condition for Free Conversation can be explained by the the self-reported focus on the screen, which is significantly higher in the LEGO-task than in the Free Conversation task. Focus on the screen is directly linked to the attention level, so this is important to consider when selecting a task.

In terms of how engaging the tasks are perceived to be, there are also clear differences. The Free Conversation yields significantly lower (but wider) ratings for how engaging the task was. This can indicate that the task was perceived boring. For the LEGO-task, it is clear that the participants felt more engaged, and this could help to make WebRTC-testing more interesting. Introducing the "game aspect" of the task conversation has a substantial impact on the emotional state and felt annoyance for the test subjects, but at the same time the engagement contributes to higher attention level throughout the experiment, as cited in section 5.2.

The findings also open interesting questions related to the type of measures that are most suitable. The results indicate that bad quality (through different types of distortions) goes hand in hand with feelings of annoyance and frustration. This impact is clear and to a certain extent mediated by the task. However, technical quality and perceived technical quality are not detrimental for positive affect (as expressed with the valence ratings in this study). Valence is overall rather high and there are no significant differences. On the other hand, there is a moderate relation between perceived quality and level of valence, but it is not as strong as the negative relation between annoyance and perceived quality. This indicates that other aspects come into the picture and the characteristics of the task, level of engagement

are definitely important to further explore. A broader implication of the findings presented, is therefore also what the most meaningful measures of QoE really are. It is clear that perceived quality ratings (the measures that are traditionally used) only show half of the picture and that it is necessary to complement them with other measures that allow to gain insights into how technical factors impact how people feel.

To summarize, the choice of conversation task is complex, and many factors have to be taken into consideration. There are advantages and disadvantages with both tasks and some differences are clear, while others are more implicit. In general, the Free Conversation task provides higher agreement of the quality ratings, but is clearly less engaging. The LEGO-task and Celebrity Name Guessing task are more engaging and implicitly, the tolerance towards impairments seems to be lower. But more implicit influence factors affect the quality ratings and their role needs to be better understood.

6.2 Suggestions for Future Work

The result in section 5.4 shows that ratings for the A condition were much higher than expected in most cases. For future research it would be interesting to dig deeper into what caused these high ratings and to replicate the experiment.

As cited in section 5.4 the test lab can create an unnatural setting. Further research is needed to validate the current findings and to investigate to which extent impact of the conversation task is also prominent in natural and more realistic settings, outside the lab. The LEGO-task may represent a natural setting either. However, video conversation increases in popularity and the QoE for new areas are not explored. We need to find a threshold for the user in all different types of scenarios. Related to this, there is also a need to develop and experiment with new types of tasks that represent common use of video conferencing in different settings (e.g., professional context) to a higher extent.

Previous research also indicated that participants who know each other discover delay and impairments in the experiment faster. In future work, it would have been interesting to further investigate whether this hypothesis is true and whether it holds for a wider range of tasks, and whether it is recommended to use people who know each other when testing QoE in WebRTC.

Schmitt, Redi, and Ceasar address large variability in audio ratings in "Towards context-aware interactive Quality of Experience evaluation for audiovisual multiparty conferencing" [35]. They have discover two groups of test subjects: One group is very sensitive to impairments in the video, so their audio quality ratings will be affected

by video interference. The other group of subjects is less bothered by impairments in the video, and they are judging audio quality as high independent on the video disruptions. Future work should try to identify trends within the data set.

Finally, I recommend the ongoing work towards evaluating QoE by means of different types of measures (traditional perceived quality measures and more affective state-related measures) should continue and I see a large potential to combine these with more behavioral (e.g., eye-tracking) and physiological measures to better understand the relation between perceptual, cognitive, affective and behavioral aspects in this respect.

Chapter A
Handed out Material in Experiment

A.1 Suggested Discussion Topics

What was the last funny video you saw?

What do you do to get rid of stress?

What is something you are obsessed with?

Who is your favorite entertainer (comedian, musician, actor, etc.)?

What's your favorite way to waste time?

Do you have any pets? What are their names?

Where did you go last weekend? What did you do?

What are you going to do this weekend?

What is something that is popular now that annoys you?

What did you do on your last vacation?

What was the last time you worked incredibly hard?

Are you very active of do you prefer to just relax in your free time?

What do you do when you hang out with your friends?

Who is your oldest friend? Where did you meet them?

What's the best / worst thing about your work / school?

If you had intro music, what song would it be? Why?

What were you really into when you were a kid?

If you could have any animal as a pet, what animal would you choose?

What three words best describe you?

What would be your perfect weekend?

A.2 Questionnaire to Fill in After Before Experiment

\Studie av interaktiv videokommunikasjon
Spørreskjema som skal fylles ut før eksperimentets start

1. I hvilket år ble du født?
   ……

2. Kjønn?
   - Kvinne
   - Mann
   - Annet

3. Hva er morsmålet ditt?

   ……

4. Har du normalt eller tilnærmet normalt syn?
   - Ja
   - Nei

5. Har du normal eller tilnærmet normal hørsel?
   - Ja
   - Nei

6. Høyeste oppnådde grad?
   - Grunnskole
   - Videregående Skole
   - Bachelorgrad
   - Mastergrad
   - PhD grad

7. Hva er yrket ditt?
   - Student
   - Hjemmeværende
   - Arbeidssøkende/permittert
   - Ufør
   - Selvstendig næringsdrivende
   - Ansatt i offentlig virksomhet
   - Ansatt i privat virksomhet
   - Annet

8. Studerer du eller jobber med lyd / videokvalitet, multimediabehandling eller et relatert felt?
    - Ja
    - Nei

9. Hvilke av følgende tjenester og applikasjoner for online videosamtaler har du brukt i løpet av den siste måneden (omtrentlig)? Flere svar er mulige.
    - Google Hangouts
    - Skype
    - appear.in
    - Facetime
    - Firefox hello
    - Tiny chat
    - Viber
    - Whatsapp
    - Profesjonell eller semi-profesjonell videokonferanse tjeneste
    - Hvis andre, spesifiser:
    - Ingen
    - Vet ikke

10. Hvis du har benyttet deg av noen av de ovenfor nevnte (eller andre) program i hvilken sammenheng?
    - Jobb
    - Samtale med familie/venner
    - Skole
    - Hvis andre, spesifiser:

11. I løpet av den siste måneden, hvor ofte (omtrent) har du deltatt i online videosamtaler ved hjelp av noen av de ovenfornevnte (eller andre) program?
    - Aldri
    - En gang
    - 2 eller 3 ganger
    - Ca en gang i uken
    - Flere ganger i uken
    - Daglig
    - Vet ikke

12. Har du i det siste bruke en applikasjon kalt “appear.in”?
    - Ja
    - Nei

13. Har du tidligere deltatt i brukerstudier eller eksperimenter om online videokommunikasjon?
    - Ja
    - Nei

A.3 Instructions for Participants

Kjære deltaker.

Takk for at du deltar i denne studien som omhandler interaktiv videokommunikasjon ved bruk av appear.in!

Studien vil foregå på følgende måte:

1. Velkommen + informasjon. Vennligst les og signér skjemaet “Informasjon til deltakere”.

2. Vennligst fyll ut forhåndsskjema, som inneholder spørsmål om bakgrunnen din

3. Instruksjoner for studiet: Det vil være åtte korte samtaler, delt i to deler.

    a. I fire av dem kan dere bestemme samtaleemnet helt selv. Det vil ved studiets start bli delt ut en liste med forslag til temaer dere kan snakke om hvis dere ikke finner samtaleemner selv.
    Merk: Dere kan snakke om hva dere vil, men prøv å unngå snakke om selve eksperimentet

    b. I de fire neste øktene får begge deltakere tilgang til en mengde legoklosser. Dere får utdelt komplementære byggeinstruksjoner, og ved å kombinere disse vil dere kunne bygge en gitt legofigur.
    Merk: Dere kan snakke om hva dere vil, men prøv å unngå snakke om selve eksperimentet

4. For alle øktene gjelder følgende:

    Når samtalen er over, vil dette annonseres i et “pop-up”vindu. Du kan da laste ned statistikken i kategorien 'WebRTC Internals', som ble åpnet automatisk ved begynnelsen av sesjonen (og som vist av eksperimentleder).

    Fyll deretter ut spørreskjemaet (på papir), relatert til betingelsene under den gjennomførte samtalen og hvordan du opplevde denne. Vennligst prøv å fylle den inn så intuitivt som mulig (det finnes ikke riktig eller feil, og du trenger ikke å tenke over svarene dine).

    Når du er ferdig, legg det utfylte spørreskjemaet i konvolutten ved siden av deg og lukk nettleservinduet (som vist i popup-vinduet). En ny økt vil starte automatisk etter en kort pause.

5. Etter åtte økter er studiet ferdig. Tusen takk for din deltagelse!

A.4 Consent Form

Studie av interaktiv videokommunikasjon ved bruk av appear.in.

Bakgrunn og formål
Forskningsprosjektet er en masteroppgave ved NTNU våren 2018. Prosjektet fokuserer på brukernes erfaringer med WebRTC-baserte videokommunikasjonstjenester, for eksempel appear.in eller Google Hangouts. I motsetning til applikasjoner som Skype, som krever brukerinnlogging og installasjon av en applikasjon, kan WebRTC-baserte videokommunikasjonstjenester kjøres i en nettleser, og i noen tilfeller kan slike tjenester til og med brukes uten brukerkonto. Bruken av slike tjenester kan imidlertid bli sterkt påvirket (både positivt og negativt) av de tekniske forholdene der samtalen foregår.

Hovedmålet i dette prosjektet er å få bedre forståelse for de tekniske og ikke-tekniske faktorene som kan påvirke brukerens oppfatning av kvaliteten på slike tjenester, og å få innsikt i hvordan disse faktorene kan korreleres. I dette prosjektet benyttes appear.in (en WebRTC applikasjon) som et instrument for å lære mer om QoE-problemstillinger i forbindelse med videokommunikasjon.

Hva innebærer en deltakelse i denne studien?
Du vil bli bedt om å delta i en rekke korte samtaler ved bruk av appear.in i en kontrollert miljø. Under disse korte samtalene får du en enkel oppgave som ikke krever noen forberedelse på forhånd. Etter hver korte samtale vil du bli bedt om å svare på spørsmål relatert til forholdene i samtalen og hvordan du opplevde det. Det vil bli gjort lyd- og bildeopptak av samtalen for analyseformål.

Hva skjer med informasjonen om deg?
Alle personopplysninger vil bli behandlet konfidensielt. Det er kun masterstudent, veileder og ansvarlig professor som vil ha tilgang til personopplysningene og eventuelle opptak. Deltakerne vil ikke gjenkjennes i masteroppgaven da alt vil anonymiseres. Du vil aldri bli identifiserbar eller gjenkjennelig i masteroppgaven. Ingen sensitive opplysninger vil bli inkludert.

Prosjektet skal etter planen avsluttes 18.06.2018. Da vil alle data slettes.

Frivillig deltakelse
Det er frivillig å delta, og du kan når som helst velge å trekke tilbake ditt samtykket uten å oppgi noen grunn. Har du ytterligere spørsmål angående prosjektet, vennligst kontakt Marie Haga (hagamarie@gmail.com, 98886742).

Samtykke til deltakelse i studien

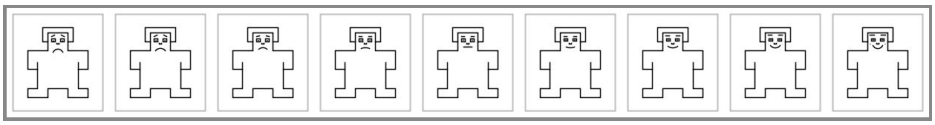
Jeg har mottatt informasjon om prosjektet og er villig til å delta.

(Navn og signatur)

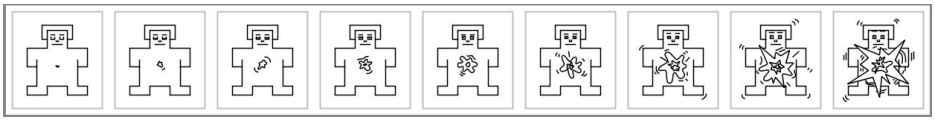
A.5 Questionnaire to Fill in After Each Condition

Vennligst prøv å svare på følgende spørsmål så intuitivt som mulig med bakgrunn i den korte samtalen du nettopp hadde.

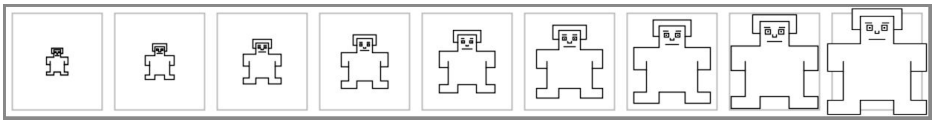
1. Hvor **ulykkelig eller lykkelig** følte du deg under samtalen? Vennligst sett en sirkel rundt bildet som best svarer til din følelse (skalaen varierer fra 1. svært ulykkelig til 9. svært lykkelig).



2. Hvor **rolig eller oppspilt** følte du deg under samtalen? Vennligst sett en sirkel rundt bildet som best svarer til din følelse (skalaen varierer fra 1. veldig rolig til 9. svært oppspilt).



3. Hvor **maktesløs eller kontrollert** følte du deg under økten? Vennligst sett en sirkel rundt bildet som best svarer til din følelse (skalaen varierer fra 1. veldig maktesløs, uten kontroll til 9. veldig dominerende, full kontroll).



4. Hvordan vurderer du den **generelle audiovisuelle kvaliteten** på økten? (samlet med kombinert lyd- og videokvalitet)? Hak av alternativet du føler passer best

- 5 - utmerket
- 4 - bra
- 3 - helt ok
- 2 - dårlig
- 1 - svært dårlig

5. Hvor **sikker** er du på svaret ditt?

- 5 - veldig sikker
- 4 - sikker
- 3 - hverken sikker eller usikker
- 2 - usikker
- 1 - veldig usikker

6. Hvordan vurderer du **videokvaliteten** på økten? Hak av alternativet du føler passer best

5 - utmerket
4 - bra
3 - helt ok
2 - dårlig
1 - svært dårlig

7. Hvordan vurderer du **lydkvaliteten** på økten? Hak av alternativet du føler passer best

5 - utmerket
4 - bra
3 - helt ok
2 - dårlig
1 - svært dårlig

8. La du merke til noen **tekniske forstyrrelser** under samtalen?

- Ja
- Nei

9. Hvor mye **innsats** måtte du legge inn for å forstå hva den andre personen fortalte deg? Hak av alternativet du føler passer best

Ingen spesiell innsats nødvendig
Minimal innsats var nødvendig
Moderat innsats var nødvendig
Betydelig innsats var nødvendig
Stor innsats var nødvendig

10. Hvor **irritert** følte du deg under samtalen? Hak av alternativet du føler passer best.

Ingen irritasjon
Minimal irritasjon
Moderat irritasjon
Vesentlig irritasjon
Enorm irritasjon

11. Hvordan **fordelte dere taletiden** under samtalen? Hak av alternativet du føler passer best.

Kun jeg snakket
Jeg snakkes mest
Vi fordelte taletiden jevnt mellom oss
Samtalepartner min snakket mest
Kun samtalepartneren min snakket

12. Under samtalen, skjedde det at du og samtalepartneren din **startet å snakke samtidig**?

- Ja
- Nei

13. Hvis ja, i hvilken grad opplevde du dette som **irriterende**?

Ingen irritasjon
Minimal irritasjon
Moderat irritasjon
Vesentlig irritasjon
Enorm irritasjon

Dette er slutten på spørreskjemaet. Takk for svarene dine.

Chapter B
Output from Wilcoxon Signed Ranks Test, Comparing Free Conversation and LEGO-task

B.1 Rated Overall Audiovisual Quality in Lego-Task Versus Free Conversation

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
FT_Overall_qual_good – Lego_Overall_qual_good	Negative Ranks	3[a]	2.50	7.50
	Positive Ranks	4[b]	5.13	20.50
	Ties	9[c]		
	Total	16		
FT_Overall_qual_AV – Lego_Overall_qual_AV	Negative Ranks	6[d]	6.17	37.00
	Positive Ranks	4[e]	4.50	18.00
	Ties	6[f]		
	Total	16		
FT_Overall_qual_A – Lego_Overall_qual_A	Negative Ranks	2[g]	3.00	6.00
	Positive Ranks	8[h]	6.13	49.00
	Ties	6[i]		
	Total	16		
FT_Overall_qual_V – Lego_Overall_qual_V	Negative Ranks	4[j]	5.00	20.00
	Positive Ranks	4[k]	4.00	16.00
	Ties	8[l]		
	Total	16		

a. FT_Overall_qual_good < Lego_Overall_qual_good
b. FT_Overall_qual_good > Lego_Overall_qual_good
c. FT_Overall_qual_good = Lego_Overall_qual_good
d. FT_Overall_qual_AV < Lego_Overall_qual_AV
e. FT_Overall_qual_AV > Lego_Overall_qual_AV
f. FT_Overall_qual_AV = Lego_Overall_qual_AV
g. FT_Overall_qual_A < Lego_Overall_qual_A
h. FT_Overall_qual_A > Lego_Overall_qual_A
i. FT_Overall_qual_A = Lego_Overall_qual_A
j. FT_Overall_qual_V < Lego_Overall_qual_V
k. FT_Overall_qual_V > Lego_Overall_qual_V
l. FT_Overall_qual_V = Lego_Overall_qual_V

Test Statistics[a]

	FT_Overall_qual_good – Lego_Overall_qual_good	FT_Overall_qual_AV – Lego_Overall_qual_AV	FT_Overall_qual_A – Lego_Overall_qual_A	FT_Overall_qual_V – Lego_Overall_qual_V
Z	-1.121[b]	-1.027[c]	-2.228[b]	-.289[c]
Asymp. Sig. (2-tailed)	.262	.305	.026	.773

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.
c. Based on positive ranks.

Figure B.1: p-values for Rated Overall Audiovisual Quality in Lego-Task Versus Free Conversation

B.2 Rated Audio Quality in Lego-Task Versus Free Conversation

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
FT_Audio_qual_good – Lego_Audio_qual_good	Negative Ranks	2[a]	2.50	5.00
	Positive Ranks	3[b]	3.33	10.00
	Ties	11[c]		
	Total	16		
FT_Audio_qual_AV – Lego_Audio_qual_AV	Negative Ranks	7[d]	5.00	35.00
	Positive Ranks	2[e]	5.00	10.00
	Ties	7[f]		
	Total	16		
FT_Audio_qual_A – Lego_Audio_qual_A	Negative Ranks	2[g]	3.00	6.00
	Positive Ranks	7[h]	5.57	39.00
	Ties	7[i]		
	Total	16		
FT_Video_qual_V – Lego_Video_qual_V	Negative Ranks	3[j]	7.33	22.00
	Positive Ranks	6[k]	3.83	23.00
	Ties	7[l]		
	Total	16		

a. FT_Audio_qual_good < Lego_Audio_qual_good
b. FT_Audio_qual_good > Lego_Audio_qual_good
c. FT_Audio_qual_good = Lego_Audio_qual_good
d. FT_Audio_qual_AV < Lego_Audio_qual_AV
e. FT_Audio_qual_AV > Lego_Audio_qual_AV
f. FT_Audio_qual_AV = Lego_Audio_qual_AV
g. FT_Audio_qual_A < Lego_Audio_qual_A
h. FT_Audio_qual_A > Lego_Audio_qual_A
i. FT_Audio_qual_A = Lego_Audio_qual_A
j. FT_Video_qual_V < Lego_Video_qual_V
k. FT_Video_qual_V > Lego_Video_qual_V
l. FT_Video_qual_V = Lego_Video_qual_V

Test Statistics[a]

	FT_Audio_qual_good – Lego_Audio_qual_good	FT_Audio_qual_AV – Lego_Audio_qual_AV	FT_Audio_qual_A – Lego_Audio_qual_A	FT_Video_qual_V – Lego_Video_qual_V
Z	-.707[b]	-1.667[c]	-1.992[b]	-.060[b]
Asymp. Sig. (2-tailed)	.480	.096	.046	.952

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.
c. Based on positive ranks.

Figure B.2: p-values for Rated Audio Quality in Lego-Task Versus Free Conversation

B.3 Rated Audio Quality in Lego-Task Versus Free Conversation

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
FT_Video_qual_good - Lego_Video_qual_good	Negative Ranks	4[a]	5.00	20.00
	Positive Ranks	6[b]	5.83	35.00
	Ties	6[c]		
	Total	16		
FT_Video_qual_AV - Lego_Video_qual_AV	Negative Ranks	5[d]	6.60	33.00
	Positive Ranks	6[e]	5.50	33.00
	Ties	5[f]		
	Total	16		
FT_Video_qual_A - Lego_Video_qual_A	Negative Ranks	2[g]	2.50	5.00
	Positive Ranks	7[h]	5.71	40.00
	Ties	7[i]		
	Total	16		
FT_Video_qual_V - Lego_Video_qual_V	Negative Ranks	3[j]	7.33	22.00
	Positive Ranks	6[k]	3.83	23.00
	Ties	7[l]		
	Total	16		

a. FT_Video_qual_good < Lego_Video_qual_good
b. FT_Video_qual_good > Lego_Video_qual_good
c. FT_Video_qual_good = Lego_Video_qual_good
d. FT_Video_qual_AV < Lego_Video_qual_AV
e. FT_Video_qual_AV > Lego_Video_qual_AV
f. FT_Video_qual_AV = Lego_Video_qual_AV
g. FT_Video_qual_A < Lego_Video_qual_A
h. FT_Video_qual_A > Lego_Video_qual_A
i. FT_Video_qual_A = Lego_Video_qual_A
j. FT_Video_qual_V < Lego_Video_qual_V
k. FT_Video_qual_V > Lego_Video_qual_V
l. FT_Video_qual_V = Lego_Video_qual_V

Test Statistics[a]

	FT_Video_qual_good - Lego_Video_qual_good	FT_Video_qual_AV - Lego_Video_qual_AV	FT_Video_qual_A - Lego_Video_qual_A	FT_Video_qual_V - Lego_Video_qual_V
Z	-.832[b]	.000[c]	-2.101[b]	-.060[b]
Asymp. Sig. (2-tailed)	.405	1.000	.036	.952

a. Wilcoxon Signed Ranks Test
b. Based on negative ranks.
c. The sum of negative ranks equals the sum of positive ranks.

Figure B.3: p-values for Rated Video Quality in Lego-Task Versus Free Conversation

B.4 Output for Felt Arousal in Lego-Task Versus Free Conversation

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
FT_Arousal_good – Lego_Arousal_good	Negative Ranks	11[a]	6.82	75.00
	Positive Ranks	1[b]	3.00	3.00
	Ties	4[c]		
	Total	16		
FT_Arousal_AV – Lego_Arousal_AV	Negative Ranks	12[d]	7.25	87.00
	Positive Ranks	1[e]	4.00	4.00
	Ties	3[f]		
	Total	16		
FT_Arousal_A – Lego_Arousal_A	Negative Ranks	11[g]	6.00	66.00
	Positive Ranks	0[h]	.00	.00
	Ties	5[i]		
	Total	16		
FT_Arousal_V – Lego_Arousal_V	Negative Ranks	12[j]	6.50	78.00
	Positive Ranks	0[k]	.00	.00
	Ties	4[l]		
	Total	16		

a. FT_Arousal_good < Lego_Arousal_good
b. FT_Arousal_good > Lego_Arousal_good
c. FT_Arousal_good = Lego_Arousal_good
d. FT_Arousal_AV < Lego_Arousal_AV
e. FT_Arousal_AV > Lego_Arousal_AV
f. FT_Arousal_AV = Lego_Arousal_AV
g. FT_Arousal_A < Lego_Arousal_A
h. FT_Arousal_A > Lego_Arousal_A
i. FT_Arousal_A = Lego_Arousal_A
j. FT_Arousal_V < Lego_Arousal_V
k. FT_Arousal_V > Lego_Arousal_V
l. FT_Arousal_V = Lego_Arousal_V

Test Statistics[a]

	FT_Arousal_good – Lego_Arousal_good	FT_Arousal_AV – Lego_Arousal_AV	FT_Arousal_A – Lego_Arousal_A	FT_Arousal_V – Lego_Arousal_V
Z	-2.852[b]	-2.955[b]	-2.952[b]	-3.078[b]
Asymp. Sig. (2-tailed)	.004	.003	.003	.002

a. Wilcoxon Signed Ranks Test
b. Based on positive ranks.

Figure B.4: p-values for Felt Arousal in Lego-Task Versus Free Conversation

B.5 Felt Annoyance in Lego-Task Versus Free Conversation

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
FT_Annoyance_good - Lego_Annoyance_good	Negative Ranks	6[a]	3.50	21.00
	Positive Ranks	0[b]	.00	.00
	Ties	10[c]		
	Total	16		
FT_Annoyance_AV - Lego_Annoyance_AV	Negative Ranks	5[d]	6.70	33.50
	Positive Ranks	6[e]	5.42	32.50
	Ties	5[f]		
	Total	16		
FT_Annoyance_A - Lego_Annoyance_A	Negative Ranks	6[g]	4.33	26.00
	Positive Ranks	1[h]	2.00	2.00
	Ties	9[i]		
	Total	16		
FT_Annoyance_V - Lego_Annoyance_V	Negative Ranks	8[j]	4.63	37.00
	Positive Ranks	1[k]	8.00	8.00
	Ties	7[l]		
	Total	16		

a. FT_Annoyance_good < Lego_Annoyance_good
b. FT_Annoyance_good > Lego_Annoyance_good
c. FT_Annoyance_good = Lego_Annoyance_good
d. FT_Annoyance_AV < Lego_Annoyance_AV
e. FT_Annoyance_AV > Lego_Annoyance_AV
f. FT_Annoyance_AV = Lego_Annoyance_AV
g. FT_Annoyance_A < Lego_Annoyance_A
h. FT_Annoyance_A > Lego_Annoyance_A
i. FT_Annoyance_A = Lego_Annoyance_A
j. FT_Annoyance_V < Lego_Annoyance_V
k. FT_Annoyance_V > Lego_Annoyance_V
l. FT_Annoyance_V = Lego_Annoyance_V

Test Statistics[a]

	FT_Annoyance_good - Lego_Annoyance_good	FT_Annoyance_AV - Lego_Annoyance_AV	FT_Annoyance_A - Lego_Annoyance_A	FT_Annoyance_V - Lego_Annoyance_V
Z	-2.232[b]	-.047[b]	-2.081[b]	-1.780[b]
Asymp. Sig. (2-tailed)	.026	.963	.037	.075

a. Wilcoxon Signed Ranks Test
b. Based on positive ranks.

Figure B.5: p-values for Felt Annoyance in Lego-Task Versus Free Conversation

Chapter C

Output from Kruskal-Wallis Test, Comparing LEGO-task and Celebrity Name Guessing

C.1 Condition G

condition = G

Kruskal-Wallis Test

Ranks[a]

	Task	N	Mean Rank
Arousal	Lego	16	16.13
	Celebrity Guessing	18	18.72
	Total	34	
Valence	Lego	16	17.81
	Celebrity Guessing	18	17.22
	Total	34	
Overall_qual	Lego	16	17.09
	Celebrity Guessing	18	17.86
	Total	34	
Audio_qual	Lego	16	17.84
	Celebrity Guessing	18	17.19
	Total	34	
Annoyance	Lego	16	17.78
	Celebrity Guessing	18	17.25
	Total	34	
Effort	Lego	16	17.81
	Celebrity Guessing	18	17.22
	Total	34	
Video_qual	Lego	16	17.81
	Celebrity Guessing	18	17.22
	Total	34	

a. condition = G

Test Statistics[a,b,c]

	Arousal	Valence	Overall_qual	Audio_qual	Annoyance	Effort	Video_qual
Kruskal-Wallis H	.592	.033	.062	.046	.032	.047	.034
df	1	1	1	1	1	1	1
Asymp. Sig.	.441	.856	.803	.830	.857	.829	.853

a. condition = G

b. Kruskal Wallis Test

c. Grouping Variable: Task

Figure C.1: Test Statistics Kruskal-Wallis Test for Condition G

C.2 Condition A

condition = A

Kruskal-Wallis Test

Ranks[a]

	Task	N	Mean Rank
Arousal	Lego	16	16.25
	Celebrity Guessing	18	18.61
	Total	34	
Valence	Lego	16	19.31
	Celebrity Guessing	18	15.89
	Total	34	
Overall_qual	Lego	16	20.78
	Celebrity Guessing	18	14.58
	Total	34	
Audio_qual	Lego	16	23.03
	Celebrity Guessing	18	12.58
	Total	34	
Annoyance	Lego	16	12.66
	Celebrity Guessing	18	21.81
	Total	34	
Effort	Lego	16	13.66
	Celebrity Guessing	18	20.92
	Total	34	
Video_qual	Lego	16	19.50
	Celebrity Guessing	18	15.72
	Total	34	

a. condition = A

Test Statistics[a,b,c]

	Arousal	Valence	Overall_qual	Audio_qual	Annoyance	Effort	Video_qual
Kruskal-Wallis H	.490	1.050	3.484	9.772	8.097	5.074	1.333
df	1	1	1	1	1	1	1
Asymp. Sig.	.484	.305	.062	.002	.004	.024	.248

a. condition = A

b. Kruskal Wallis Test

c. Grouping Variable: Task

Figure C.2: Test Statistics Kruskal-Wallis Test for Condition A

C.2.1 Condition AV

C.3 Condition AV

condition = AV

Kruskal-Wallis Test

Ranks[a]

	Task	N	Mean Rank
Arousal	Lego	16	19.44
	Celebrity Guessing	18	15.78
	Total	34	
Valence	Lego	16	15.38
	Celebrity Guessing	18	19.39
	Total	34	
Overall_qual	Lego	16	17.19
	Celebrity Guessing	18	17.78
	Total	34	
Audio_qual	Lego	16	19.66
	Celebrity Guessing	18	15.58
	Total	34	
Annoyance	Lego	16	15.31
	Celebrity Guessing	18	19.44
	Total	34	
Effort	Lego	16	17.75
	Celebrity Guessing	18	17.28
	Total	34	
Video_qual	Lego	16	14.59
	Celebrity Guessing	18	20.08
	Total	34	

a. condition = AV

Test Statistics[a,b,c]

	Arousal	Valence	Overall_qual	Audio_qual	Annoyance	Effort	Video_qual
Kruskal-Wallis H	1.192	1.421	.033	1.710	1.629	.021	2.804
df	1	1	1	1	1	1	1
Asymp. Sig.	.275	.233	.856	.191	.202	.885	.094

a. condition = AV

b. Kruskal Wallis Test

c. Grouping Variable: Task

Figure C.3: Test Statistics Kruskal-Wallis Test for Condition AV

C.4 Condition V

condition = V

Kruskal-Wallis Test

Ranks[a]

	Task	N	Mean Rank
Arousal	Lego	16	19.28
	Celebrity Guessing	17	14.85
	Total	33	
Valence	Lego	16	15.84
	Celebrity Guessing	17	18.09
	Total	33	
Overall_qual	Lego	16	15.91
	Celebrity Guessing	17	18.03
	Total	33	
Audio_qual	Lego	16	15.41
	Celebrity Guessing	17	18.50
	Total	33	
Annoyance	Lego	16	18.56
	Celebrity Guessing	17	15.53
	Total	33	
Effort	Lego	16	20.34
	Celebrity Guessing	17	13.85
	Total	33	
Video_qual	Lego	16	16.22
	Celebrity Guessing	17	17.74
	Total	33	

a. condition = V

Test Statistics[a,b,c]

	Arousal	Valence	Overall_qual	Audio_qual	Annoyance	Effort	Video_qual
Kruskal-Wallis H	1.781	.470	.431	.922	.906	4.166	.224
df	1	1	1	1	1	1	1
Asymp. Sig.	.182	.493	.512	.337	.341	.041	.636

a. condition = V

b. Kruskal Wallis Test

c. Grouping Variable: Task

Figure C.4: Test Statistics Kruskal-Wallis Test for Condition V

Chapter D
Output from Kruskal-Wallis Test, Comparing Free Conversation and Celebrity Name Guessing

D.1 Condition G

condition = G

Kruskal-Wallis Test

Ranks[a]

	Task	N	Mean Rank
Valence	Free conversation	15	17.67
	Celebrity Guessing	18	16.44
	Total	33	
Arousal	Free conversation	15	11.57
	Celebrity Guessing	18	21.53
	Total	33	
Overall_qual	Free conversation	15	17.53
	Celebrity Guessing	18	16.56
	Total	33	
Video_qual	Free conversation	15	18.63
	Celebrity Guessing	18	15.64
	Total	33	
Audio_qual	Free conversation	15	18.20
	Celebrity Guessing	18	16.00
	Total	33	
Effort	Free conversation	15	15.07
	Celebrity Guessing	18	18.61
	Total	33	
Annoyance	Free conversation	15	13.50
	Celebrity Guessing	18	19.92
	Total	33	

a. condition = G

Test Statistics[a,b,c]

	Valence	Arousal	Overall_qual	Video_qual	Audio_qual	Effort	Annoyance
Kruskal-Wallis H	.139	9.165	.110	.951	.534	2.447	7.130
df	1	1	1	1	1	1	1
Asymp. Sig.	.709	.002	.740	.329	.465	.118	.008

a. condition = G

b. Kruskal Wallis Test

c. Grouping Variable: Task

Figure D.1: Test Statistics, Kruskal-Wallis Test for Condition G

D.1.1 Condition AV

condition = AV

Kruskal-Wallis Test

Ranks[a]

	Task	N	Mean Rank
Valence	Free conversation	16	15.25
	Celebrity Guessing	18	19.50
	Total	34	
Arousal	Free conversation	16	14.91
	Celebrity Guessing	18	19.81
	Total	34	
Overall_qual	Free conversation	16	16.06
	Celebrity Guessing	18	18.78
	Total	34	
Video_qual	Free conversation	16	14.59
	Celebrity Guessing	18	20.08
	Total	34	
Audio_qual	Free conversation	16	17.50
	Celebrity Guessing	18	17.50
	Total	34	
Effort	Free conversation	16	19.03
	Celebrity Guessing	18	16.14
	Total	34	
Annoyance	Free conversation	16	15.69
	Celebrity Guessing	18	19.11
	Total	34	

a. condition = AV

Test Statistics[a,b,c]

	Valence	Arousal	Overall_qual	Video_qual	Audio_qual	Effort	Annoyance
Kruskal-Wallis H	1.589	2.135	.724	2.842	.000	.773	1.075
df	1	1	1	1	1	1	1
Asymp. Sig.	.207	.144	.395	.092	1.000	.379	.300

a. condition = AV

b. Kruskal Wallis Test

c. Grouping Variable: Task

Figure D.2: Test Statistics Kruskal-Wallis Test for Condition AV

D.2 Condition A

condition = A

Kruskal-Wallis Test

Ranks[a]

	Task	N	Mean Rank
Valence	Free conversation	16	20.38
	Celebrity Guessing	18	14.94
	Total	34	
Arousal	Free conversation	16	10.97
	Celebrity Guessing	18	23.31
	Total	34	
Overall_qual	Free conversation	16	24.44
	Celebrity Guessing	18	11.33
	Total	34	
Video_qual	Free conversation	16	22.91
	Celebrity Guessing	18	12.69
	Total	34	
Audio_qual	Free conversation	16	25.34
	Celebrity Guessing	18	10.53
	Total	34	
Effort	Free conversation	16	10.66
	Celebrity Guessing	18	23.58
	Total	34	
Annoyance	Free conversation	16	10.88
	Celebrity Guessing	18	23.39
	Total	34	

a. condition = A

Test Statistics[a,b,c]

	Valence	Arousal	Overall_qual	Video_qual	Audio_qual	Effort	Annoyance
Kruskal-Wallis H	2.711	13.583	15.779	10.917	20.013	17.694	16.087
df	1	1	1	1	1	1	1
Asymp. Sig.	.100	.000	.000	.001	.000	.000	.000

a. condition = A

b. Kruskal Wallis Test

c. Grouping Variable: Task

Figure D.3: Test Statistics Kruskal-Wallis Test for Condition A

D.3 Condition V

condition = V

Kruskal–Wallis Test

Ranks[a]

	Task	N	Mean Rank
Valence	Free conversation	16	16.84
	Celebrity Guessing	17	17.15
	Total	33	
Arousal	Free conversation	16	13.53
	Celebrity Guessing	17	20.26
	Total	33	
Overall_qual	Free conversation	16	16.31
	Celebrity Guessing	17	17.65
	Total	33	
Video_qual	Free conversation	16	16.81
	Celebrity Guessing	17	17.18
	Total	33	
Audio_qual	Free conversation	16	17.94
	Celebrity Guessing	17	16.12
	Total	33	
Effort	Free conversation	16	17.13
	Celebrity Guessing	17	16.88
	Total	33	
Annoyance	Free conversation	16	15.13
	Celebrity Guessing	17	18.76
	Total	33	

a. condition = V

Test Statistics[a,b,c]

	Valence	Arousal	Overall_qual	Video_qual	Audio_qual	Effort	Annoyance
Kruskal-Wallis H	.009	4.178	.178	.013	.326	.006	1.395
df	1	1	1	1	1	1	1
Asymp. Sig.	.926	.041	.673	.911	.568	.938	.237

a. condition = V

b. Kruskal Wallis Test

c. Grouping Variable: Task

Figure D.4: Test Statistics Kruskal-Wallis Test for Condition V

References

[1] Trilogy-LTE, "How webrtc is revolutionizing telephony," 2014. http://blogs.trilogy-lte.com/post/77427158750/how-webrtc-is-revolutionizing-telephon.

[2] S. Möller and A. Raake, *Quality of experience: advanced concepts, applications and methods*. Springer, 2014.

[3] D. Ammar, K. De Moor, and P. Heegaard, "Quality of experience-assessment of webrtc based video communication," *ERCIM NEWS*, no. 105, pp. 42–43, 2016.

[4] K. De Moor, S. Arndt, D. Ammar, J.-N. Voigt-Antons, A. Perkis, and P. E. Heegaard, "Exploring diverse measures for evaluating qoe in the context of webrtc," in *Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on*, pp. 1–3, IEEE, 2017.

[5] S. Sector and O. Itu, "Series p: Terminals and subjective and objective assessment methods methods for objective and subjective assessment of speech quality," *ITU*, 2013.

[6] S. Sector and O. Itu, "Series p: Telephone transmission quality, telephone installations, local line networks. subjective quality evaluation of audio and audiovisual multiparty telemeetings," *ITU*, 2012.

[7] scientificamerican, "\"can videoconferencing replace traveling?\"." https://www.scientificamerican.com/article/can-videoconferencing-replace-travel/, 2009.

[8] M. web docs, "Webrtc api." https://developer.mozilla.org/en-US/docs/Web/API/WebRTC_API, 2017.

[9] H. F. Johnsen, "Start en videosamtale kun med ett klikk," 2015. https://www.misc.no/service/appear-in.jsp.

[10] D. Ammar, K. De Moor, and P. Heegaard, "An experimental platform for qoe studies of webrtc-based multi-party video communication," 2016.

[11] E. Fosser and L. O. D. Nedberg, "Quality of experience of webrtc based video communication," Master's thesis, NTNU, 2016.

[12] S. Sector and O. Itu, "Series e: overall network operation telephone service service operation and human factors quality of telecommunication services: concepts models objectives and dependability planning-use of quality of service objectives for planning of telecommunication networks," *ITU*, 2008.

[13] K. Brunnström, S. A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, M.-C. Larabi, *et al.*, "Qualinet white paper on definitions of quality of experience," 2013.

[14] K. De Moor, F. Mazza, I. Hupont, M. R. Quintero, T. Mäki, and M. Varela, "Chamber qoe: a multi-instrumental approach to explore affective aspects in relation to quality of experience," in *Human Vision and Electronic Imaging XIX*, vol. 9014, p. 90140U, International Society for Optics and Photonics, 2014.

[15] S. Sector and O. Itu, "Recommendation itu-t p.10/g.100, vocabulary for performance, quality of service and quality of experience," *ITU*, 2017.

[16] A. Vakili and J.-C. Grégoire, "Qoe management for video conferencing applications," *Computer Networks*, vol. 57, no. 7, pp. 1726–1738, 2013.

[17] G. Berndtsson, M. Folkesson, and V. Kulyk, "Subjective quality assessment of video conferences and telemeetings," in *Packet Video Workshop (PV), 2012 19th International*, pp. 25–30, IEEE, 2012.

[18] P. ITU-T RECOMMENDATION, "Subjective video quality assessment methods for multimedia applications," 1999.

[19] M. Schmitt, S. Gunkel, P. Cesar, and P. Hughes, "A qoe testbed for socially-aware video-mediated group communication," in *Proceedings of the 2nd international workshop on Socially-aware multimedia*, pp. 37–42, ACM, 2013.

[20] D. Ammar, K. De Moor, M. Xie, M. Fiedler, and P. Heegaard, "Video qoe killer and performance statistics in webrtc-based video communication," in *Communications and Electronics (ICCE), 2016 IEEE Sixth International Conference on*, pp. 429–436, IEEE, 2016.

[21] W. Wu, A. Arefin, R. Rivas, K. Nahrstedt, R. Sheppard, and Z. Yang, "Quality of experience in distributed interactive multimedia environments: toward a theoretical framework," in *Proceedings of the 17th ACM international conference on Multimedia*, pp. 481–490, ACM, 2009.

[22] N. Unuth, "Mean opinion score (mos): a measure of voice quality," 2018. https://www.lifewire.com/measure-voice-quality-3426718.

[23] M. Fiedler, S. Möller, P. Reichl, and M. Xie, "Qoe vadis?(dagstuhl perspectives workshop 16472)," 2018.

[24] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[25] F. Brauer, M. S. Ehsan, and G. Kubin, "Subjective evaluation of conversational multimedia quality in ip networks," in *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pp. 872–876, IEEE, 2008.

[26] J. Wang, F. Yang, Z. Xie, and S. Wan, "Evaluation on perceptual audiovisual delay using average talkspurts and delay," in *Image and Signal Processing (CISP), 2010 3rd International Congress on*, vol. 1, pp. 125–128, IEEE, 2010.

[27] B. Belmudez, S. Moeller, B. Lewcio, A. Raake, and A. Mehmood, "Audio and video channel impact on perceived audio-visual quality in different interactive contexts," in *Multimedia Signal Processing, 2009. MMSP'09. IEEE International Workshop on*, pp. 1–5, IEEE, 2009.

[28] S. Möller, *Assessment and prediction of speech quality in telecommunications*. Springer Science & Business Media, 2012.

[29] O. Dalland, *Metode og oppgaveskriving for studenter*. Gyldendal akademisk, 2007.

[30] C. Jones, "Advantages and disadvantages of qualitative and quantitative research," 2017. https://classroom.synonym.com/advantages-disadvantages-of-qualitative-quantitative-research-12082716.html.

[31] S. R. Brown and L. E. Melamed, *Experimental design and analysis*. No. 74, Sage, 1990.

[32] C. f. I. i. R. Grand Canyon University and Teaching, "Benefits and limitations of experimental research," 2015. https://cirt.gcu.edu/research/developmentresources/research_ready/experimental/benefits_limits.

[33] D. Kowalczyk, "Non-experimental and experimental research: Differences, advantages & disadvantages," 2012. http://study.com/academy/lesson/non-experimental-and-experimental-research-differences-advantages-disadvantages.html.

[34] A. Field, *Discovering statistics using IBM SPSS statistics*. sage, 2013.

[35] M. Schmitt, J. Redi12, and P. Cesar12, "Towards context-aware interactive quality of experience evaluation for audiovisual multiparty conferencing," in *Proc. 5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)*, pp. 54–58, 2016.

[36] J. Guilford, "Fundamental statistics in psychology and education 4th ed," 1965.

[37] M. Schmitt, S. Gunkel, P. Cesar, and D. Bulterman, "Asymmetric delay in video-mediated group discussions," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*, pp. 19–24, IEEE, 2014.

[38] K. D. Moor, M. Fiedler, P. Reichl, and M. Varela, "Quality of experience: From assessment to application (dagstuhl seminar 15022)," *Dagstuhl Reports*, vol. 5, no. 1, pp. 57–95, 2015.