NTNU – Trondheim
Norwegian University of
Science and Technology

Investigating QoE in a Cloud-Based Classroom Response System

A Real-Life Longitudinal and Cross-Sectional
Study of Kahoot!

Marthe Thorine Sunde
Anlaug Gårdsrud Underdal

Master of Science in Communication Technology
Submission date: May 2014
Supervisor: Poul Einar Heegaard, ITEM
Co-supervisor: Katrien De Moor, ITEM

Norwegian University of Science and Technology
Department of Telematics

Title: Investigating QoE in a Cloud-Based
Classroom Response System -
A Real-Life Longitudinal and
Cross-Sectional Lab Study of Kahoot!

Student: Marthe Thorine Sunde and Anlaug Underdal

Problem description:

The converged telecom, IT and media industries have traditionally been dominated by a Quality of Service (QoS) and technology-centered discourse. Now there is a growing awareness that users, their experiences and the broader socio-economic context in which these experiences take place, cannot be ignored any longer. The key differentiators are no longer only the technological excellence and optimized QoS, but increasingly also enabling pleasurable and positive experiences, meeting users' expectations. In literature, the concept Quality of Experience (QoE) has been introduced in this respect.

The fast growing Cloud industry is causing intense cost-driven competition and providers of Cloud computing must for this reason rely on the quality of their services to acquire consumers. If performance levels and the overall experience that is offered, do not reach users' expectations with respect to the utility and/or enjoyment of the application or service, as defined in Qualinet White Paper on Definitions of Quality of Experience, users might abandon the service or refuse adoption. It is for this reason important for providers to obtain accurate insights into the QoE related to the use of a system, service or application, from a user perspective in addition to the service perspective.

The goal of this thesis work is to investigate the user-perceived QoE of Kahoot!, a Cloud gaming learning platform, designed to be used in classrooms. The general objective of this thesis is to investigate to what extent a specific QoS parameter (delay), affects the QoE perceived by users of such a cloud-based application. More specifically the following research questions have been set: To which degree does delay impact QoE and how does the presence of others influence the experience of fairness?

Experiments involving actual users will be conducted in a real-life setting as well as in a controlled lab environment. The two different setups are to be conducted for their complementary characteristics. For the real-life testing, the ecological validity is higher but control over possible influencing factors will be lower. For the lab environment, different factors regarding delay will be more controllable, but the test environment is artificial, it does not reflect the natural user context.

The real-life experiment will be conducted within a course at Norwegian University of Science and Technology (NTNU) which provides a large test panel consisting of students, using a longitudinal setup. The longitudinal setup is chosen to investigate whether and how QoE and the related expectations change over time. In this part of the experiment, different levels of delays will be investigated throughout a series of tests. The delay in each test will be equal for every participating student.

The lab experiment is to be conducted on a smaller test panel of students. The experiment will take place in a controlled lab environment where the network connections of each participant will be manipulated, giving connections of different quality to the students.

The following tasks are planned:

- Review state of the art of Cloud services and applications, as well as QoE.
- Discuss possible influential factors and evaluate relevant related work on clouding application and for the *Kahoot!* application in specific.
- Investigating the influence of network-related parameters on Gaming QoE.
- Investigate co-experiences: how does the presence of others influence experienced fairness and QoE?
- Investigate whether and how QoE changes over time under fixed network settings.

Responsible professor: Poul Heegaard, ITEM
Supervisor: Katrien De Moor, ITEM

Abstract

Due to advantages like quick deployment, (almost) unlimited storage, cost efficiency as well as easy access, administration and maintenance, Cloud Computing is gaining enormous attention by both researchers and corporations. This paradigm has boomed over the last years and in this context a wide range of services has migrated to the cloud. One type of service that is strongly growing in importance is Cloud Gaming, which combines the concepts of Cloud Computing and Online Gaming. Unlike traditional computer games, the game itself is located in the cloud and not on the end device used for playing. Consequently the performance of the network has become a crucial factor that may have a strong influence on the game experience. Therefore, it should be investigated and better understood.

For providers of Cloud Gaming services and platforms, and for application developers, it is important to provide and maintain a good service quality to acquire and retain users. For this reason the paradigm of Cloud Gaming is largely dependent on their demands and willingness to use the application or service. As a result, users determine the success or failure of an application. Put differently: without gamers there are no use for the game. If users are unsatisfied they might refuse to adopt the service and switch to another provider, even though from a technical perspective the service is functioning optimally. For this reason it is important for providers and developers to obtain accurate insight into Quality of Experience (QoE) from a user perspective, and not only from a pure service perspective. Even though QoE often is presented as a crucial concept when migrating to the cloud, there are few studies that examine cloud QoE from the user's point of view.

The goal of this thesis work is to investigate the user-perceived QoE of the cloud gaming application Kahoot!. The application is a Classroom Response System which can be played on any device with a browser. Users obtain scores depending on their answering time, thus the response speed is essential. The general objective of this thesis is to investigate to what extent a specific QoS parameter, namely delay, affects the QoE perceived by users of a cloud-based application.

To reach this objective, a series of subjective user tests have been conducted in a real-life setting as well as in a lab environment.

The real-life experiment (N = 175) was conducted within a course at Norwegian University of Science and Technology (NTNU), using a longitudinal setup. The longitudinal setup was chosen to investigate whether and how QoE and the related expectations changed over time. In this part of the experiment, levels of no-, moderate- and high delay were introduced throughout a series of 4 tests.

The lab experiment with a cross-sectional setup was conducted with a smaller test panel of students (N = 21), in a controlled lab environment where the network connection of each participant was manipulated, giving connections of different quality to the students. This experiment was carried out to complement the Longitudinal study, and to find how users experience being given unequal conditions and whether this creates feelings of unfairness among the users. QoE has been evaluated based on the new definition of QoE proposed by Qualinet [1], which is presented in Section 2.1.2, and for this reason QoE has been evaluated through subjective-measures of delight, annoyance and quality. Because of the impact external factors have on users, measurement of QoE is difficult. QoE is multidimensional and thus incorporates the influence of non-technical aspects such as user characteristics and the context of use, it is hard to analyze all the external factors as they are subjective and complex to document.

The result of these user studies has shown that an unfair setting where some students experience delay while others does not, enhances the feelings of annoyance among the affected users while it increases the feelings of delight among the users who are not affected by delay, "every man for himself". In addition, when given a reference point, students are affected by this, changing their QoE of a delay setting.

This thesis work is an attempt to make a contribution to the literature on QoE in the context of Cloud gaming and CRS, but the findings need to be further explored and validated in follow-up research. Moreover, new questions have been raised, which can guide directions for future research.

Sammendrag

Grunnet fordeler som rask distribusjon, (nesten) ubegrenset lagringsplass, kostnadseffektivitet samt enkel tilgang, administrasjon og vedlikehold, har Cloud Computing (nettsky) mottatt stor oppmerksomhet fra både forskere og bedrifter. Teknologien har opplevd stor vekst de siste årene, noe som har ført til at et bredt spekter av tjenester har migrert til skyen. En slik tjeneste, som stadig øker i betydning er Cloud Gaming, som kombinerer konseptene Cloud Computing og Online Gaming. I motsetning til tradisjonelle dataspill, er selve spillet lokalisert i skyen og ikke på klienten eller spillkonsollen. På grunn av dette har nettverkets ytelse blitt en viktig faktor, som har sterk innflytelse på spillopplevelsen. For de nevnte grunnene er Cloud Gaming en tjeneste det er interessant å se nærmere på.

For leverandører av Cloud Gaming tjenester og plattformer, og for programvare-utviklere (applikasjons-utviklere), er det viktig å kunne tilby og opprettholde god tjenestekvalitet for å tilegne seg nye og beholde eksisterende brukere. Cloud Gaming er derfor avhengig av brukerens krav og villighet til å benytte seg av en applikasjon eller tjeneste og som et resultat av dette er det brukeren som bestemmer om applikasjonen blir en suksess eller fiasko. Med andre ord: uten spillere er det ikke bruk for et spill. Dersom brukere er misfornøyde kan det føre til at de lar være å tilpasse seg en tjeneste og bytter til en annen leverandør, selv om tjenesten fra et teknisk perspektiv fungerer optimalt. Av denne grunn er det viktig for leverandører og utviklere å få nøyaktig innsikt i "Quality of Experience"(QoE) fra brukerens perspektiv, og ikke bare fra et rendyrket tjenesteperspektiv. Selv om QoE ofte blir presentert som et viktig begrep når tjenester migrerer til skyen, er det få studier som undersøker QoE av skytjenester fra et brukerperspektiv.

Målet med denne avhandlingen er å undersøke brukeropplevd QoE av sky-spillet Kahoot!. Applikasjonen er et Classroom Response System (CRS) som kan spilles på enhver enhet med en nettleser. Brukere får poeng ut ifra svartid, som kan være avgjørende for hvem som vinner spillet. Det generelle målet med denne avhandlingen er å undersøke til hvilken grad en bestemt QoS-parameter, forsinkelse, påvirker QoE opplevd av brukerne av en slik sky-basert applikasjon.

For å nå dette målet, har en rekke subjektive brukertester blitt gjennomført i et naturlig miljø samt i et lab miljø.

Forsøket i det naturlige miljøet (N=175) ble utført i forelesninger ved Norges Teknisk-Naturvitenskapelige Universitet (NTNU), og ble gjennomført ved hjelp av gjentagende tester over en tidsperiode på fire uker. Dette oppsettet ble benyttet for å undersøke hvordan og om QoE og forventninger til tjenesten forandret seg over tid. Her ble forsinkelser av ingen- moderat- og høy grad introdusert for studentene.

Lab forsøket ble gjennomført som et tverrsnitts-studie og ble utført på et mindre testpanel (N=21), i et kontrollert lab miljø, der nettverkstilkoblingen til hver student ble manipulert slik at studentene fikk oppleve ulik tjenestekvalitet. Forsøket ble gjennomført for å komplimentere forsøket utført i det naturlige miljøet, og for å undersøke om ulik nettverkskvalitet fører til følelser av urettferdighet blant brukerne. QoE har blitt evaluert basert på den nye definisjonen for QoE foreslått av Qualinet [1], som er presentert i avsnitt 2.1.2. På denne måten har QoE blitt evaluert ved hjelp av subjektive-evalueringer av glede, irritasjon og opplevd teknisk kvalitet. Måling av QoE er vanskelig på grunn av påvirkning av eksterne faktorer, QoE er multi-dimensjonalt da det blant annet omfatter påvirkning av ikke-tekniske aspekter som brukerkarakteristikker og kontekst. Det er vanskelig å analysere alle eksterne faktorer da disse er subjektive og komplekst å dokumentere.

Resultatet fra brukerstudiene har vist at i et urettferdig miljø, der noen studenter opplever forsinkelser og andre ikke gjør det, styrkes følelsen av irritasjon blant brukerne som opplever forsinkelse, mens følelsen av glede er forsterket blant studenter som ikke er påvirket av forsinkelse. Det kom også frem at studentene ble påvirket av tidligere opplevelser, som førte til endring i QoE da lik forsinkelse ble opplevd igjen på et senere tidspunkt.

Denne avhandlingen er et forsøk på å gi et bidrag til litteraturen innen QoE i sammenheng med Cloud Gaming og CRS, resultatene funnet her krever likevel videre forskning og økt validitet gjennom oppfølgings studier. Utover dette har nye spørsmål dukket opp, noe som kan påvirke retninger for fremtidig forskning.

Preface

This thesis is original and independent work by Anlaug Underdal and Marthe Thorine Sunde. The thesis is the final contribution to the Master's degree in Communication Technology at the Norwegian University of Science and Technology (NTNU).

The goal of this thesis is to investigate user-perceived QoE of the cloud gaming application Kahoot!. The general objective of this thesis is to investigate to what extent the QoS parameter delay affects the QoE perceived by users of the cloud-based application.

Gratitude is given to Alf Inge Wang and Morten Versvik at Mobitroll for providing useful information regarding Kahoot!, and for showing interest in the project. Katrien De Moor and Poul Heegaard for contribution and guidance throughout the project, as well as Professor Kjersti Moldeklev and her students in TTM4100 for participating in the testing and making this study possible.

Contents

Appendices

List of Figures

List of Tables

List of Acronyms

CRS Classroom Response System.

FPS First Person Shooter.

HD High Definition.

HTTP HyperText Transfer Protocol.

MMORPG Massively Multiplayer Online Role-Playing Game.

MOS Mean Opinion Score.

NIST National Institute of Standard and Technology.

NRK Norsk rikskringkasting.

NTNU Norwegian University of Science and Technology.

QoE Quality of Experience.

QoS Quality of Service.

RPG Role Playing Game.

RTS Real-time Strategy.

RTT Round-trip time.

SAM Self Assessment Scale.

SPSS Statistical Package for the Social Science.

TCP Transmission Control Protocol.

TTL Time To Live.

Wi-Fi Wireless Fidelity.

Chapter 1 Introduction

Most students know that keeping focus through an early Monday class can be hard. Kahoot! is a newly developed cloud-based Classroom Response System (CRS) where use of an interactive game as a tool can change the learning process to a large extent. Due to its game-based elements, Kahoot! can be considered a Cloud Game and is defined as such for this thesis work. The tool is designed to keep the attention of students and increase engagement in a unique way. Kahoot! is characterized as a cloud-based learning tool and a part of EdTech learning technology[1], a rapidly developing technology [2].

Cloud-based CRS is made possible through the priciples of Cloud Computing, which includes delivery of software, infrastructure and storage over the Internet based on user demands. Due to advantages like quick deployment, (almost) unlimited storage, cost efficiency in addition to easy access, administration and maintenance, Cloud Computing is gaining enormous attention by both researchers and corporations. This paradigm has boomed over the last years and in this context a wide range of services has migrated to the Cloud.

One type of service that is strongly growing in importance is Cloud Gaming, which combines the concepts of Cloud Computing and Online Gaming. A cloud game refers to a game that is located on a cloud server rather than on the gamer's end-device. Cloud Gaming has experienced a tremendous growth the last years, much due to the explosive demand and adoption of tablets and smart phones [3]. The technology of Cloud Gaming enables game playing on various devices since the computational burden of the game is located on Cloud servers. However, as a consequence the performance of the network has become a crucial factor that may have a strong influence on the game experience. Therefore, it is interesting and highly important to look into different Quality of Service (QoS) parameters to investigate and better understand how technical aspects influence Quality of Experience (QoE)

[1] EdTech refers to educational technology and is the study of facilitating e-learning.

of the service. From a technical perspective, when dealing with networks, QoS is useful to guarantee a certain level of performance and to measure related aspects as delay, availability and bandwidth, as QoS focuses on the objective attributes of the network.

Literature on Cloud Computing [4, 5] suggests however that QoE and not only QoS will be a crucial concept in the context of managing quality in the Cloud. According to Schatz et al. [6] the fear of losing consumers to other services or service providers has led to an increased focus on QoE. Users can be seen as the ultimate barometers: if they do not want to use a service, the service is likely to fail. As a result, QoE is typically evaluated through subjective testing, involving real users. Even though QoE is often presented as the guiding paradigm for managing quality in the cloud [4, 5], there are still few studies examining QoE in relation to Cloud Games, from the user's perspective. For this reason this thesis work can make a relevant contribution to the research on QoE in regards of Cloud Gaming.

The goal of this thesis work is to investigate the user-perceived QoE of the cloud gaming application Kahoot!. The application is a Classroom Response System which can be played on any device with a browser. Users obtain scores depending on their answering time, thus the response speed is essential. The general objective of this thesis is to investigate to what extent a specific QoS parameter, namely delay, affects QoE perceived by users of a cloud-based application.

One of the developers of Kahoot! Alf Inge Wang (September 12th, 2013), stated that "the greatest challenge in regards of QoE for the application is slow network". For this reason it is interesting to introduce network constraints and observe how it affects the QoE and the positive and negative emotional states characterized by "delight" and "annoyance", as perceived by the users of Kahoot!. As a result, the objective of this thesis work is to investigate the impact of the QoS parameter delay in regards to Cloud Gaming, looking into how it affects the QoE perceived by users of Kahoot!. The following main research questions have been set: To which degree does delay impact QoE and how does the presence of others influence the experience of fairness?

To answer the main research questions addressed in this thesis work, user experiments has been conducted in both a real-life setting (with a longitudinal setup) and in a controlled lab setting (with a cross-sectional setup), these test will hereafter be referred to as Longitudinal and Cross-Sectional tests. As there is a need to study QoE in a real-life environment due to influencing factors which can be difficult to include in a controlled lab environment, the Longitudinal testing was conducted as part of lecture in TTM4100 at NTNU, quizzing students on the curriculum presented in previous lectures, using Kahoot!. Students were then asked to fill in a questionnaire containing different self-report QoE measures and inquiring after other influencing factors experienced throughout the quiz. The Longitudinal experiment was repeated four times adding different delays in order to find how the students responded to different quality levels and if they noticed the changes in quality as the tests were done as a longitudinal study in weekly intervals.

A Cross-Sectional lab experiment was conducted to complement the Longitudinal study, carried out using students from the same course as the Longitudinal experiment. The students were invited into the lab, where computers (answering devices for the Kahoot! quizzes) were emulated with different delays. The goal of this test was to find how how users react to changes in quality and in terms of perceived fairness. After each test the students were to answer the same questionnaire as given in the Longitudinal testing.

The reason for looking into fairness in a lab setting is to be able to emulate delay on all devices. Fairness is a topic where little research has been done and is an important factor when investigating cloud-based CRS. It is interesting to investigate how students react when subjected to different technical conditions than co-located participants and to find how this affects feelings of delight and annoyance.

Finally the two test setups will bee compared to find how users are affected by delay, when all contestants of a game are affected equally, or in an unfair setting. In addition, we investigate whether the overall perception of the application changes due to the introduction of delays and thus, how (in)tolerant users are in the case of a high delay in the context of this type of service (Kahoot!).

The remainder of this thesis work is structured as follows: Chapter 2 presents related work and theory regarding essential topics. Chapter 3 provides an overview of methodology and the technical setup of the study. In Chapter 4, results and observations are presented and then further investigated and discussed in Chapter 5. Chapter 6 looks into limitations and future work. Finally Chapter 7 summarizes the findings and gives concluding remarks.

Chapter 2 Theoretical Background and Related Work

In this chapter the theoretical background of this thesis work will be presented. The chapter starts by introducing the concepts of QoS and QoE in regards to Cloud Gaming. In addition, a brief overview of relevant related work in the context of Cloud Gaming will be given.

2.1 Definitions and Influencing Factors

There are different perspectives on the measurement and assessment of quality of a service. This section is looking into different perspectives of evaluating quality related to Cloud Gaming, namely Quality of Service and Quality of Experience.

2.1.1 Quality of Service

The most common way of measuring the performance of a service is by considering the network and the service itself. When or how often a service fails and how good the quality of the service is at its best. When in the realm of networks, QoS is a set of standards and mechanisms to ensure high quality performance for systems and applications [7]. There are multiple views and definitions of QoS, the following definition has been issued by the ITU :

> "Totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service" [8].

QoS is related to the end-to-end network service and by this meaning the performance aspects of physical systems [1]. It has for many years been the way quality is assessed in the area of telecommunication. However, this purely technical perspective does not give adequate insights into how a service or application is experienced by users, and which factors have a major impact in this respect. In literature, the concept of QoE has been introduced to capture this highly subjective user-perspective.

While QoS evaluates the actual service delivered, QoE refers to the user-perceived experience of the service. QoE is a different way of measuring performance, to be able to better adapt services to users' preferences and facilitate the best possible QoS as well as creating services that will stay popular on the market.

According to Qualinet[1], QoE is highly dependent on QoS [1] and the technical aspects of a system's performance may have a significantly impact on QoE dimensions. It is for this reason highly relevant to take into account a system's QoS when evaluating its QoE. As Siller et al. explained, QoE can be evaluated using weighted factors given by QoS metrics such as jitter, delay and packet loss [9].

2.1.2 Quality of Experience

Although QoS parameters are highly relevant in the context of Cloud Gaming, the overall theme of this study is to investigate QoE-related issues of Cloud services from a user's point of view. For this reason it is necessary to also establish an understanding of what QoE is in regards of Cloud Computing. As mention above, QoS originates from the telecommunication industry, so does QoE, but QoE is multidimensional and should therefore be considered from a multi-disciplinary perspective. QoE stands out from QoS because of its different nature and focus, and incorporates the possible influence of non-technical aspects such as user characteristics, the context of use [1], and is inherently subjective and individual.

QoE has become more important over the last decades and Cloud applications and services represent an important application domain for QoE research. Literature focuses on what users expect and which factors influence the quality of their experience when using cloud-based services [4, 5, 10, 11]. As of today, the most widespread definition of QoE is derived from ITU [12] and states that QoE is:

> "[The] overall acceptability of an application or service, as perceived subjectively by the end user."

According to this definition, QoE is the subjective perception of the quality of an application. This definition gives a rather vague concept of the 'overall acceptability' as a measure of QoE and in light of this a new and more explicit definition has emerged. A more holistic conceptualization of QoE was proposed by Qualinet [1] with the following working definition:

[1] Qualinet is a European Network of Excellence on Quality of Experience in Multimedia Systems and Services (www.qualinet.eu).

> "Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state."

This means that QoE is defined as an emotional state and evaluation of QoE must take into account the positive and negative emotions resulting from playing. This definition points explicitly to the possible influence of human-related factors, as well as system specific and context-related factors. This definition will be the basis for further evaluation of QoE in this thesis work.

As a new definition on QoE is emerging and covers more aspects, the way to evaluate the term of QoE must as well be re-assessed. According to De Moor et al. traditional QoE measures need to be reconsidered and extended [13]. The Qualinet definition presented above includes the users' degree of delight and annoyance, thus a single focused quality aspect is not sufficient when evaluating QoE. In the wake of the new QoE definition, the understanding of QoE has changed. It implies that QoE is no longer only about satisfying expectations related to the utility of a service or application [13], but also about how users feel, how experiences with technology involve and move people emotionally [14]. It could be assumed that the degree of delight perceived by a user will decrease as the network quality deteriorates, similarly it could be assumed that the degree of annoyance will increase, as was found in a previous study done by Sunde [15]. As QoE is inherently subjective and thus also refers to feelings, expectations, personal relations and motivations, etc., it is complex and difficult to measure, in addition it is fundamentally important to include actual users when attempting to evaluate and measure QoE.

The new definition does not include clear guidelines on how to evaluate QoE and currently, most studies in the literature still make use of the Mean Opinion Score (MOS), which is an averaged measure of perceived quality, to subjectively measure QoE. Questions can be made if this scale is effective enough to represent states of delight and annoyance. According to De Moor et al. the traditional assessment of QoE is insufficiently taking influencing factors into account and QoE is not evaluated in terms of experienced affect: the measures that are traditionally used for evaluating QoE do not provide accurate insight into delight or annoyance [13]. In addition, Hoßfeld et al. indicates that MOS cannot solely be used for QoE management as influencing factors, such as user diversity also needs to be taken into account [4].

In an attempt to put the new definition on QoE into practice, this thesis work will use self-report measures based on multiple items (to ensure reliability and robustness) and measured on Likert-scales to evaluate QoE. After reliability analyses, new variables for degree of delight and annoyance will be computed. In addition, two different test setups will be used to consider the wide range of influencing factors that potentially have an impact on QoE. First a real-life longitudinal study, the importance of conducting such a study in a real-life setting is to include all influencing factors of how a game is usually played. As the setting is preventing control of end devices as well as influencing factors, a cross-sectional lab study is to be conducted to complement the real-life study. More on the questionnaire and use of the Likert scale will follow in Section 3.4, Questionnaires and Subjective Measures.

2.1.3 Factors Influencing QoE

From Qualinet [1] it is said that "QoE is part of the complete ecosystem for the media industry, consisting of creativity (Content), technology (Deliver and Interaction), market/finance (Business models) and users (Usage)". It is also acknowledged that there are multiple factors that may influence QoE. An influence factor is defined as:

> "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"[1].

It is also important to add that it is not just QoS parameters that influence the experience perceived when it comes to Cloud Gaming. Such influencing factors can among other be human-, system- or context related.

Influencing Factors in Game Play

Game players interpret and reflect experience in light of their own desires, anticipations and previous experience [16]. A game can for instance be interpreted as fun, delightful, challenging and victorious until a friend effortlessly makes a better record, then the experience may be reinterpreted more as a waste of time [16]. The Longitudinal setup in this thesis work facilitates such previous experiences by introducing the same condition at two different occasions with other conditions of delay introduced in between, where the conditions are to be presented a week apart from each other.

When interacting with a game, the user's character, skills and needs can influence the experience. Experience is context-related, meaning the same activity might be delightful in one context (playing a game with friends) but boring in another setting (playing a game alone) [17]. Other human factors like personal relationships between the players can also affect the experience. Therefore, both tests in this thesis work include students with relation to each other as they are classmates. Zander et al.

state that "users of Cloud Gaming may tolerate higher QoS degeneration if they have strong relationships to other players, as they become more captivated by the game" [18].

Gajadhar et al. [19] explored the influence of social setting on player experience in digital games. Previously, Ravaja et al. have reported findings indicating that playing against co-located human players elicits engagement and more positively emotional responses than playing against a virtual adversary [20]. Similar results appeared in Gajadhar et al.'s results where participants involved in co-located play against a human co-player reported significantly more positive effect and less tension. Further on Gajadhar et al. believe that the opportunity for social interaction and sharing experiences with others in co-located play enhances game enjoyment. These findings are specifically interesting for this thesis work as Kahoot! is a game intended for classrooms and a social context.

In other literature, these more enjoyment and engagement-related dimensions are not considered. For instance, Jarschel et al. focused specifically on QoS parameters [21]. They argued that only packet delay and loss are relevant QoS parameters to Cloud Gaming and QoE. The time it takes before a user's action is executed and the results that are perceived, are affected by the delay. Other QoS parameters like jitter, packet re-ordering or duplication, to name a few, result in the application not being able to display a video or execute inputs in time [21]. Real-time constrains in Cloud Gaming cannot wait an arbitrary amount of time for one packet to be delivered, therefore packets will be dropped, resulting in loss [21].

Although it would be interesting to investigate the impact of multiple QoS parameters, it has been decided to focus on the impact delay has on QoE in this thesis work. The decision of evaluating delay will be elaborated on in Chapter 3.

Delay as an Influencing Factor

In literature, multiple studies examine games where delay has been added to find how this affects QoE. Zander et al. performed a lab study where artificial network delay and packet loss were introduced during each game session [18]. Their findings shows that players of Quake3[2] have an increasing desire to leave the game as the latency reach 200ms. In [21] it is suggested that delay should be no more than 100ms, this suggestion is confirmed in other literature [22, 23]. According to Claypool et al. [22] different types of games require different thresholds for maximum tolerable delay. Shea et al. [24] presented that the maximum delay a player can tolerate before QoE begins to degrade is 100ms in a First Person Shooter (FPS) game, 500ms in a Role Playing Game (RPG) and 1000ms in a Real-Time Strategy (RTS) game. What

[2] Quake 3 is a multi-player focused FPS video game intended for PC, Dreamcast and PlayStation.

reoccurs in these studies is that delay has been added to online games including animation, avatars and moving objects. For this reason, these applications are more sensitive to delay than other games genres not including these features. It is for this reason interesting to look into the game genre CRS, to find how this genre responds to delay and how high a delay will interfere with QoE.

Chen et al. [25] have analyzed the relationship between QoE and network delay in online games, as an attempt to obtain an overall evaluation of QoE. Their findings show that violation of expectations is a direct cause of degradation of QoE. When a gamer expects an action in a game and this action is delayed, the gamer's experience is violated. If, in addition the player perceives inconsistency between her game environment and the environment of other gamers (due to dissimilar levels of delay), it will lead to perceived unfairness between the gamers [25]. As mentioned a number of factors affect the playability and QoE in Cloud Games. Network transmission impairments like delay, are according to [25] one of the least controllable factors affecting a games playability. For this reason it is crucial for developers to understand and comprehend the effect delay has on QoE when developing games intended for the Cloud.

2.2 QoE in the Context of Cloud Gaming

Cloud Computing refers to the idea of running applications on a server in a data center rather than on individual computers. It facilitates the use of providers computing resources, by giving access to their storage, services and their equipment. Thus, it is cost saving as there is no need to acquire software, and in addition it provides maintenance, security and mobility. Cloud Computing is, according to the National Institute of Standards and Technology (NIST), defined as following:

> "...a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [26].

As the industry of Cloud Computing evolves, a trend of moving more and more services away from the end devices is occurring [21]. With the evolution of mobile networks and devices, the concept of Cloud Gaming is made possible [27].

Cloud Gaming is a Cloud server based approach where Cloud servers carry the burden of executing the gaming engines. Players use a thin client to input game commands which are used to communicate with the Cloud server, thus relieving the end device from the storage and computational burden of the game [24], facilitating rich, multi-player gaming using mobile devices [21]. In order to provide a good experience to the players of Cloud Gaming, the operations mentioned must happen within milliseconds [24], minimizing the delay of these operations is therefore a fundamental challenge for developers of Cloud Games. Users no longer have to purchase powerful hardware to run games and games can be run from almost any device with an Internet connection. This makes it possible to make games available to anyone with a suitable device at a low cost, as well as to reduce the risk of piracy for the software owner as the software never leaves the Cloud [21]. Due to this shift of the computational burden of games, Cloud Gaming may have the stringent demands on network QoS to date, according to Jarschel et al. [21]. For this reasons it is highly relevant to use objective QoS metrics when evaluating the quality of a Cloud Game.

Chen et al. [28] have demonstrated that QoS is a potential indicator of user satisfaction in regards to time spent in an online game. They investigated a so called MMORPG[3], which is a designation of an online role playing game where large amounts of players interact with each other in a virtual world. Chen et al. decided to investigate this game genre because of its slow-past nature (compared to First Person Shooter (FPS) games). They asserted that if they found that network QoS deterioration frustrate MMORPG players, it will definitely also frustrate players of other faster game genres [28]. In collaboration with a provider of a popular MMORPG, Chen et al. reviewed the game traffic over two days (a total of 15 140 game sessions). They compared the network performance with the game play time of each session. Their findings indicated that the more serious the network deterioration is, the sooner players exit the game. In other words, the deterioration of the network causes players to reject the game. By looking into different network impairments (network latency, network delay variations and network loss rate) their research shows that network delay variations are the major reason for player departure [28]. Based on this, it can be said that QoS does affect the quality experienced by the users of Cloud Gaming.

Möller et al. [11] tested a MMORPG by adding a delay, in total they tested six different delays. From their results it was found that setting a higher bandwidth creates a higher variation in quality perceived while setting a lower bandwidth creates a more constant quality. This makes sense as the connection is a best effort service. The lower bandwidth with the more constant network quality was rated higher in regards of the quality of perceived video stream, while the higher bandwidth was rated higher in respect to the input sensitivity by the test subjects. As mentioned above, [28] claims high network delay variations cause players to leave the game.

Depending on the type of game, Cloud Gaming creates new and different challenges.

[3] Massively Multiplayer Online Role-Playing Game

Games including video, as First Person Shooter games or MMORPG, requires both a relative high constant downlink bandwidth as well as low latency [21]. The game used in this thesis work is in a simpler game genre, a CRS with simple static pages including some animations, and is more tolerable to delay than the games mentioned above. A CRS game is different from most Cloud games, but some similarities that can be found is that players receive a reward for doing something right and punished for doing something wrong [29]. In CRS a player receives points for answering correct and fast, while in a more regular Cloud game such as a MMORPG the player is rewarded for completing a task.

2.3 Classroom Response Systems

Interactivity in classrooms has been presented as an important component of learning and is considered a key to success in traditional classrooms [30, 31]. In this setting, oral questioning and answering is the most common form of interaction between students and lecturer [32]. According to Siau et al. [32] this type of interaction faces several obstacles, amongst other the students' reservations to speak out in class. A CRS, also known as Audience Response System, Interactive Response System, Personal Response System, or Student Response System [33], is intended to motivate the students to interact in class by using transmitters (clickers or other thin clients) to capture student votes instantly and display results in class. The system provides anonymity for students and therefore lowers the threshold for participating.

Siau et al. [32] conducted a study where they investigated the interactivity in a course at a university before and after the introduction of CRS. The CRS was designed to improve questioning and answering in the classroom and provide instant feedback to the teacher and students [32]. Results from the study shows that the introduction of CRS significantly increased the interactivity in class. Students participating in the study stated that use of CRS helped to promote class participation and enabled them to assess their understanding compared to other students. In addition, students stated that anonymity increased the students' willingness to participate in class. Another benefit of the CRS mentioned by the students were the elements of fun and play which made the lectures more interesting [32].

Suchman et al. [33] evaluated the impact of a CRS by comparing two similar courses to see if use of a CRS increased student learning, confidence, attendance, and the lecturer's ability to respond to students' misconceptions [33]. The motivation behind this research done by Suchman et al. was the increasing enrollments in courses at colleges and the negative impact that increasing class size has on students grades and course completion rates [34]. According to Vygotsky [35] students need to interact with others in order to make sense of new information prior to internalizing it, but large classes are, according to [33], not organized so that interaction between the instructor and students or among students can take place easily. Suchman et al.'s [33] study found that students attending a course where a CSR was used throughout the course expressed higher confidence levels in their learning and knowledge and indicated that they interacted more with other students than students in the course where use of CRS was only sporadically. In addition, the lecturer of the course where a CRS was used, recorded more modifications to the lecture content due to results from the CRS [33]. These findings are comparable to findings in other studies which shows that use of CRS increase students' attendance, attentiveness, enthusiasm and in-class participation [36, 37, 38]. Other studies have as well showed that students report more enjoyment in class because of CRS and that student confidence in their own learning increases [39].

Little work has however been done on QoE in regards of cloud-based CRS, as most literature is focusing on Cloud Games and online video games. Therefore, it is interesting to look into this topic of how the cloud-based CRS Kahoot! is affected by delay. Some aspects that are interesting to look into is how delay affects the general QoE and how it affects the fairness of the game. As fairness is a topic where very little work has been done in regards of gaming and in particular Cloud Gaming, it is a topic in need of more research. This thesis will therefore look into fairness and how it is affected by delay in a setting of co-located players, and how this affects the delight and annoyance of the players.

Kahoot! can be played on any device with a browser where users can participate in quizzes, discussions and surveys without having to create user accounts. The real-life attribute of the application facilitates easy displaying and retrieval of users' results. A more detailed description of Kahoot! can be found in Chapter 3. Other games that are similar to Kahoot! are Socrative [4] and Clicker [5]. Both these games are based on the same principles as Kahoot!, but have slightly different features. While Clicker is a way of collecting answers, where the lecturer writes the questions and possible answers on the board for the students to answer, Socrative and Kahoot! are cloud-based and more sophisticated games. Questions are integrated in these applications and it is possible to download the results including the questions and what each participant answered.

The specific nature of a social context may significantly influence a player's game experience [40], which is highly relevant in regards of cloud-based CRS. Poels et al. [41] looked into when and why a person is gaming, game immersion, imaginative immersion and how feelings are connected to the effort put into the game. In the context of CRS this can be translated into the effort each students has put into the subject and if they feel they are able to keep track with the progress of the

[4] http://www.socrative.com/

[5] http://clicker.math.ntnu.no/

lecture. CRS is a useful tool for the lecturer responsible to see whether students are paying attention and to evaluate their understanding. From the study done by Poels et al. it was reported that players put more effort into the game when they play against co-located friends, which is the case for the CRS used in this thesis work. Further on, [41] states that a setting playing against co-located friends creates a higher tendency to take 'revenge', which might be a factor creating a competitive and positive learning environment.

Chapter 3

Methodology and Experimental Setup

The research method used in this thesis work is a subjective user study (one experimental and one semi-experimental study were conducted) on how students are affected by delay playing the CRS Kahoot!. The application has been tested by emulating delay on different devices during Kahoot! sessions. Feedback from the test participants was collected through subjective measures. The following chapter presents the methodological setup of the conducted user studies as well as how to interpret and process feedback.

3.1 System Description and Technical Setup

3.1.1 Kahoot!

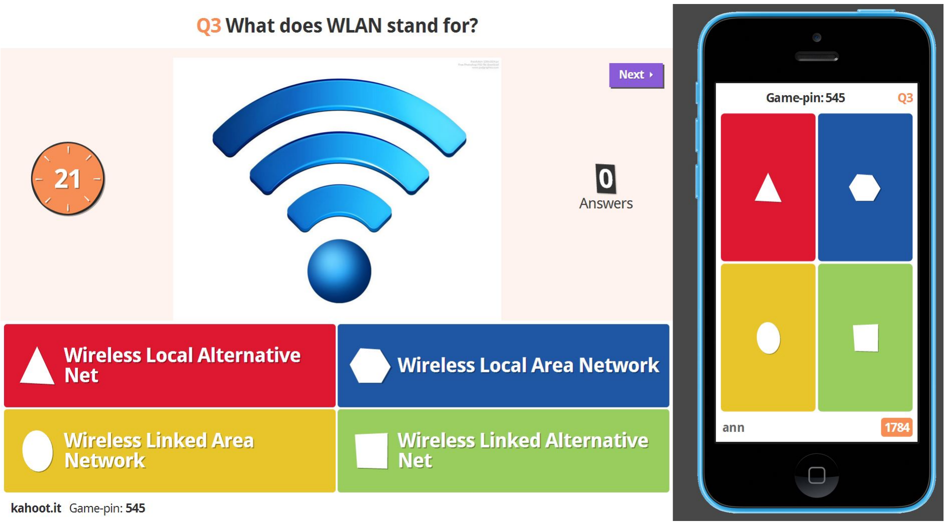
Kahoot! is an up-and-coming game-based learning platform and CRS intended for schools and universities. The application has been developed by Mobitroll, which is a collaboration between Norwegian University of Science and Technology (NTNU) and the British company We Are Human [42]. Kahoot! was first introduced in lectures at NTNU in 2012, and is still in its beta version. The application makes use of blended learning, which combines face-to-face classroom methods with computer-mediated activities, creating an environment where students learn in part through online delivery and self-directed learning [43]. In 2014 Kahoot! was announced the winner of the technology accomplishment of the year by Teknisk Ukeblad [44], and by March 2014 Kahoot! has been played by over 3 million players, and is fast growing with 150 000 new users per week [45].

In an interview with NRK[1], Alf Inge Wang, the creator of Kahoot! describes Kahoot! as "a social learning-game, created for a classroom environment to activate and motivate students, test their knowledge, repeat important elements from a lecture and create a fun and creative diversion from a traditional lecture" [46]. This study

[1] State-owned national broadcasting company in Norway

has chosen to focus on Kahoot! as the application is an interactive game-based CRS used with co-located participants as well as it is functioning through the Cloud and resides in a field where little research has been done.

Kahoot! is a tool that can create quizzes, discussions and surveys by using any device with a web browser. Figure 3.1 displays a screenshot from the monitor and answering device when playing Kahoot!. In a classroom setting Kahoot! is intended to engage the students by involving them in an interactive way by emphasizing elements of fun and play. By projecting a quiz on a canvas or screen in a classroom, students can, without needing an account, join the quiz with their personal device. Questions are answered in real-time and the participants play against each other to achieve the highest score. The fastest answer, given it is correct, collects the most points and as soon as everyone attending the quiz has submitted a response, the scores will appear on the canvas. In addition to making the learning more interactive and varied, the platform also provides the teacher with a greater understanding of the students' current knowledge.



Q3 What does WLAN stand for?

Wireless Local Alternative Net

Wireless Local Area Network

Wireless Linked Area Network

Wireless Linked Alternative Net

Figure 3.1: Screenshot of the monitor and an answering device playing a Kahoot!.

As mentioned Kahoot! is a cloud-based learning platform, which is a field where little research has been done. It is interesting to look into how QoE of the application is affected when subjected to network restraints, in particular delay. For this reason the network constraint delay and Kahoot! has been chosen as the main research areas for this thesis work.

3.1.2 Delay

The goal of this thesis work is to find how delay affects users' QoE. The expansion of the Internet is fast increasing, as more and more people are going online and as the traffic for each user is increasing with the development of new services and applications, delay is of big concern to service providers. Cloud Computing is a paradigm created to deal with this among other problems, by placing the service closer to the user, creating shorter traffic paths, to avoid congestion in the network. For this reason it is interesting to find how cloud services can cope with delay. Accordingly delay was added to different user settings to find how it affected the users. Are users affected differently if all contestants of a game are affected equally, or in a case where some are affected and others are not, creating an unfair setting? How high a delay is tolerable in regard to this type of service (Kahoot!)? These issues will be further discussed in Chapter 5.

After discussion with the developers of Kahoot!, delays used for testing have been chosen carefully by looking at real-life scenarios. In Figure 3.2 a plot of actual latency for Kahoot! is printed. The plot includes latency from a period of 20 days, where one bar is the maximum delay experienced over three hours. As can be seen from the figure, a delay of 10000ms is observed even though it is not that frequently. A delay of 5000ms happens rather frequently and is a realistic scenario.

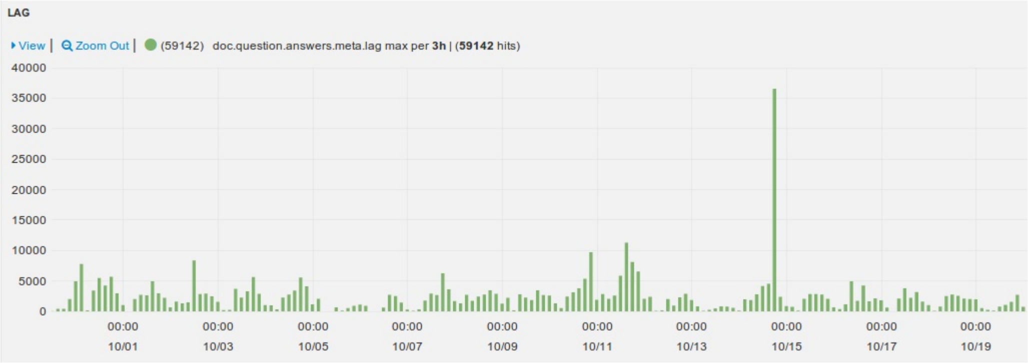


Figure 3.2: Data showing maximum latency from Kahoot!

Three different delays were chosen for this study and can be found in Table 3.1. The reason for adding delays up to 9000ms ± 1000ms can be explained as Kahoot! is a low demanding application. A delay of this degree added to a different environment could lead to a much more fatal result than for this application.

	Delay
No delay	$0ms$
Moderate delay	$5000ms \pm 1000ms$
High delay	$9000ms \pm 1000ms$

Table 3.1: Delays used for testing

A delay of 5000ms ± 1000ms can be considered a high delay, but as this study is looking into a delay rhat is even higher, the delay has been categorized as a *moderate delay* and will be categorized as such throughout this thesis.

Delay was set by manipulating the network using NetEm. NetEm will be explained in section 3.1.3.

The Unix command ping was used to monitor the delay. Ping uses the IMCP protocol's mandatory ECHO_REQUEST to evoke an ICMP ECHO_RESPONSE from the host [47].

```
$ping -c 3 getkahoot.com
```

`-c 3` makes the ping stop after sending 3 ECHO_REQUEST packets. The result from the ping includes the number of bytes sent, IP address of the host, sequence number, Time-To-Live and the time it took the packet to reach the client. Finally a ping statistic shows the minimum, average and maximum time it took for a packet to traverse the network, as well as the Round-Trip-Time.

3.1.3 NetEm

The introduction of different delays in a realistic scenario is done by controlling and manipulating the network conditions. This was done by using NetEm, which provides network emulation functionality, a way to simulate the properties of a network. NetEm facilitates the emulation of variable delay distribution, packet loss, packet re-ordering and packet duplication [48]. The ability to emulate variable delay is how the tool has been used in the testing related to this thesis. NetEm enables real-world scenarios which creates an environment to evaluate the performance of connected devices, services and applications. The network emulator appears to be a network in order for end-systems to be attached to the emulator and behaves as if it was connected to a network with the characteristics decided by the tool. The network emulation tool NetEm is enabled in the kernel of Linux operative system.

The tool is one of the most commonly used network emulators in the research world, being implemented as a queuing discipline in Linux, the tool is easy to deploy [49].

NetEm consist of two components, a small kernel module for a queuing discipline and a command-line utility to configure it. The command-line utility communicates with the kernel through the net-link socket interface [50].

To manipulate the network and create a delay, the following NetEm commands has been used:

```
$sudo tc qdisc add dev eth0 root netem delay 9000ms 1000ms
```

The binary command `tc` was used to delay traffic. In Linux `tc` manipulates traffic control settings that allows control over packets being sent by the computer. `qdisc` is short for "queuing discipline", packets sent from the kernel are enqueued to the qdisc corresponding with the wanted interface `eth0`. `root` provides access to all commands and files and gives the ability to alter the system. In the example above a delay of 9000ms ± 1000ms was added to the outgoing packets leaving the external interface `eth0`.

```
$sudo tc qdisc del dev eth0 root netem
```

The second command was used to disable the NetEm delay by deleting the previous rule. In the real-life Longitudinal test delay was emulated using NetEm on the device running the Kahoot!, while in the Cross-Sectional lab test delay was emulated on each device used by the test participant during the quizzes.

3.2 Real-Life Longitudinal Testing

The real-life testing was set up as a longitudinal study, looking into how users in a real-life setting reacted to different delays and how they experienced changes as they were presented in weekly intervals. Conducting this study as a real-life experiment increases the ecological validity of the results as real influencing factors were present. This testing is an important contribution to the field of studying QoE in a real-life environment as very little work has been done within the subject (QoE) and in particular the real-life setting. As the setting includes multiple limitations in regards of controlling the outcome of the study as well as different influencing factors, a lab study has been conducted to complement the research. QoE is very dynamic and for this reason this study attempts to go beyond the limitations of one test setup, by complementing the real-life experiment with a lab study.

The reason for choosing a longitudinal setup was to find how test participants were affected by previous experiences and whether their perception changed due to this. This was done by presenting multiple delays to the test participants over a period of time and repeating one of the conditions in the final test.

As mentioned delays were chosen after conversations with the creators of Kahoot!. Table 3.2 shows the chosen delay and in which test they were presented.

Test number	**Delay**	**Description**
1	0ms	No delay
2	$9000ms \pm 1000ms$	High delay
3	$5000ms \pm 1000ms$	Moderate delay
4	$0ms$	No delay

Table 3.2: Test number with assigned delay used in real-life Longitudinal testing.

Repeating conditions gives an insight into the results in a different way than when representing multiple conditions only once. It is realistic to take into account the return of a previous condition as well as the results becoming more credible.

3.2.1 Real-Life Longitudinal Test Environment

The test was conducted as part of lecture in TTM4100 at NTNU, which takes place in one of the largest auditorium on campus, R1, with a capacity of 478 persons, Figure 3.3 gives an overview of the auditorium. The auditorium is equipped with all the facilities necessary to perform the test: a canvas, sound system, desktop PC, Wi-Fi connection, to mention some.

The technical specifications of the room includes three base stations as described in Figure 3.3. These base stations support two frequencies 802.11b/g (2.4GHz) and 802.11a (5GHz). Each base station is using one channel from each frequency to prevent noise, from 2.4GHz channel 1, 6 and 11 while channel 64, 116 and 140 from 5GHz. The wired network connection available in front of the room (used with the computer running the Kahoot!) is a 100Mbit connection running on a twisted pair cable connected to the NTNU network.

As a rule of thumb and as long as the users are not downloading big amounts of data there should be no problem connecting 20-30 users per channel, in total the network should be able to support about 180 connections, assuming the connections are spread equally among the base stations. These are not absolute numbers, but used as best-practice to guarantee a stable network connection. On the other hand, there has been an event where about 400 users were connected to the network in this location, some slow network was reported on, but there were no problem to connect all the users. For this reason, it is impossible to set a maximum number of connections in the location (from correspondence with Vidar Stokke, IT-departement NTNU, May 16th 2014).

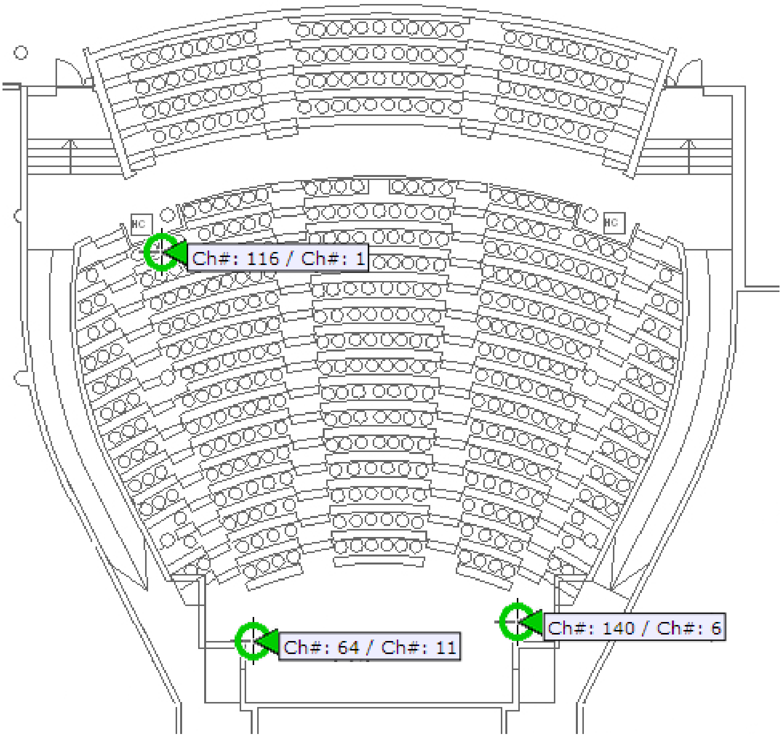


Figure 3.3: Layout of real-life Longitudinal test environment.

As mentioned, testing in a real-life setting can lead to the occurrence of unexpected events. When emulating delay there is always possibilities that external influencing factor can affect the intended test procedure. During both tests where delay was emulated, the connection to the Kahoot! server was lost. It can only be speculated on why this occurred, but as there are network constraints in regards of the number of connections, this might have played a role in provoking the occurrence. In particular this might happen when the traffic occurs in chunks, which happens when all users have the time to decide on their answer, before being able to answer. The effect the disruption had on the test completion will be further discussed in Chapter 4.

3.2.2 Real-Life Longitudinal Pilot Test

Before the Longitudinal test was carried out, a series of small pilot tests were conducted. A pilot test gives a good indication on where the main test could fail and is necessary to make sure that there are no technical problems.

The large sample size of the Longitudinal setup made it complicated to imitate the main test in the pilot. The pilots were executed using different devices, and apart from a substantially smaller test group, were conducted under the same circumstances as planned for the main test. The participants recruited for the pilots were not part of the original test panel. The pilot was run a number of times to verify that the intended levels of delay behaved as expected, in particular testing how the emulated delay affected the execution of the test. As well the questionnaire was tested to find if any question was unclear or other changes had to be made.

As a result of the pilot, the questionnaire was altered, adding the Norwegian translation of all questions. For the emulated delays, no errors were discovered in any of the pilots and the delays decided on in correspondence with the makers of Kahoot! were kept as planned.

3.2.3 Real-Life Longitudinal Sample Description

All students entering the class of TTM4100 were invited to participate during lecture, the number of students registered for the course was 381, with a distribution of 16.3% females and 83.7% males. All participants were students from NTNU, and in the range between 19 and 30 years old, with a mean of 22 years. A total of 175 students participated in one or more of the tests, among these 21% females and 79% male, 124 students participated in the first test, 110 in the second, 85 in the third and 49 in the last test. The distribution of gender did not change significantly throughout the tests.

All students are studying for a masters or bachelor degree in different technologies and are in their second or third year of their study. 66% is students of Computer Science or Informatics. All participants were using their own device for the answering process, 56% was using a laptop, 35% an Android phone or iPhone. This test group was chosen as Kahoot!'s main user base consists of students, they are young, up to date regarding the technology of today, and have an interest in technology due to their study. Robinson et al. [51] used a similar test panel in their "Youth Lab Subjective Testing" where the age group was chosen because of their knowledge and interest in newer technology.

3.2.4 Test Procedure in Real-Life Longitudinal Setting

Four weekly tests were conducted as part of lecture in TTM4100, Communication, Services and Network, a class taught by Kjersti Moldeklev at NTNU. The class is normally taught using Kahoot!, thus it is a known application to the students. The goal was to interfere as little as possible to how the lecture normally is carried out, with the exception of a short questionnaire after the Kahoot! quiz. The purpose of the questionnaire was to gather information about the participants and how they experienced the quiz as well as the performance of the application, Kahoot!.

After a short introduction to the lecture, a Kahoot! consisting of ten questions related to the previous lecture was played, with a timer of 30 seconds per question. The students were not informed that delay was to be introduced, in order to not influence their experience or expectations beforehand in any way. In two out of four tests different network delays were introduced. This was done using NetEm (as previously explained in section 3.1.3) on the computer running the Kahoot! platform. Adding delay to the monitor caused a delay on the traffic leaving the monitor, creating a delay between the timer on the monitor and when the answering devices were able to register a response. If the delay set was five seconds, the possible answering time of 30 seconds will be decreased to 25 seconds, as students would be prohibited from answering during the first five seconds, leading to possible frustration as well as decreasing the possible answer score. The delays added in the different lectures can be found in Table 3.2. To create an incentive for the students to participate as well as to create a more competitive setting it was announced that after the final test, cinema tickets would be given to the three students with the best total score, as well as one random participant. To receive a ticket the student would have to participate in the Kahoot! as well as the questionnaire.

During the quiz, the professor stopped to explain the answer of each question after answers from the audience had been registered. Finally after finishing the quiz, all participants were asked to answer a questionnaire in regards of QoE. The questionnaire will be discussed in more detail in section 3.4.

3.3 Cross-Sectional Lab Testing

Due to lack of control over influencing factors in the real-life Longitudinal setting, a decision was made to conduct a cross-sectional test in a controlled lab environment, complementing the Longitudinal tests. A cross-sectional study is a study that takes place at one specific point in time. This test was conducted to find how different delays affect users in a more controlled test environment, to look into fairness and to evaluate how users react when other users obviously experience better or worse conditions than themselves. As mentioned in Section 2.1.3 the social context is a possible influence factor, but it is not clear to which extent and to which direction it influences QoE [20]. In relation to Kahoot! this is relevant as Kahoot! is used in the social environment of a classroom.

Executing the Cross-Sectional test in a controlled lab environment made it possible to manipulate the network condition of the devices used by the test participants, which was unrealistic in the real-life Longitudinal setting. Adding delay on the test participants' devices enabled further investigation of the experience of unfairness. 21 users, where all except two was enrolled in the class of TTM4100, were invited to participate in the Cross-Sectional testing.

3.3.1 Cross-Sectional Test Environment

The test was set up in a computer lab, Sahara at NTNU. The room was set up with computers running Ubuntu and a layout as described in Figure 3.4. All computers were connected to the NTNU network through 100Mbit twisted-pair network cables.

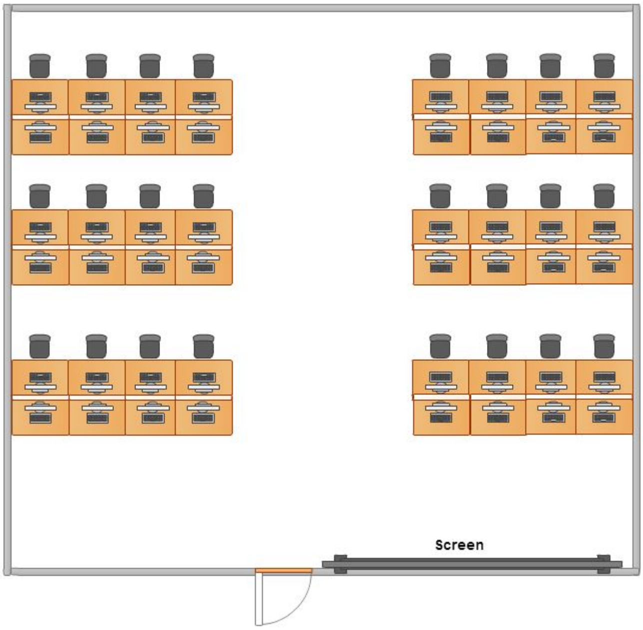


Figure 3.4: Layout of Cross-Sectional test environment, students competing on different delays was seated next to each other.

A projector was set up, screening the questions and answering options of the Kahoot! quiz on a canvas. Three Kahoot! quizzes were played during the test. Each quiz consisted of seven miscellaneous questions of low difficulty level with a timer giving the students 30 seconds to answer each question. Per quiz each student was able to receive a maximum score of 7000 points. The reason for choosing more simple questions was for the students to be able to answer quickly. This might make the participants more aware of delays and was done to trigger possible feelings of unfairness. The students were not aware that delays were added, but lead to believe they were testing the application Kahoot! itself.

3.3.2 Pilot Test

A pilot test was conducted ahead of the actual testing, to evaluate whether the proposed methods and instructions were appropriate and clear. The main goal of the pilot test was to test the different emulated delays to investigate how the higher delays affected the possibility to answer the quiz, trying to make sure the errors of the Longitudinal test would not reoccur. In addition, the pilot test was used to prepare a smooth test execution.

The pilot test was carried out in the same room and under same circumstances as planned for the main test, but on a smaller scale. A group of three recruited pilot test subjects conducted the test and were not part of the original test panel. The test resulted in deciding to stick with the originally planned delays as well as some minor changes in the questionnaire.

3.3.3 Cross-Sectional Environment Sample Description

As mentioned, the test was carried out on a test panel consisting of 21 students. 9 female (43%) and 12 male (57%). All participants were in the range between 17 and 27 years old, with a mean of 22 years old. All participants, with the exception of two were students of technology, studying for their masters or bachelor degree at NTNU. The last two participants were students in their second year of secondary school. All students (except from the two) were enrolled in TTM4100 this semester and had already participated in (parts of) the Longitudinal testing. All participants had used Kahoot! before and knew how the application works. The reason for including two students from secondary school was the lack of available students from the class of TTM4100.

3.3.4 Test Procedure in Cross-Sectional Setting

Ahead of testing, each answering device (computers running Ubuntu) were set to one of the three chosen delays as can be found in Table 3.1 on page 18, giving a number of test participants competing on each delay. The number of students competing on each delay as well as the different delay orders are described in Figure 3.3. Computers were tagged with the names of the test participants and the test participants were asked to sit down by the computer with their name on it. The approach used in the

Cross-Sectional test differs from the real-life Longitudinal testing by adding delay to the device used by the participants to answer the quiz, instead of on the device sending out and receiving the questions/answers. This created an unfair setting where participants experienced different delays. According to [25] unfairness refers to the degree of difference among players in the same gaming environment. Delays were assigned to computers in a systematic order to make sure that test participants were situated close to participants with different conditions, ensuring the experience of unfairness.

No. of students	5	2	3	4	4	3
Test 1	no	no	moderate	moderate	high	high
Test 2	moderate	high	no	high	no	moderate
Test 3	high	moderate	high	no	moderate	no

Table 3.3: Delay order students were competing on during testing.

A canvas screening the Kahoot! quiz questions with answer options was presented in front of the class for students to answer. After the quiz, all test participants were to answer a questionnaire (presented in section 3.4) reporting their perceived QoE. The test participants were then asked to leave the room for a couple of minutes. Before re-entering the room, each answering device was set to a different delay and the users were asked to sit down by the same computer they just left. A new Kahoot! quiz was then carried out in the same manner as the first quiz with a corresponding questionnaire. The students were again asked to leave the room for recomputing of delay. In total, three Kahoot!s with corresponding questionnaires were conducted to make sure that all participants experienced all different delays. This was done to make it possible to identify and investigate possible order effects.

3.4 Questionnaires and Subjective Measures

3.4.1 Questionnaire

A questionnaire was created to gather standard demographic information (age, gender, study program, etc.), as well as subjective measures of QoE from the Kahoot! quiz. In all tests, the questionnaire was answered after playing a Kahoot!. Multiple questions used the Likert scale, ranging from 1 to 5; with 5 being strongly agree/extremely; 4 agree/fairly; 3 neutral/moderately; 2 disagree/slightly; and 1 strongly disagree/not at all. The Likert scale is a technique for measuring attitudes [52] set forward by Rensis Likert. According to SurveyMonkey [53], the Likert scale is the most popular and reliable used approach to scaling attitude in survey research.

When creating the questionnaire, inspiration was taken from [41] which used a focus group methodology to explore feelings and important aspects influencing users' experiences while playing a game. The questionnaire included questions regarding how the test panel felt while playing the Kahoot! quiz. Questions including similar emotions such as for example, frustrated and annoyed were used to evaluate whether the panel was consistent while answering the questionnaire.

The developed questionnaire aimed to gather feedback relevant for QoE from the students. To operationalize this, the questionnaire consisted of different blocks and intended to gather feedback regarding delight, annoyance, quality, and positive and negative emotions the impact of co-located participants had.

In the working definition of QoE described by Qualinet [1], the degree of delight and annoyance are described as important characteristics of QoE (definition can be found in Chapter 2). For this reason, variables describing emotions of delight and annoyance are part of the questionnaire. The delight and annoyance blocks were comprised of adjectives regarding the impact the Kahoot! session had on the students emotions. As an example, the students were asked to rate in which degree they felt happy, entertained, frustrated, bored and so on, during the Kahoot! session.

In order to evaluate perceived quality during the Kahoot! session, various statements were included and students were asked to agree or disagree to these. The statements referred to the technical quality aspects of Kahoot!, as an example one statement claimed that Kahoot! functioned optimally, another that the overall technical quality was unacceptable. Negations were included deliberately in order to investigate the reliability and consistency of the answers to these statements.

For the block regarding the positive and negative effect that co-locates participants had, the students were instructed to indicate to which extent they (dis)agreed with the various statements regarding this effect. Some of the statements chosen were for example, whether the student felt that other participants stressed them, made him or her laugh or more competitive. This block was included to examine whether the presence of co-located participants influenced their perception of fairness during the session. In addition, students were asked how they believed they performed compared to their neighbors. This was asked to find whether they paid attention to co-located participants and whether they, in particular for the Cross-Sectional testing, noticed the setting of unfairness as students were playing on different network conditions.

In addition to the above mentioned blocks, the questionnaire collected feedback on the students overall impression of Kahoot! and on their experience of delay between the monitor and their own device. A complete list with all the adjectives and statements used in the different blocks are listed in Table 3.5 on page 30, and the questionnaire can be found in Appendix D.

Task Performance

As the questionnaire is a self-report measure, the score from the Kahoot! quiz is used as a more objective measurement and a task performance indicator. The results from the quiz can confirm whether students answered all questions of the quiz and how they scored, as well as answering time. The answering time can partly indicate to which degree the students were affected by delay as a high answering time can be an indication of delay.

The self-report questionnaire included questions asking how students believed delay affected their score, and whether their final quiz score was as expected or in line with their effort. Together with the objective measurement of their actual score and the answering time, the self-measurement could indicate how a student was affected by the condition of the Kahoot! session.

3.4.2 Overview of Techniques and Procedures Used for Analyzing Data

To examine the data collected through the self-report questionnaire, different tests were run on the data to investigate where significant differences occur and to support visually inspected findings.

The tool SPSS created by IBM was utilized to prepare data and check for consistency, reliability, significant differences and significant correlations in the collected data from the questionnaires. A range of analysis techniques and tests were conducted to analyze the data. These include: Cronbach's alpha, Friedman ANOVA, Wilcoxon signed-rank test, the Mann-Whitney U test and Spearman's rank-order correlation were applied to results to analyze the data.

Some of the tests are looking for significant differences or correlations, these might be found high but not be significant or significant yet low. Significant in this sense means that the probability that the obtained test statistic just occurred by chance, is lower than 0.05. However, when a result is not significant, it does not mean that there is definitely no effect or no difference/correlation between the investigated conditions, it only means that the effect that was found in the investigated data is not substantial enough to argue that it may not be based on chance.

SPSS was also used to create boxplots for better visualization of the findings. These tests will be further explained in the next sections and the results will be presented in Chapter 4.

Statistical Analysis Software (SPSS)

IBM SPSS Statistic is an analytical software for statistical calculations [54]. The tool enables comprehensive analytical processes including descriptive statistics, cross tabulations, frequency analysis, exploratory data analysis, ANOVA, non-parametric tests and more [55].

Using SPSS the frequencies of responses from a questionnaire can be investigated to prepare data for further analyze and to systematically try to eliminate errors. The frequency function in SPSS creates frequency tables as well as a variety of graphs and chart which can be examined to analyze the data.

Cronbach's Alpha

Cronbach's alpha analyses internal consistency and indicates how closely related items are as a group [55]. Such items can be adjectives representing similar emotions or feelings such as happy and satisfied or different sentences asking similar questions, examples can be found in Table 3.5.

When using multiple Likert questions in a questionnaire that form a scale, Cronbach's alpha is useful to determine whether the scale is reliable or not, and whether the items in the questionnaire measures the same variable, and can thus be grouped together or not. When constructing a questionnaire, it is necessary to include consistency checks. This is done by including statements intended to measure the same underlying construct but using a different wording and including negations. The next step is then to check whether the test participants were consistent in their answers and whether these multiple items measure the same underlying construct or not.

A high alpha value indicates that the items are measuring the same latent variable which implies that the items within a category are consistent [55]. As shown in Table 3.4 an alpha value above 0.6 is considered as acceptable and if multiple items share a value above this, these items can be computed into a new variable. For this variable the score is the average of the ratings for all included items making the measure more robust as it is not based on one single item. If a number of items receive an alpha value above 0.6 a new variable is calculated by summarizing the items of each participant and dividing the result by the number of consistent items.

Cronbach's alpha	**Internal consistency**
$\alpha \geq 0.9$	Excellent
$0.7 \leq \alpha < 0.9$	Good
$0.6 \leq \alpha < 0.7$	Acceptable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

Table 3.4: Cronbach's alpha values representing levels of internal consistency.

The Cronbach's alpha also provides an overview of how the removal of one item will affect the alpha value [54]. This overview is useful to investigate whether an item is not consistent with the others and should be removed from the category. In Table 3.5 multiple items used in the questionnaire are listed. From each test result Cronbach's alpha was run on these categories to find if the items were consistent and could be translated into new variables. The items that are part of Degree of Delight and Degree of Annoyance were measured on a scale ranging from 1 to 5, with 1 indicating 'not at all' delighted/annoyed, and 5 indicating 'extremely' delighted/annoyed. The remaining items in Table 3.5, were also measured on a 5-point scale, but with 1 indicating 'strongly disagree' and 5 indicating 'strongly agree'.

Degree of Delight	Happy, Amused, Satisfied, Delighted, Entertained
Degree of Annoyance	Tense, Bored, Irritated, Annoyed, Frustrated, Disappointed
Evaluation of Quality	Kahoot! functioned optimally, The quality of Kahoot! was good, The technical quality of Kahoot! was as it should be, I did not experience delay between monitor and device
Positive Influence	Made me laugh, Made me feel good, Made me more competitive
Negative Influence	Scared me, Stressed me, Distracted me, Embarrassed me

Table 3.5: Items to be checked for internal consistency.

As mentioned, Degree of Delight and Annoyance are important indicators of

QoE, and are among the variables that were evaluated using Cronbach's alpha as can be seen in Table 3.5. The same has been done for variables of positive and negative influences in regards of the impact of others. As stated in Chapter 2, QoE is the subjective perception of quality, for this reason the items regarding the overall technical quality of Kahoot! in the questionnaires were combined into the new variable called Evaluation of Quality. When using the term Evaluation of Quality further on in this thesis work, it is oriented towards the technical quality and acceptability of Kahoot!. According to the most widespread definition of QoE presented in Chapter 2, QoE is the overall acceptability of a service as perceived subjectively by the user [8].

Friedman ANOVA

Because of the measurement level of the subjective measures (i.e., considered as ordinal variables) and because non-parametric data do not allow the use of parametric tests, non-parametric tests were used for the statistical analysis of the data. Non-parametric tests make fewer assumptions about the data and strictly speaking, when analyzing non-parametric data, typical descriptors such as variance, mean and standard deviation cannot be not used. Instead, non-parametric tests are based on the idea of ranking the data.

The Friedman ANOVA test was used when testing differences between multiple conditions where the same participants participated in all conditions. To run a Friedman ANOVA test in SPSS the following assumptions must be considered; the test group is measured on three or more different occasions, the group is a random sample from the population, the dependent variable should be measured at the ordinal or continuous level and the sample does not need to be normally distributed [54]. In circumstances were these requirements are met, Friedman ANOVA test will be used to determine whether significant differences exists between different conditions. The test gives a significance level, also called a p-value, if this value is less than 0.05 it means that there exists a significant difference between two or more variables in the sample.

The Friedman ANOVA test was used after items with high internal consistency ($\alpha > 0.6$) from the Cronbach's alpha test were combined into a new variable, and investigate whether there existed any significant differences between the three levels of delay, i.e. *no delay*, *moderate delay* and *high delay.*

Wilcoxon Signed-Rank Test

In cases where dependent variables are measured at an ordinal or continuous level and when two related conditions are compared, the Wilcoxon signed-rank test can be applied [54]. In this thesis a 5-point Likert scale is used, thus dependent variables are measured at an ordinal level. The Wilcoxon signed-rank test is a statistical comparison of two dependent samples and compares two sets of scores where individuals have been subjected to both conditions.

The Wilcoxon signed-rank test determines whether there is a significant difference between the two conditions tested by providing a p-value similar to the one found by Friedman ANOVA. The p-value determines how similar or different two conditions are, if the p-value is less than 0.05 there is a significant difference between the two conditions tested.

Mann-Whitney U Test

As mentioned, the Wilcoxon signed-rank test requires that all participants have participated in both conditions to be tested. As there was no guarantee all students participating in the real-life Longitudinal test were to participate in all the tests scheduled, a Mann-Whitney U test was run on the unrelated variables of the sample, where the same users did not participate in two tests that are compared. The Mann-Whitney U Test is used to compare differences between two independent groups where the dependent variable is either ordinal or continuous, but not normally distributed [54].

The Mann-Whitney U test in SPSS generates a similar table of ranks as the Wilcoxon signed-rank test and provides a comparison of two conditions. As well as finding whether there are significant differences between the different distributions, the test compares mean ranks as the distributions used for the tests have different shapes. Again a significant difference is found if the p-value is below 0.05.

Spearman's Rank Correlation Coefficient

To explore possible relations between the alternative measures used to evaluate QoE in this thesis work, correlation analyses were used for some of the measures from the questionnaire as well as the subjective measures of QoE from Table 3.5.

Spearman's rank correlation coefficient is a non-parametric statistic and can be used when the data have violated parametric assumptions such as non-Normality [55]. The data is ranked before applying the Pearson's equation to the ranks. The Pearson equation calculates the *Pearson correlation coefficient* and indicates to which degree two variables correlate (ρ=1 perfect correlation, ρ= -1 perfect negative correlation, ρ= 0 indicates no linear relationship)[55]. The test results tell whether there is a low or clear correlation between the ratings of the different variables as well as whether the correlation is significant. A correlation is found to be significant if the significance level is below 0.05. When a correlation is found significant, this means that it is unlikely that the correlation happened by chance. The test indicates whether or not there is a correlation and in which direction, but it does not tell anything about the causality in the observed relationship.

Summary of Statistical Analysis

Method	Conditions	Applied in thesis
Cronbach's alpha (*Internal consistency*)	Items measured on equal scale	Determine if scale is reliable and items can be combined into new variables
Friedman ANOVA (*Significant differences*)	Compares 3 or more related samples, Dependent ordinal variables, Does not need to be normally distributed	Analyze differences between test results in Cross-Sectional test
Wilcoxon signed-rank (*Significant differences*)	Compares two related samples, Dependent ordinal variables, Dist. symmetrically shaped	Analyze differences between related test results
Mann-Whitney U test (*Significant differences*)	Compares two independent samples, Dependent ordinal variables, Determine if shape of the distributions is equal or not	Analyze differences between unrelated test results
Spearman's rank correlation coefficient (*Correlation*)	Ordinal variables, Monotonic relationship	Measure strength of association between two ranked variables

Table 3.6: Summary of statistical analysis. The text written in *italics* is the purpose of the test.

3.4.3 Methods of Presenting Statistics

Boxplots

Boxplots generated in SPSS can be used as visualization for viewing how data are distributed. Due to the use of boxplots in this thesis work a short review on how to interpret these plots will be given. Figure 3.5 is an example of a boxplot showing the distribution of gender and will be used to illustrate how to interpret such a plot. In this example a group of students have been asked to rate their experience of Degree of Delight in relation to playing Kahoot!.

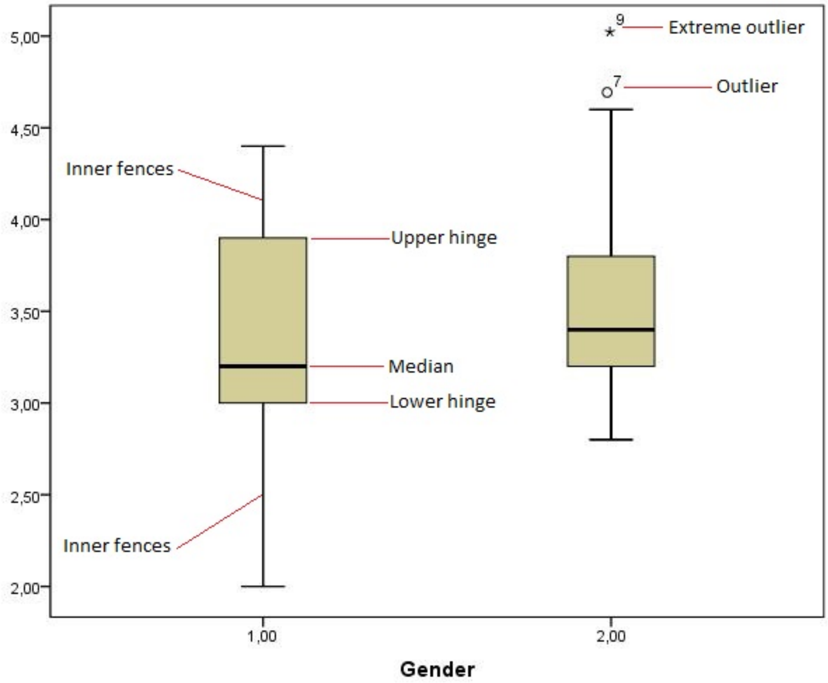


Figure 3.5: How to interpret boxplots.

As can be seen in the figure the y-axis denotes a scale ranging from two to five. In this thesis work the y-axis is indicating the steps of the Likert scale. In this example the x-axis denotes gender, where one indicates male and two female.

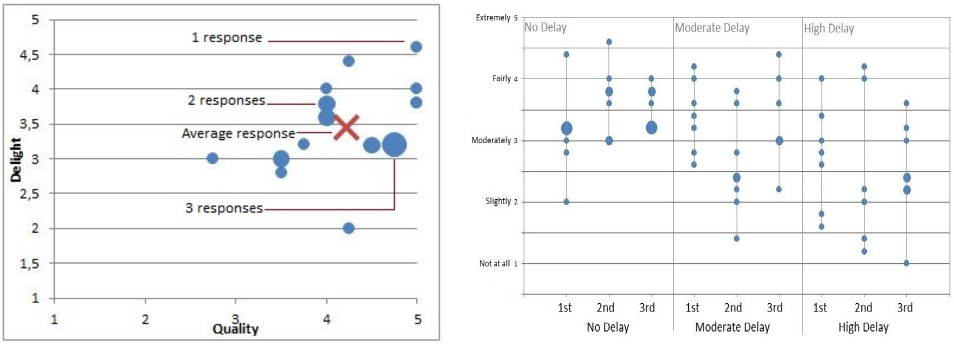
The bold horizontal line in the middle of the two boxes is the median, meaning that half of the samples have a value greater than the median, and the other half a lower value. The bottom line of the box, namely the *lower hinge*, is the 25th percentile and indicates that 25 percent of the data set has a value below this line. The top line of the box, the *upper hinge*, is the 75th percentile indicating that 25 percent of the samples have a value greater than this line. Thus 50% of the samples lies within the box which is referred to as the *interquartile range.* As can be seen in Figure 3.5 the interquartile range, for males is bigger than the interquartile range for females, indicating that the experience of delight varied more for males than for females.

The lines in the boxplot resembling a 'T' which extends the boxes, are called *inner fences*. These extends 1.5 times the interquartile range, or, if no samples have values in that range, to the minimum or maximum value of the samples. It is expected that about 95% of the sample resides within the inner fences. In cases where the median is located in the center of the interquartile range and the two inner fences are equal in length, the distribution is symmetric. When this does not occur it indicates a potential positive or negative skewness of the distribution. In this example the inner fences extend further from the interquartile range for males than females, which is another indication that delight varied more for males than for females.

If there exists values greater or less than the inner fences these values will be denoted by circles or asterisk and is defined as *outliers*. Outliers residing more than 1.5 times the interquartile range from the lower or upper hinges are illustrated as circles and outliers more than three times the interquartile range away are illustrated by asterisk, and is referred to as *extreme outliers*. The numbers next to the outliers represent the case number in SPSS and can be used to further analyze these events.

Dotted Plots

To visualize correlations and investigate possible order-effects dotted plots has been used. Figure 3.6 displays example plots for correlation (3.6a) and order-effect (3.6b).



(a)

(b)

Figure 3.6: How to interpret dotted plots.

Figure 3.6a displays how evaluation of perceived technical quality and delight correlates. As some responses might end up on the same point in the plot, different sized dots are used to mark this. As marked in Figure 3.6a, a smaller dot represents one response, a slightly larger dot represents two responses and a large dot represents three responses, this is also the case for the order-effect plots as presented in Figure

3.6b. The large red cross presents the correlation between average evaluation of perceived quality and delight.

The example plot in Figure 3.6a displays responses where both quality and delight has been rated high, as the majority of the dots resides in the upper right corner above a rating of 3 on both axis, showing there exist a degree of correlation between ratings of quality and delight. If the dots had been located in the bottom right corner, quality would have been rated high while annoyance was rated low, creating a negative correlation between the variables. How high a correlation is, can be found by investigating corresponding Spearman's rank correlation coefficient, the test also finds if the correlation is significant or not, as explained in Section 3.4.2.

Figure 3.6b investigates possible order-effect by presenting how students rated delight for each delay grouped by when the delay was introduced (the Cross-Sectional test introduces different delays to students through a series of three tests). As mentioned, different sized dots are used, marking if multiple students rated delight equally. Delight is measured on the Y-axis and the higher the dot is located, the higher is the measured delight. What is interesting to investigate in this plot is the similarities and differences within each delay, an example can be seen in the *no delay* section of the plot, where delight is rated higher and is more concentrated when presented last (after being presented with *high-* and *moderate delay*) than when *no delay* was presented first.

Chapter 4 Results and Observations

In this chapter results from tests in the Longitudinal and Cross-Sectional settings are to be presented and attempt to answer the research questions raised for this thesis work, investigate user perceived QoE and to what extent delay affects perceived QoE. The Longitudinal study investigates whether and how QoE changes over time under different network conditions, while the Cross-Sectional study examines how presence of others influences the experience of fairness.

The chapter starts by looking into results from the Longitudinal tests before proceeding with results from the Cross-Sectional testing. The results will be further discussed in Chapter 5.

4.1 General Findings

In both test settings, students were asked about their general impression of Kahoot!. Regardless of setting (Longitudinal/Cross-Sectional) or network condition, more than 70% reported they have a positive impression of the application, characterizing it as a useful learning tool creating a nice diversion during lecture. It is interesting to find that delay does not affect the impression of Kahoot! and this will be further discussed in Chapter 5.

It has been investigated to which extent there are found differences in how males and females rate their degree of delight or annoyance as well as perceived technical quality. Results from these analyzes showed no interesting differences and will for this reason not be further elaborated on for either settings. In addition, for the real-life Longitudinal setting, the findings would not have been strong enough to draw conclusions due to the skewed distribution of gender (approximately 20% females).

4.2 Real-Life Longitudinal Testing

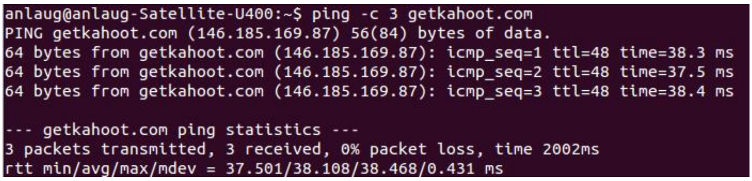
As mentioned previously, the Longitudinal setting consisted of four tests conducted over a period of four weeks. The first and fourth test was carried out without adding any delay, while the second and third test were administered with delays of respectively $9000ms \pm 1000ms$ and $5000ms \pm 1000ms$. For all tests the majority of the students were in the range between 20 and 30 years old with a mean of 22 and the gender distribution was on average 21% female and 79% male.

This section starts by presenting the two tests conducted without delay (the first and the forth test), followed by presenting the tests were delays were emulated (the second and third test).

4.2.1 No Delay

As mentioned, the condition of *no delay* was tested at two occasions, the first and the last out of four tests. Both tests were conducted as intended without any remarks or interrupts, and the average answering time related to the Kahoot!s were accordingly 7.2 (test 1) and 6.0 (test 4) seconds.

The condition of these two tests is characterized as "normal" as no delay was emulated. To find what "normal" indicates, a ping was done to getkahoot.com on the computer running the Kahoot!. This ping is displayed in Figure 4.1, which shows that an average round-trip time for a "normal" condition of approximately 38ms.



```
anlaug@anlaug-Satellite-U400:~$ ping -c 3 getkahoot.com
PING getkahoot.com (146.185.169.87) 56(84) bytes of data.
64 bytes from getkahoot.com (146.185.169.87): icmp_seq=1 ttl=48 time=38.3 ms
64 bytes from getkahoot.com (146.185.169.87): icmp_seq=2 ttl=48 time=37.5 ms
64 bytes from getkahoot.com (146.185.169.87): icmp_seq=3 ttl=48 time=38.4 ms

--- getkahoot.com ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2002ms
rtt min/avg/max/mdev = 37.501/38.108/38.468/0.431 ms
```

Figure 4.1: Ping from the condition *no delay*, showing an average RTT of 38.108ms.

First Test with No Delay

In the first test that was conducted, a total of 124 students participated in the Kahoot! quiz and 98 of these answered the questionnaire afterwards.

When asked whether the students answered all the questions of the Kahoot! quiz 90% said yes. The remaining indicated that the reason for not answering all questions was due to human error, for example one student showed up late missing the first question, another student unintentionally closed the browser with his Kahoot! and had to re-enter the quiz. 34% reported that they agreed or strongly agreed to have experienced delay between the monitor and their device. As no delay was emulated in this test and users were using their own devices it is impossible to confirm their statement: did users experience delay, or did the application function differently than expected by the users? The students were also asked to indicate whether they would have received a higher score if the technical quality of the Kahoot! session had been better. 56% disagreed or strongly disagreed, indicating the technical quality of Kahoot! did not have an impact on their scores. 14% agreed or strongly agreed that their scores would have been better if the technical quality of the Kahoot! session had been better. Why as many as 14% had this claim is interesting and will be further investigate in Chapter 5.

Last Test with No Delay

Ahead of test number four, also without delay, the students had been introduced to two tests where different delays had been emulated. In test number four, 49 students participated in the Kahoot! quiz and 34 out of these completed the questionnaire. The main reason for the low attendance in this test can be explained by the overall low attendance in class at this time of the semester. Conversations with the lecturer of TTM4100, explained that the descending number of students participating in class throughout the semester is a normal trend observed every semester. The main reason for this could be the intensive workload with multiple projects that these students were working on at the time of testing and the fact that participation in lecture is voluntarily.

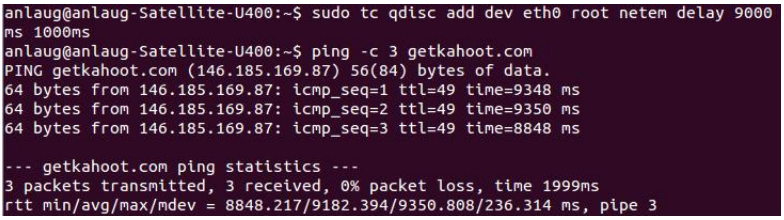
When asked whether the students answered all the questions in the Kahoot! quiz of the fourth test, 82% answered yes. One person claimed he left the quiz due to technical difficulties while the remaining did not answer all questions due to human errors. Thus, out of 34 responses only one indicated technical issues which implies that the overall quality of Kahoot! was good.

Whereas in the first no delay test, 14% of the students indicated that they agreed or strongly agreed to the statement saying they would have received a higher score if the technical quality of the Kahoot! session had been better, no one agreed to this in the fourth test. 15% of the students in test four answered neutral to this question while the remaining 85% disagreed or strongly disagreed to this statement indicating that the technical quality had nothing to do with their results of the Kahoot! quiz. In addition, 12% reported that they experienced delay between the monitor and their devices, compared to 34% in test one. One possible explanation might be that this difference in answers is influenced by the introduction of delay in the two previous tests.

4.2.2 High Delay

In the second test conducted, a delay of 9000ms ± 1000ms was emulated on the computer running the Kahoot! A total of 110 students participated in the Kahoot! quiz and 108 of these answered the questionnaire afterwards.

After launching the Kahoot! quiz and letting the students connect to the quiz, delay was emulated on the output from the computer running the quiz (displaying the questions). How delay was added and a corresponding ping to getkahoot.com showing an average round-trip time of 9182.394ms, can be seen in Figure 4.2.



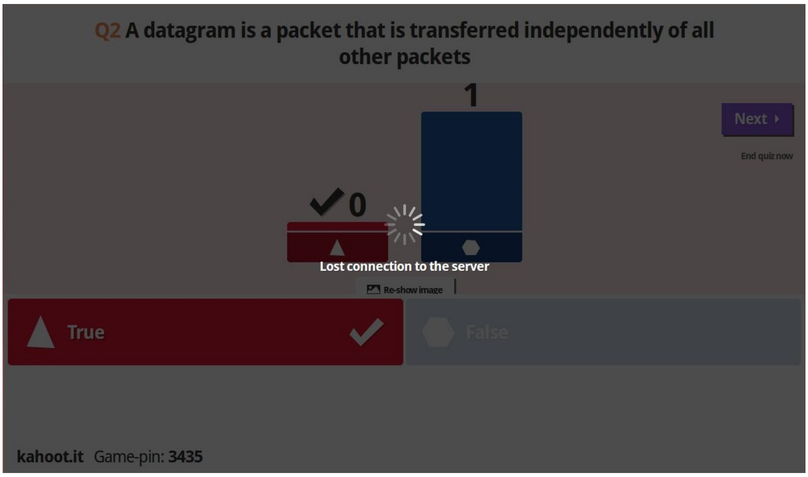
```
anlaug@anlaug-Satellite-U400:~$ sudo tc qdisc add dev eth0 root netem delay 9000
ms 1000ms
anlaug@anlaug-Satellite-U400:~$ ping -c 3 getkahoot.com
PING getkahoot.com (146.185.169.87) 56(84) bytes of data.
64 bytes from 146.185.169.87: icmp_seq=1 ttl=49 time=9348 ms
64 bytes from 146.185.169.87: icmp_seq=2 ttl=49 time=9350 ms
64 bytes from 146.185.169.87: icmp_seq=3 ttl=49 time=8848 ms

--- getkahoot.com ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 1999ms
rtt min/avg/max/mdev = 8848.217/9182.394/9350.808/236.314 ms, pipe 3
```

Figure 4.2: Ping from the condition of *high delay*, showing an average RTT of 9182.808ms.

As the first question was displayed, noises of irritation could be heard from the audience and from the main screen it could be seen that no responses had yet been registered. By the time the timer in Kahoot! reached 30 seconds, a total of 18 answers were registered with an average answering time of 18.5 seconds. This was due to the emulated delay, not giving the students the ability to answer as quickly as they would have, without delay. When preceding to the second question a message popped up on the canvas saying "lost connection to the server", more people sounded frustrated and someone shouted that they were not able to answer. A total of three responses were registered. Figure 4.3 shows a screenshot of the error message appearing on the canvas. To reconnect with the server, a quick decision was made to remove the delay, allowing students to reconnect and to keep the flow of the lecture.

To reconnect each student had to refresh their browser and re-login to the quiz using the game-pin. By the next question most of the students were able to reconnect with the quiz using the same username as they used before connection was lost. This is a good example of how control over unexpected influencing factors is lacking in tests conducted in a real-life setting. To avoid interrupting the lecture in a too great manner it was decided not to add any delays before the last question. Before launching the final question (question 10) delay was emulated again, for the students to keep the experience of the delay fresh in memory before answering the questionnaire. By paying attention to the canvas it was observed that most of the answers were registered only a couple of seconds before the timer ran out. 58 out of

110 participants were able to answerer this question in time, with an average answer time of 25.4 seconds.



Q2 A datagram is a packet that is transferred independently of all other packets

Lost connection to the server

Figure 4.3: Error message displayed on the canvas when loosing connection to the Kahoot! server.

Even though the test was not conducted according to plan, the situation of the lost connection was a consequence of the delay and is not an unrealistic scenario. The results can still be found interesting, and as the loss of connection made a major impact on the students and as quality was rated low by most participants, the results have been treated as a condition of *high delay.* In correspondence with one of the creators of Kahoot!, Morten Versvik, he states that it is interesting for them to find how QoE is affected in events of total disconnection (September 13th, 2013).

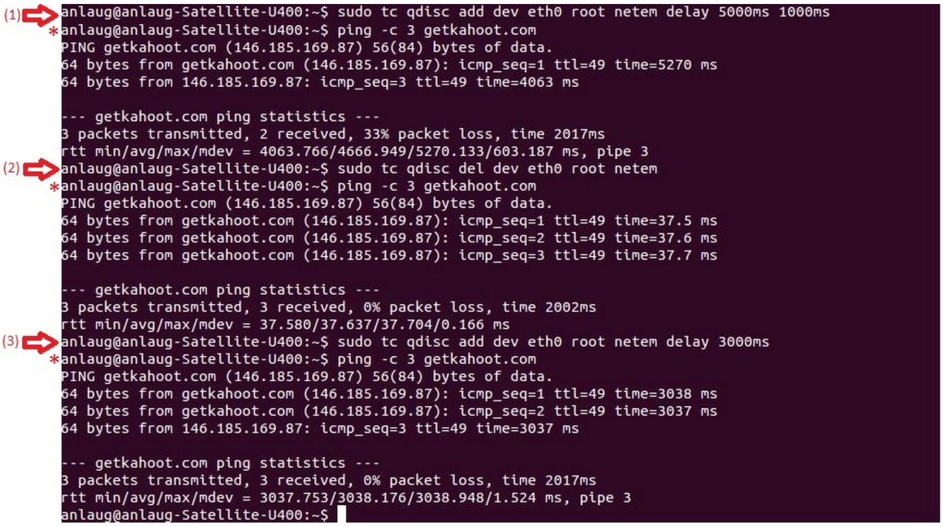
When asked whether the students answered all the questions of the Kahoot! quiz as few as 7% answered yes. 46% stated they left the quiz because of technical difficulties and 24% said they were thrown out of the quiz. Multiple students commented that they experienced server issues. As a result, a total of 72% agreed or strongly agreed they would have received a higher score if the technical quality of the Kahoot! session had been better while 12% disagreed or strongly disagreed to this, meaning their score reflected their knowledge.

86% reported they agreed or strongly agreed to have experienced delay between the monitor and their answering device. Multiple students commented on a high delay between the monitor and their device which made them stressed to answer or having to guess due to the short answering time left when their device displayed the answering options. In addition, multiple students commented on being thrown out of the quiz multiple times.

4.2.3 Moderate Delay

In the third test conducted, a delay of 5000ms ± 1000ms was emulated on the computer running Kahoot!. A total of 85 students participated in the Kahoot! quiz and 53 of these answered the questionnaire afterwards.

After letting the students connect to the quiz, delay was emulated. The first question was displayed as expected and the students were able to submit their answers. By the time scores from the first question were displayed and students prepared for the next question, the device running the Kahoot! quiz lost connection to the server, as happened in the test with *high delay*, causing all participants to loose connection. To solve the problem, delay was removed for the students to be able to reconnect, making sure all students were able to answer the next question and precede the lecture. This is supported by the result of the Kahoot! quiz as 70 answers were registered for the second question. In an attempt to carry out the rest of the test with delay but without interruptions, a delay of 3000ms was emulated and kept successfully throughout the quiz. The average answering time per question of this test was 12.4 seconds. As the change in delay settings happened quickly after losing connection, it is assumed students were not affected by this in the same extent as when this happened during the condition of *high delay*. The delay set was lower than the originally planned setting, as can be seen in Table 3.2 on page 20, but can still be considered a *moderate delay* after comparisons with the maximum latency from Kahoot! as displayed in Figure 3.2 on page 17.



```
anlaug@anlaug-Satellite-U400:~$ sudo tc qdisc add dev eth0 root netem delay 5000ms 1000ms
anlaug@anlaug-Satellite-U400:~$ ping -c 3 getkahoot.com
PING getkahoot.com (146.185.169.87) 56(84) bytes of data.
64 bytes from getkahoot.com (146.185.169.87): icmp_seq=1 ttl=49 time=5270 ms
64 bytes from 146.185.169.87: icmp_seq=3 ttl=49 time=4063 ms

--- getkahoot.com ping statistics ---
3 packets transmitted, 2 received, 33% packet loss, time 2017ms
rtt min/avg/max/mdev = 4063.766/4666.949/5270.133/603.187 ms, pipe 3
anlaug@anlaug-Satellite-U400:~$ sudo tc qdisc del dev eth0 root netem
anlaug@anlaug-Satellite-U400:~$ ping -c 3 getkahoot.com
PING getkahoot.com (146.185.169.87) 56(84) bytes of data.
64 bytes from getkahoot.com (146.185.169.87): icmp_seq=1 ttl=49 time=37.5 ms
64 bytes from getkahoot.com (146.185.169.87): icmp_seq=2 ttl=49 time=37.6 ms
64 bytes from getkahoot.com (146.185.169.87): icmp_seq=3 ttl=49 time=37.7 ms

--- getkahoot.com ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2002ms
rtt min/avg/max/mdev = 37.580/37.637/37.704/0.166 ms
anlaug@anlaug-Satellite-U400:~$ sudo tc qdisc add dev eth0 root netem delay 3000ms
anlaug@anlaug-Satellite-U400:~$ ping -c 3 getkahoot.com
PING getkahoot.com (146.185.169.87) 56(84) bytes of data.
64 bytes from getkahoot.com (146.185.169.87): icmp_seq=1 ttl=49 time=3038 ms
64 bytes from getkahoot.com (146.185.169.87): icmp_seq=2 ttl=49 time=3037 ms
64 bytes from 146.185.169.87: icmp_seq=3 ttl=49 time=3037 ms

--- getkahoot.com ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2017ms
rtt min/avg/max/mdev = 3037.753/3038.176/3038.948/1.524 ms, pipe 3
anlaug@anlaug-Satellite-U400:~$
```

Figure 4.4: Ping from the condition of *moderate delay*. (1) Setting the original delay of 5000ms ± 1000ms. (2) Deletion of the original delay. (3) Setting new delay of 3000ms.
* Shows where ping to getkahoot.com has been executed.

Figure 4.4 shows a screenshot of the terminal on the computer running the Kahoot! during the test. What can be found in this screenshot is: (1) the original delay set (5000ms ± 1000ms) followed by a ping confirming the delay. After losing the connection to the Kahoot! server, (2) the delay was deleted. Finally, (3) a delay of 3000ms was set and a final ping, confirming the new delay.

From the result of the questionnaire it was found that only 13% answered all quiz questions while 77% reported they left the quiz due to technical difficulties or were thrown out of the quiz. Similar to the test with *high delay*, losing connection to the server had a severe impact on the users. 81% reported that they agreed or strongly agreed to have experienced delay between the monitor and their device, while 68% agreed or strongly agreed they would have obtained a higher score if the technical quality of the session had been better. 4% disagreed or strongly disagreed that the technical quality had an impact. Technical difficulties are expected to affect the users' impression of the application, yet surprisingly, more than 70% still characterize Kahoot! as a useful learning tool and a nice diversion during lecture while 10% consider Kahoot! to be too time consuming.

In the last question of the questionnaire, students were able to comment on their experience. Multiple students commented on a high delay between their device and the monitor and that they lost connection to the server, forcing them to refresh as well as some anger in regards of the delay and correct answers not being registered.

4.2.4 Statistical Analysis

As the data set from the real-life Longitudinal testing consists of self-reported measures, it can be assumed some answers are disreputable. In a data cleaning phase, frequencies were analyzed using SPSS, resulting in removal of some participants' answers. As an example, one participant strongly disagreed with all the statements throughout the questionnaire. These inconsistent responses were removed from the data pool to ensure quality of the data.

As previously mentioned, Cronbach's alpha was used to investigate whether there existed internal consistency between the items listed in Table 3.5 on page 30, Table 4.1 list the alpha values found for all four tests. As mentioned, an alpha value above 0.6 confirms that items measure the same construct and can be computed into a new subjective measures of QoE. Accordingly five new variables were created; Degree of Delight; Degree of Annoyance; Evaluation of Quality; Positive Influence and Negative Influence.

	Test 1 no delay	**Test 2** high delay	**Test 3** moderate delay	**Test 4** no delay
Degree of Delight	0.874	0.906	0.918	0.933
Degree of Annoyance	0.700	0.785	0.747	0.806
Evaluation of Quality	0.684	0.795	0.485*	0.887
Positive Influence	0.642	0.388*	0.797	0.791
Negative Influence	0.727	0.834	0.833	0.862

Table 4.1: Results from Cronbach's alpha test in real-life Longitudinal testing.
* indicating $\alpha < 0.6$ and new variables cannot be calculated.

These new dependent variables were then used to test for significant differences between the conditions of delay. As the test was conducted in a class where participation was voluntarily and as mentioned, the number of students participating in class was fast decreasing, the data set had to be split up to be able to run analyses for both the related and unrelated test samples.

The related data set consists of data from students who participated during multiple conditions while the unrelated set consists of answers from students who participated in only one of the conditions being compared. The Wilcoxon signed-rank test and Mann-Whitney U test were run on the corresponding sets to locate significant differences. These test results can be found in Appendix A.

As can be seen in Table 4.1 the alpha value from the Cronbach's alpha test was not high enough for the items in the variable Evaluation of Quality of *moderate delay* (Test 3), and the items in the variable of Positive Emotions of *high delay* (Test 2). Due to this, these variables for the mentioned sections could not be computed and could therefore not be used for subsequent analyses.

4.2.5 Subjective Measures of QoE

This section presents boxplots illustrating the ratings of the subjective measures of QoE for each test. The plots consist of all test participants from each testing and the numbers of contestants are therefore not equal. Test 1 consisted of 98 students, test 2 of 108, test 3 of 53 and finally 34 students participated in test 4. Moreover, not all participants attended all tests. A total of 25 test participants did experience a normal condition before experience at least one of the delays and the final test of *no delay*. For this reason the results presented are more descriptive illustrations, leading to a number of assumptions that need to verified in future research. These results can therefore not be used to draw general conclusions. Anyhow the results presented give an indication of what the result might look like if test had been conducted on a class with mandatory attendance and will be supported by results from the Wilcoxon signed-rank test (related measures) and Mann-Whitney U test (unrelated measures) as can be found in Appendix A.

Distribution of Degree of Delight

In Figure 4.5, a boxplot illustrating how students rated Degree of Delight is displayed. The plot shows that the results are highly spread within each condition when expressing Degree of Delight, as illustrated by the widespread inner fences.

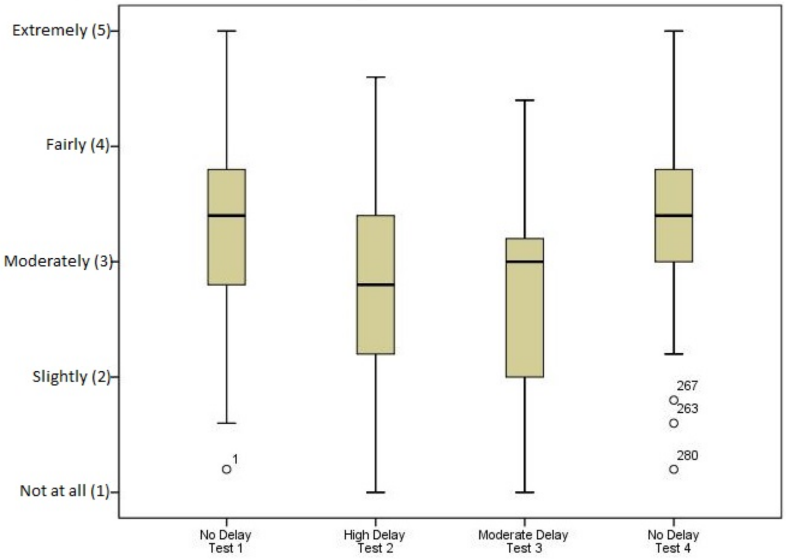


Figure 4.5: Distribution of Degree of Delight in real-life Longitudinal tests. Extremely indicates a high degree of delight.

The results from the Wilcoxon signed-rank and Mann-Whitney tests describe the related and unrelated measures and looks for significant differences between the self-reported measures when considering the delay conditions.

Between the two conditions of *no delay* no significant differences were found, both when considering the related and the unrelated sample. As mentioned, when a result is not significant, it does not mean that there is definitely no effect or no difference between the investigated conditions, it means that the effect found in the investigated data is not substantial enough to argue that it may not be based on chance. When comparing the two conditions of *no delay* as seen in the boxplot, it can be argued that there are some differences between the two conditions. These differences could be a result of the previously experienced conditions with delay or the fact that the sample size of test four is much smaller than that of test 1. However, the medians for these two conditions are equal, 3.40, which can be seen as an indication that Degree of Delight is not highly affected by previous experience of delay.

When looking at the results for Degree of Delight presented in Table A.1 in Appendix A, it can be seen that significant differences were identified and this was the case for both the related and unrelated measures between the two conditions of delay. This could indicate that delay does affect delight in some manner. As there are not found a significant difference for the related measures between the condition of *high delay* and the final test of *no delay*, it could maybe be argued that this significant difference may have been a coincidence, and does not give a clear indication that delight is affected by delay for this setting. This might be a result of the relatively small sample size and for this reason conclusions cannot be drawn.

For all test it can be assumed that the similar evaluations of delight, might be due to delight being affected by external factors of the real-life setting including co-located participants.

Distribution of Degree of Annoyance

Figure 4.6 describes a plot of how students rated Degree of Annoyance throughout the four different tests conducted.

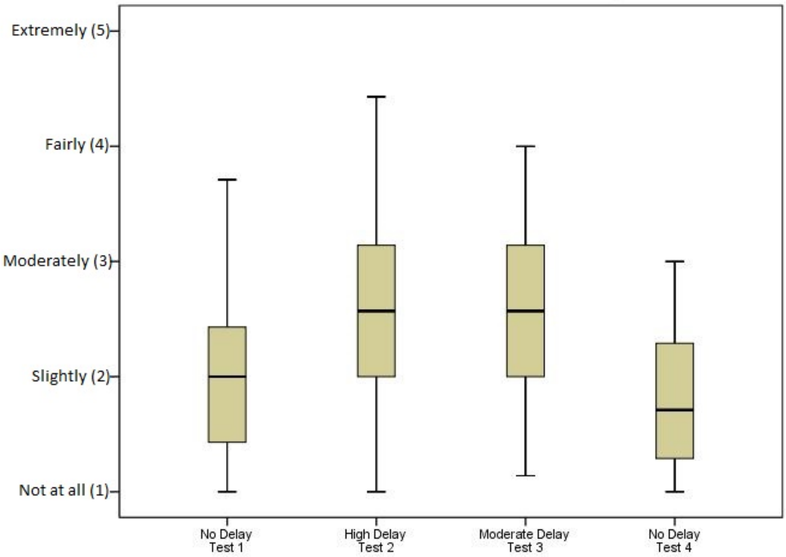


Figure 4.6: Distribution of Degree of Annoyance in real-life Longitudinal tests. Extremely indicates a high degree of annoyance.

It can be found that no one used the highest grade of the scale when describing Degree of Annoyance, but the lower part of the scale is represented in all conditions. Regardless of condition the annoyance of some participants was thus not increased because of the delay.

In the two conditions of *no delay*, the results from the Wilcoxon and Mann-Whitney tests presented in Appendix A, indicated no significant differences between the results from the two tests. Anyhow, it is interesting to point out that the results from the last test have a smaller range and a lower median value than in test one. What can be seen is that the range of the results is decreasing and it can be said that users probably were affected by their experience of the higher delay settings, even though tests were done over a longer period of time, and even though the number of participants had decreased.

The results presented in Appendix A Table A.2, point to significant difference in rated annoyance between the setting of *no delay* and the settings where delay was emulated. No significant differences in rated annoyance could be found between the two settings with emulated delay, or between the two settings without delay. Together this gives leverage to state that annoyance in some manner is affected by the emulated delay, and more specifically, that annoyance increases when a delay is experienced.

Distribution of Evaluation of Quality

Figure 4.7 displays a plot showing how students evaluated the technical quality of Kahoot!. This boxplot only includes three plots as the Cronbach's alpha value of Evaluation of Quality in *moderate delay* was too low to be calculated into a new dependent variable.

As expected, technical quality was rated higher for the conditions of *no delay* than the condition were delay was emulated. Again there is a trend showing that experience matters as the final condition of *no delay* shows an experience of higher technical quality than in the first test. This is supported by the results presented in Appendix A, Table A.3 showing a significant difference between the three conditions of delay for the related measures, and which corresponds with the results presented in the figure below. A possible explanation for the significant difference in the rating of perceived technical quality between the two conditions of *no delay*, is the test subjects' previous experience with delay and its influence on their reference and expectations, increasing the evaluation of experienced quality for the last test.

It can be seen in Figure 4.7 that there are some outliers in the results from the first test. These outliers could be users with a high demand for quality and assumption of how they believe the application should work at its best.

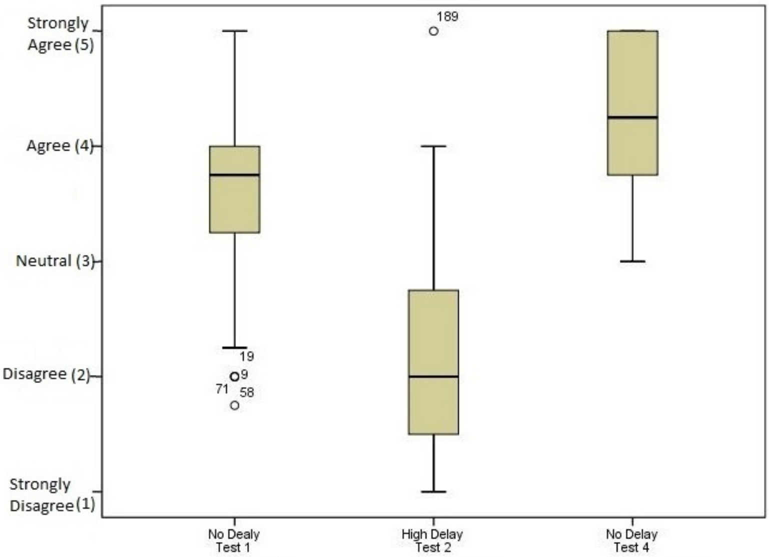


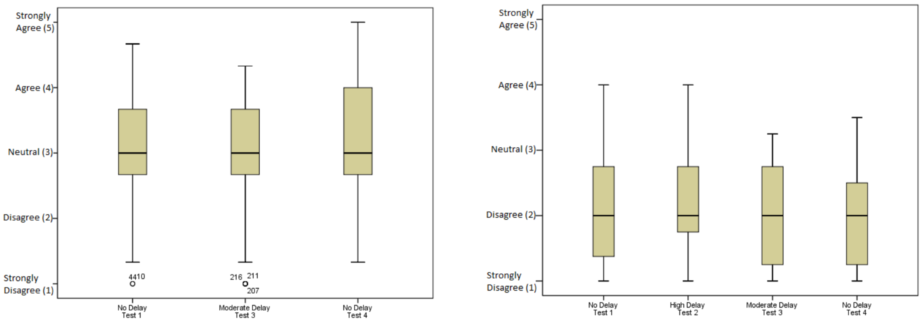
Figure 4.7: Distribution of Evaluation of Quality in real-life Longitudinal tests. Strongly agree indicates that quality was evaluated as good.

When considering the test with *high delay*, there is one outlying value where Evaluation of Quality was rated as good. By looking through the result it was found that this person showed up late to class, probably after the second question where almost everyone lost connection to the server. As mentioned, delay was turned off due to this incident before turned back on again ahead of the last question. For this reason this student might not have noticed the delay in a similarly strong manner as his classmates and rated the quality as good.

Distribution of Positive and Negative Influences

Figure 4.8 displays how students rated positive and negative emotions influenced by co-located participants. As the Cronbach's alpha value of Positive Influences in *high delay* was too low to be calculated into a new variable, the boxplot in Figure 4.8a only includes three plots.

As can be seen in Figure 4.8a, the medians for all conditions rating Positive Influences are equal (3.0 - neutral), indicating consistency in the test participants' responses. Also the medians of Negative Influences, presented in Figure 4.8b are equal (2.0 - disagree). By looking at the results from Wilcoxon signed-rank and Mann-Whitney U tests, as can be found in Appendix A, only few significant differences were found. Within Positive Influences, significant differences have been found for the related measures between the first and the third test and for the unrelated measures between the third and fourth test. For Negative Influences, the only significant difference that could be observed, is situated between the first and second test.



(a) Distribution of Positive Influence in real-life Longitudinal tests.

(b) Distribution of Negative Influence in real-life Longitudinal tests.

Figure 4.8: Distribution of Positive and Negative Influence of co-located participants. Strongly agree indicates that student were strongly positively/negatively influenced by co-located participants

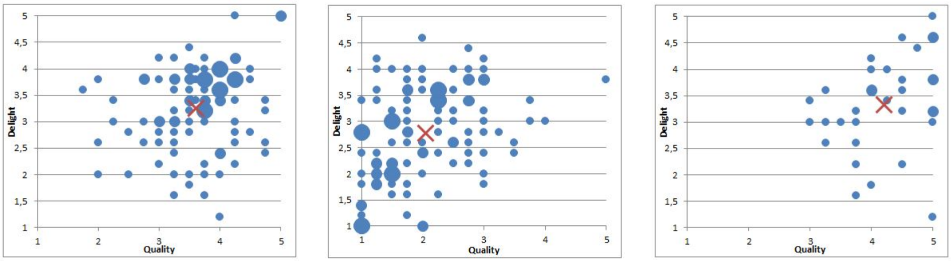
It is worth mentioning that the upper and lower fences for the Positive Influences use the entire scale, meaning that the students were conflicted regarding this topic, as some participants were not at all influenced in a positive way and others felt highly positively influenced by others. The distribution of Negative Influences also make use of a large part of the scale, but the main distribution is located in the lower part of the scale, indicating that test participants were in general not negatively influenced by their co-located participants.

Together with the few significant differences and the range of answers which are highly spread across the scale, this indicates that the influence of co-located participants did not change considerably for the different conditions of delay. Thus, it can be stated that the emulated delay or the previous experience did not affect positive or negative experiences influenced by others in a high manner. For this reason positive and negative influences will not be further investigated for this test setting.

4.2.6 How the Self-Reported Measures Correlate

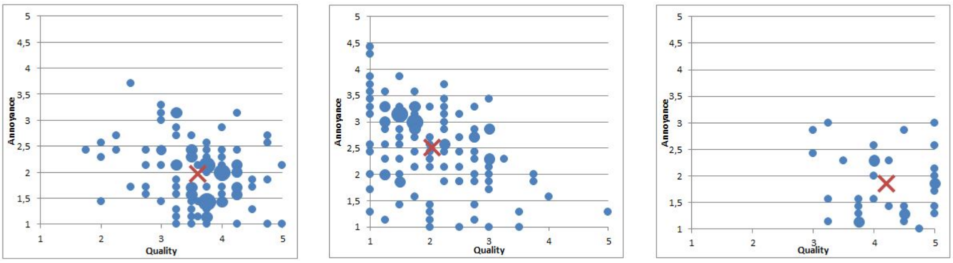
As mentioned Spearman's rank-order test has been used to look for correlations and explore the alternative measures used to evaluate QoE in this thesis work. The results from the tests done can be found in Appendix C, Section C.1.

In Figure 4.9 the correlation between perceived technical quality and delight, and in Figure 4.10 perceived technical quality and annoyance has been plotted for the different conditions of delay. As the quality variable for the third test did not produce a high enough alpha value in the Cronbach's alpha test, these rest results are not presented.



(a) Test1, no delay N=96 (b) Test2, high delay, N=106 (c) Test4, no delay, N=33

Figure 4.9: Correlation between perceived technical quality and delight. 5 indicates high experienced quality or high feelings of delight. Cross marking average delight and quality.



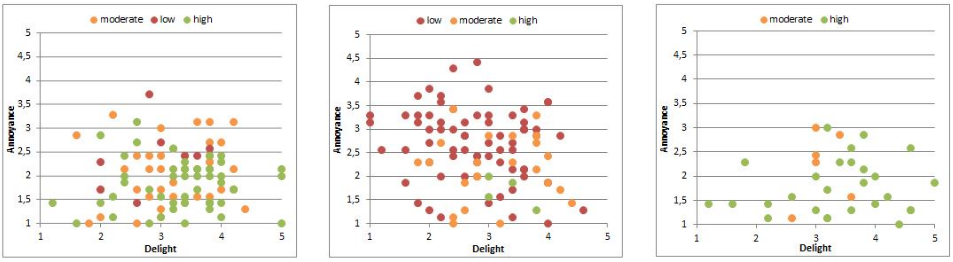
(a) Test1, no delay, N=96 (b) Test2, high delay, N=106 (c) Test4, no delay, N=33

Figure 4.10: Correlation between quality and annoyance. 5 indicates high experienced quality or high feelings of annoyance. Cross marking average annoyance and quality.

There was found a significant, yet low correlation between the self-reported quality and delight for the second (see Figure 4.9a) and fourth test (Figure 4.9c) (Spearman's $\rho = 0.374$ and 0.405 respectively). The correlation analyses also yielded significant negative correlations between the self-reported measurements of quality and annoyance for the first (fig.4.10a) and second test (fig. 4.10b) (Spearman's ρ = -0.223 and -0.381 respectively), yet again the correlation is considered low. From this it can be stated that both annoyance and delight are affected by perceived technical quality in some manner but as the correlation between quality and delight or annoyance is low, it can be said that other influencing factors plays a role as well. Though, as found in results already presented, it can be seen for both delight and annoyance that experience has changed the evaluation of the delay setting. Even though the number of students has decreased, it can be seen in the figures above, that the overall feeling of delight has increased from the first to the last test of *no delay*, and that overall annoyance has decreased.

When looking at the results presented in Section C.1, it can be observed a significant correlation ($p<0.01$) between delight and positive influences for all tests (Spearman's ρ = 0.279, 0.598 and 0.760 respectively). This implying that these variables are influencing each other in some way, as delight is increasing, so is the positive influences of co-located partcipants.

Figure 4.11 presents the correlation between feelings of delight and annoyance and is color coded in respect to how students rated perceived technical quality. From the results presented in Appendix C, Section C.1, it can be seen that no significant correlations were found between delight and annoyance for any of the tests. Anyhow, the plots presented in Figure 4.11 confirms that annoyance seems to be affected by delay in a stronger manner than delight, and that the previous experiences seem to have an influence on the students as there is a decrease in annoyance as well as an increase in perceived technical quality from test 1 presented in 4.11a to test 4 presented in 4.11c, both where *no delay* was emulated.



(a) test1, no delay, N=96

(b) test2, high delay, N=106

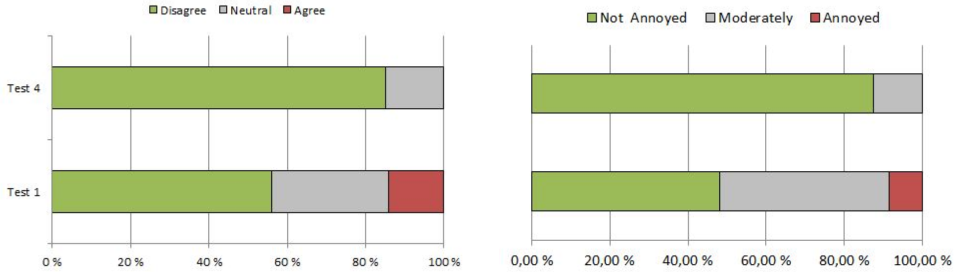
(c) test4, no delay, N=33

Figure 4.11: Correlation between delight and annoyance. 5 indicates high feelings of delight or annoyance, color coded by perceived technical quality.

4.2.7 Longitudinal Effect

The goal of introducing a longitudinal study was to present different delays to a group of students over a period of time, before presenting them with a delay they already had experienced. This was to investigate whether students change their mind over time and whether experience has influenced their evaluations of QoE. In particular, it is interesting to emphasize the two tests conducted without delay emulated, especially to see whether the last test without delay were affected by the previous tests where delay was emulated.

Figure 4.12a presents a bar chart with the two conditions of *no delay*, where students answered if they believed their score would have been higher if the technical quality of the Kahoot! session had been better. As can be seen in the bar chart, the results from the first test include the entire scale. (The Likert scale has been simplified merging the end values of the scale creating a cleaner chart.) Even though the majority of the students disagreed, meaning these students did not believe quality of the session affected their score, a total of 14% claimed the technical quality impacted their score in a negative manner. By comparing these results to the result from test four it can be seen that no one believed their score was affected by technical difficulties. Even though the same delay condition was presented in both tests, there were changes in responses and there are reasons to believe that this change is due to the new experience of the users, giving them a better understanding of different conditions of quality. The charts in Figure 4.12 includes 124 test participants for test one, and only 25 participants of test four. As the gap in numbers of participants is high, these results can only be seen as illustrative and could differ if the test group of test four had been greater, as it would have if participation had been mandatory. Anyhow, it is interesting to explore the occurrence of similar trends.



(a) Participants describing if delay affected their Kahoot! score in the conditions of *no delay*.

(b) How feelings of annoyance is rated for the conditions of *no delay*

Figure 4.12: Comparing test 1 and 4, in both tests *no delay* was emulated. The Likert scale has been simplified merging the end values. The test participants for test four in these plots have previously experienced both a condition of *no delay* and a condition of delay, resulting in 25 test participants. In test number one 124 students participated.

Figure 4.12b confirms the claim presented above, that there is a difference in how students evaluate the setting of *no delay* in the first and the last test. The percentage of students feeling annoyed has decreased from 8.5% to none. By looking at the trend presented from the results above, it can be said that for a longitudinal study done in weekly intervals, where the same delay is presented 3 weeks apart, the experience of students has changed in some degree as an effect of experience.

4.3 Cross-Sectional Testing

The Cross-Sectional test was conducted one week after the real-life Longitudinal testing, which most of the participants had previously attended. As mentioned earlier, the Cross-Sectional test was conducted on 21 students and consisted of three Kahoot!s where students were given different delay settings. Delay was emulated on the answering devices to ensure an unfair setting. The test played out as intended creating an obvious unfair setting for the users.

This section looks into how emotions of delight and annoyance, perceived technical quality and positive influences of co-located participants were affected by the different delay conditions.

To be able to better understand and analyze the results, they have been sorted with respect to the order of the conditions of delay experienced by the users, before looking into fairness and how the setting and different conditions of delay affected QoE.

4.3.1 No Delay

When experiencing *no delay*, all students reported they were able to answer all Kahoot! questions. The results indicate an average answering time of 5.3 seconds. This result is as expected as the questions chosen were of a low difficulty level and participants should be able to answer in a short amount of time. The questionnaire shows that frustration was low and the overall satisfaction was high, creating an environment where the students were concentrated and in a competitive mode.

No one in the test panel reported to have experienced any delay between the monitor and their device. As each of the three tests were conducted with students competing on three different delays, it was expected that the participants completing the Kahoot! without delay would perform better than the students with delays emulated on their device. 48% of the participants believed that they performed better than their co-located participants and only 5% believed they performed poorer in the condition of *no delay.* The average score of *no delay* was about 4700 point out of 7000 possible.

It has been found that all students were able to answer correctly to the second question of the quiz when no delay was added to their device. On this question an average answering time of 2.6 seconds was registered giving an average score of 965 point out of 1000 possible. This shows what is ideal and how the application works at its best.

4.3.2 Moderate Delay

With a delay of 5000ms ± 1000ms, all users were able to answer all questions from the Kahoot!. 23% reported that they performed poorer than their neighbors while 50% reported that they did not notice how others performed compared to themselves. The average answering time registered was 10.5 seconds and an average quiz score of 3760 points. 45% agreed or strongly agreed they would have received a higher score if the technical quality of the session had been better.

By looking at question two, where again all students answered correctly when subjected to *moderate delay*, an average answering time of 9.6 seconds per question was registered giving an average score of 850 points. By comparing this with the result from *no delay* it can be found that the students loses about 100 points per question when there is a delay of 5000ms ± 1000ms in the network. This indicates that delay affects the participants' performance, as was also found in the Longitudinal tests.

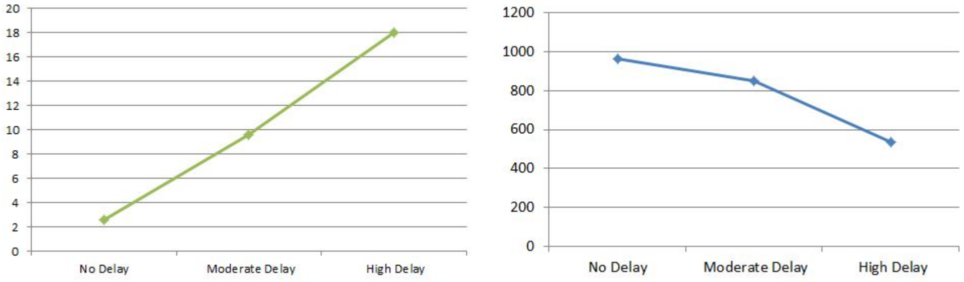
54% reported a moderate or fairly degree of frustration but no one felt an extreme degree of frustration. 68% stated that they did experience delay between the monitor and their device. Still concentration and competitiveness were rated high: 50% described their level of concentration as high, 59% indicated that their level of competitiveness was high.

In the comment section of the questionnaire, students commented on a delay of about 5 seconds and how it created frustration. One student stated: "I experienced a delay of about five seconds, and had to wait in frustration while everybody else was able to answer, losing my opportunity to achieve top score". This is an accurate observation, as it is in line with the actual delay emulated.

4.3.3 High Delay

When competing with an emulated delay of 9000ms ± 1000ms, the participants reported high frustration (67% fairly or extremely) and low satisfaction (62% not at all or slightly satisfied). All participants reported that they experienced delay between the monitor and their device and only 33% managed to answer all questions while the remaining 67% reported that they ran out of time, were not able to answer due to the display of the message saying "slow network connection" or left the quiz due to technical difficulties. All students experiencing the condition of *high delay* were able to answer some questions except from one student who was obstructed by the message "slow network connection" throughout the quiz and never had the ability to answer. The other students who experienced the message had a small window of about three seconds where they were able to answer before the screen went grey, displaying the message. Multiple students reported on this window creating stress and frustration as they had very little time to answer the question.

From the Kahoot! result it was found that the average answering time was 17.7 seconds per question and an average total score of 3036 points. When looking at question two it has been found that 5 out of 21 students were not able to answer. The registered average answering time was 18 seconds with average score on this question of 536 points. By comparing this result with the corresponding results from each delay, as has been visualized in Figure 4.13, it can be stated that the delay is affecting the answering time and points collected in a high manner. A total of 90% agreed or strongly agreed they would have received a higher score if the technical quality of the Kahoot! session would have been better.



(a) Average answering time of q2 from the Kahoot! for each condition of delay.

(b) Average score obtained for q2 of the Kahoot! for each condition of delay.

Figure 4.13: Visualization of the relation between delay, average answering time and average score obtained for the Kahoot! quiz. Question two from each quiz has been used for the visualization as all students able to answer this question, answered correctly.

Throughout the quiz, observations were done of students wondering what to do as they were not able to answer, followed by lost interest in the questions. One student was observed wiggling the network cable and its connection to his computer. From the look of it he believed this was the week link of the connection and he was trying to hold it in the right angle to keep the quality of the connection from dropping.

In the next section, results presented from all conditions of delay will be analyzed using the statistical tool SPSS.

4.3.4 Statistical Analysis

In the Cross-Sectional tests conducted, all participants participated in all conditions creating a related data set. SPSS has been used to look for internal consistency and significant differences to find how delay affects QoE through feelings of delight and annoyance, evaluation of technical quality as well as how the influence of co-located participants affects positive and negative emotions, using Cronbach's alpha, Friedman ANOVA and Wilcoxon signed-rank tests. A description of the methods used can be found in Table 3.6 on page 33.

The Cronbach's alpha test was run on items listed in Table 3.5 on page 30 looking for internal consistency and to figure out if the items could be computed into new variables of subjective measures of QoE. Table 4.2 presents the results from these tests, where an α-value above 0.6 confirms internal consistency. As all resulting Cronbach's alpha values are above 0.6, new subjective measures of QoE were computed for all cases tested.

Subjective Measures of QoE	**No Delay**	**Moderate Delay**	**High Delay**
Degree of Delight	0.743	0.910	0.919
Degree of Annoyance	0.782	0.813	0.821
Evaluation of Quality	0.863	0.834	0.864
Positive Influence	0.783	0.806	0.824
Negative Influence	0.824	0.689	0.828

Table 4.2: Results from Cronbach's alpha for Cross-Sectional test in lab environment. For $\alpha > 0.6$ internal consistency is considered high, and new variables can be calculated.

Friedman ANOVA was then run to compare the subjective measures of QoE with corresponding variables from the different delay settings. From the results it was found that there are no significant differences between the different delays for the variable of Negative Emotions (p-value=0.07) while for all other variables, significant differences were found. A table containing the output from Friedman ANOVA can be found in Appendix B. From this it can be stated that for the Cross-Sectional testing, the negative influence of co-located participants is not affected in a high manner by the introduction of delay, and Negative Influences will not be further investigated.

The Wilcoxon signed-rank test was run on the four variables where the Friedman ANOVA test located significant differences, and the results can be found in Table 4.3. From the Wilcoxon signed-rank test it was found significant differences for most cases related to how students rated delight, annoyance, technical quality and positive influences by co-located participants.

Table 4.3a shows that significant differences were found for most cases when comparing the variables of each condition of delay, the only exception can be found when comparing the variable of Positive Influences between the conditions of *no delay* and *moderate delay*, which has been marked in the table as not significant. This can be confirmed by looking at the medians presented in Table 4.3b where the medians of Positive Influence in *no delay* and *moderate delay* are equal.

		Degree of Delight	**Degree of Annoyance**	**Evaluation of Quality**	**Positive Influence**
No Delay vs. Moderate Delay	T	4	4	0	6
	P	0.034	0.009	0.000	ns
No Delay vs. High Delay	T	3	3	0	2
	P	0.002	0.001	0.000	0.004
High Delay vs. Moderate Delay	T	3	3	1	2
	P	0.014	0.006	0.000	0.002

(a) Results from Wilcoxon signed-rank test in the Cross-Sectional test. P < 0.05 indicates significant difference and T-values refers to the test statistics. P > 0.05 is not significant and marked by ns.

	Degree of Delight	**Degree of Annoyance**	**Evaluation of Quality**	**Positive Influence**
No Delay	3.2000	1.5714	4.2500	3.3333
Moderate Delay	3.0000	2.1429	2.7500	3.3333
High Delay	2.4000	2.5714	1.5000	3.0000

(b) Median values of each condition. Medians are a measure of the central tendency of the data set and give the 50th percentile of the distribution.

Table 4.3: Wilcoxon signed-rank test results and medians. As there were found no significant differences between the ratings of negative influences, this variable has not been included

4.3.5 Subjective Measures of QoE

The following section presents boxplots illustrating the ratings for the subjective measures of QoE created after running the Cronbach's alpha analyses. The variables presented includes experienced delight and annoyance, perceived technical quality and positive influences affected by co-located participants, as significant differences were found for these variables (see Appendix B). Results from the Wilcoxon signed-rank test (Table 4.3) together with boxplots have been used when evaluating whether the mentioned measures of QoE have been affected by delay. As the 21 test subjects participating in this part of the study took part in all three quizzes, the box plots are more representative then the plots presented for the Longitudinal part of the study.

Distribution of Degree of Delight

Results from the Wilcoxon signed-rank test listed in Table 4.3a show that students rated Degree of Delight significantly higher in the condition of *no delay* compared to the two other conditions of delay. The test also shows that ratings were significantly higher for *moderate delay* compared to *high delay.* Based on these results it can be concluded that for the Cross-Sectional study, as delay increases, Degree of Delight significantly decreases. As can be seen in Table 4.3b, the median for Degree of Delight is higher in *no delay* than in the condition of *high delay*, and the self-reported delight remains relatively high, despite the fact that there are indications that delay plays a role.

Figure 4.14 displays the three different conditions of delay in a boxplot, illustrating the distribution of feelings of delight.

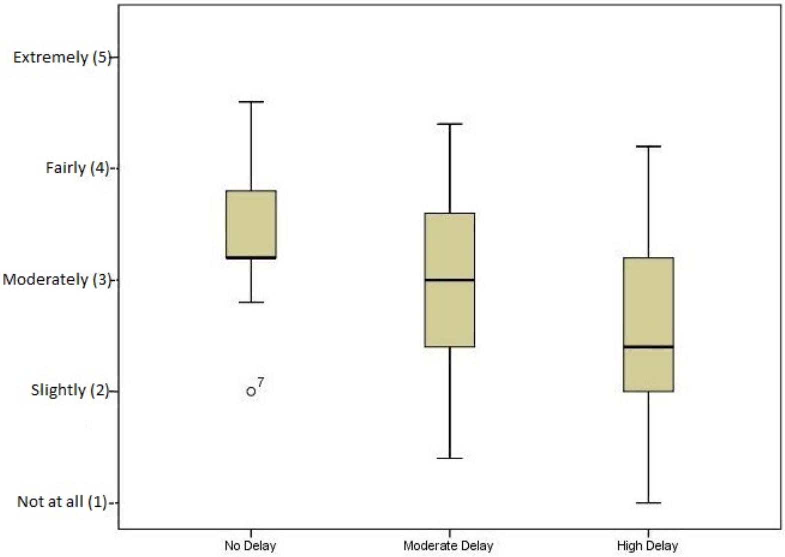


Figure 4.14: Distribution of Degree of Delight from the three conditions of the Cross-Sectional test, N=21 per condition. Extremely indicating a high degree of delight.

As can be observed in the boxplot, the interquartile range for the condition of *no delay* is smaller than for the two other conditions, thus the variation of samples were less in this condition. This means that students were more aligned in their expression of delight when no delay was emulated. This is also indicated by the median which resides on the lower hinge of the box. One participant deviates from the main distribution, indicated by the outlier below the inner fence, meaning that this participant rated Degree of Delight out of unison with the other students. Further investigation showed that this student characterizes Kahoot! as too time consuming, and it can be assumed his attitude in regards of the application goes hand in hand with a lower feeling of delight.

The boxplot illustrating Degree of Delight of *moderate delay* shows an indication of symmetric distribution as the median resides approximately in the center of the interquartile range and the inner fences are of similar lengths. Compared to *no delay* the inner fences of *moderate delay* and *high delay* extends further from the interquartile range, meaning students made use of a larger part of the scale. Unlike *moderate delay*, the median of *high delay* is gravitating towards the bottom hinge, indicating a skewness in the distribution where a larger part of the samples reside in the lower part of the scale.

As mentioned, there is a significant difference in rated delight and the conditions where delay has been emulated, but as can be seen in Figure 4.14, the distribution is increasing for the conditions with delay emulated. This supports the finding that delight is not as affected by delay, since the delight of multiple students stays relatively high through all tests. This also supports the earlier statement saying delight is affected by multiple influencing factors.

Distribution of Degree of Annoyance

As can be seen in Table 4.3a on page 57 all p-values found by the Wilcoxon signed-rank test for Degree of Annoyance were below 0.05 indicating a significant difference between all three conditions of delay. Thus the introduction of different delays significantly changed the test participants' evaluation of annoyance. Table 4.3 shows that students felt significantly more annoyed and frustrated in conditions where delay was emulated. The boxplots in Figure 4.15 clearly shows changes in the distribution of answers.

In accordance with medians presented in Table 4.3, Figure 4.15 illustrates the overall low level of annoyance in the condition of *no delay.* With some exceptions as shown by the outliers, the students answered somewhat unison, which is illustrated by the centered median and the similar lengths of the inner fences. When moving to *moderate delay*, a greater variation in the responses can be observed. As for Degree of Delight, the condition of *moderate delay* is somewhat more symmetrically distributed than *high delay.* This occurrence might be due to participants experiencing either *high* or *no delay*, or both, before they experienced *moderate delay.* It is interesting to investigate how the order of experienced delay affected the students and this will be further elaborated on later in this chapter. The boxplot clearly illustrates that as the delay increases, so do the inner fences, meaning that students used larger parts of the scale in higher delays.

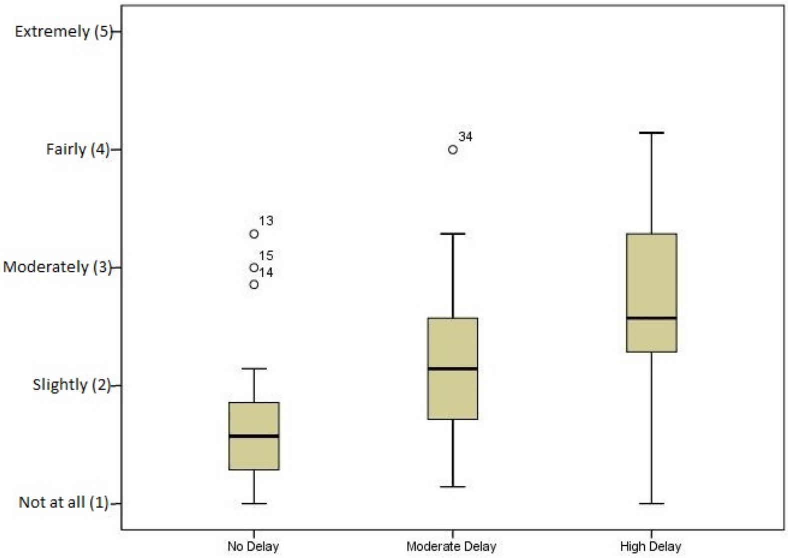


Figure 4.15: Distribution of Degree of Annoyance in the Cross-Sectional test, N=21 per condition. Extremely indicating a high degree of annoyance.

When investigating the responses associated with the outliers of *no delay*, as can be seen in Figure 4.15, it was found that these students had a high competitive nature: they indicated to feel highly competitive, as well as tense, concentrated and nervous in the questionnaire. As mentioned in Section 2.1.3, the characteristics of the user are defined as a factor influencing the QoE of a user [1]. By looking at the outlier of *moderate delay* it was found that this student (34) is the same student that represents one of the outliers (13) in the *no delay* condition.

Compared to *no delay*, the upper fence of *high delay*, here representing the maximum value of the sample, is much higher in this condition. This means that the maximum values (except for the outliers) rises drastically from *no delay* to *high delay*. Simultaneously, it can be observed that the lower part of the scale is represented in all conditions, which can be compared to the results found for the same variable in the Longitudinal study. Although the conditions of delay change, there are still participants who do not perceive feelings of annoyance, while annoyance in general was increasing for the higher levels of delay. This might be explained by the fact that users were given delays in different orders to avoid the order effect, creating a setting where users might be affected by the given order of delays. Some participants who rated annoyance low, meaning they did not feel annoyed in the condition of *high delay*, could have experienced this condition first.

Distribution of Evaluation of Quality

As can be found for the Longitudinal study, Evaluation of Quality is rated significantly different for all conditions of delay as described in Table 4.3 on page 57. The medians for this category are shown in Table 4.3b and presents a clear trend that delay has a great impact on the perceived technical quality. In *no delay*, the median was 4.25, meaning that the students to a great extent agreed with statements saying that the overall quality of Kahoot! was good, functioned optimally and was as it should be. For the condition of *high delay*, the median was 1.50, meaning that the students disagree with statements claiming the quality was good. The apparent difference in medians among the conditions of delay illustrates the effect delay has on perceived quality. It seems that delay has an impact on all self-report measures of QoE, but more so for the quality aspect than for the feelings of delight and annoyance.

As illustrated in Figure 4.16, the changes in Evaluation of Quality between the conditions are clearly visible. It can be assumed that delay has a greater impact on measures of quality as it does not involve highly subjective emotions from playing, but rather the more objective aspects of the performance of Kahoot!.

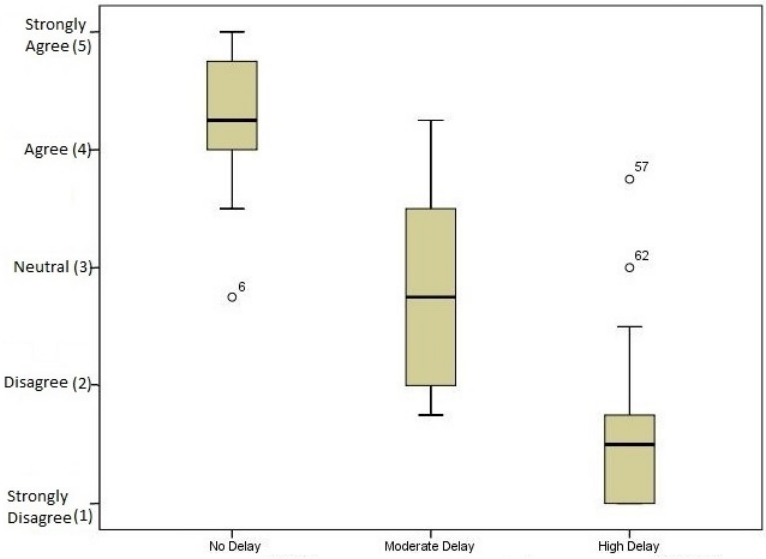


Figure 4.16: Distribution of Evaluation of Quality in the Cross-Sectional test, N=21 per condition. Strongly agree indicates that quality was evaluated as good.

Figure 4.16 shows a high rating of quality in the condition of *no delay*. The distribution in this condition is relatively small indicating that the perception of quality to a certain degree was consistent among the students. The same trend can be observed for *high delay*, but in the opposite end of the scale. *Moderate delay* on the other hand varies more within the distribution and this observation emphasizes the effect previous experience had as *moderate delay* was experienced after either *no* or a *high delay*.

The results presented in Figure 4.16 can be emphasized by the results presented in Figure 4.17, which displays the percentage of students who did or did not notice delay between the monitor and their device during the different quizzes. This figure clearly displays how no student perceived delay when no delay was emulated on their device while all students experienced delay when in the condition of *high delay*.

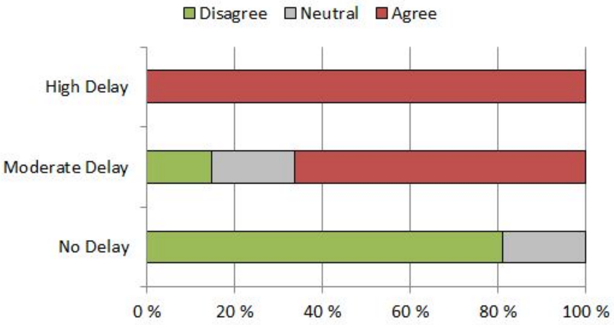


Figure 4.17: Experience of delay between monitor and device.

Distribution of Positive Influences

Figure 4.18 presents how students rated whether the influences created by co-located participants affected them in a positive way. The idea of asking these questions was to explore whether students were influenced by other students, seated in close proximity, experiencing a higher or lower quality level than themselves.

What can be seen from the plot in Figure 4.18, is that the medians for *no delay* and *moderate delay* are 3.33 and equal, the medians are presented in Table 4.3 on page 57. For both conditions, the plot shows a couple of outliers. By studying the self-reported measures, it was found that the two outliers from the condition of *no delay* are results from the same participants as the outliers in *moderate delay*. The results show that neither of the students representing these outliers did pay attention to how other students performed, which likely is the reason why the influence of others did not affect them in a positive or any manner.

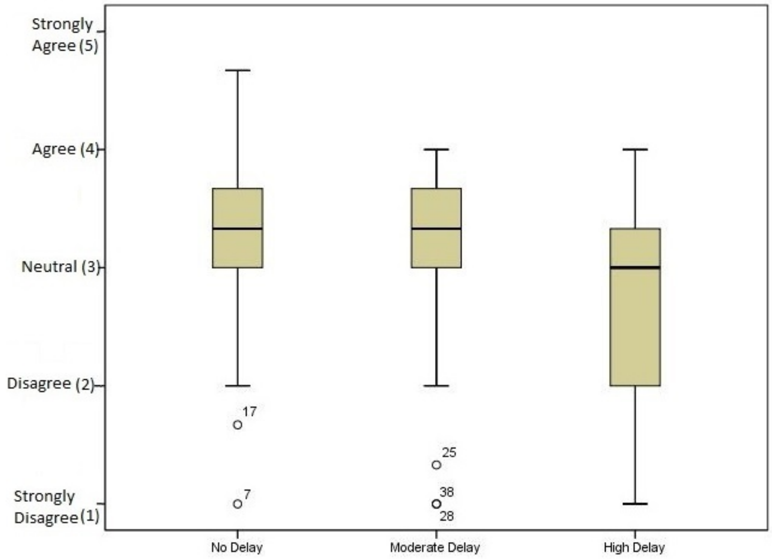


Figure 4.18: Distribution of self-reported Positive Influences in all conditions of the Cross-Sectional test, N=21 per condition. Strongly agree indicates that the students were strongly positively influenced by co-located participants.

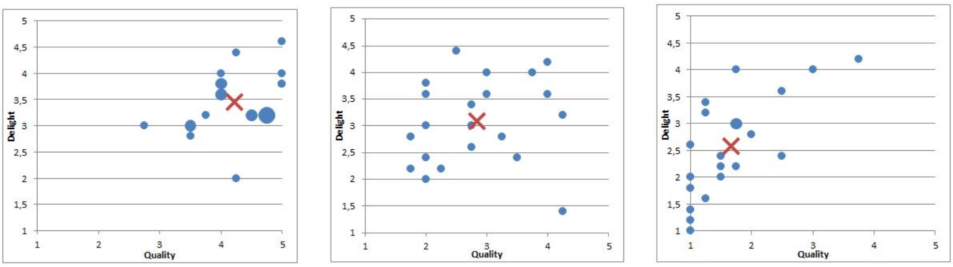
As it can be seen in Table 4.3 on page 57, there were significant differences between how positive influences were affected during *high delay* compared to the other two delay settings. This can also be observed in Figure 4.18, where the lower hinge is drawn towards the lower part of the scale. It is interesting to compare this plot to the plots presenting delight and annoyance (Figure 4.14 and 4.15): in the *high delay* condition, a larger part of the scale has been used and this is the case for all subjective measures of QoE. The reason for this might be that students disagree on how the delay influenced their experience even though as can be seen in Figure 4.16 most students agreed to experience a degraded quality setting.

4.3.6 How the Self-Reported Measures Correlate

Spearman's rank-order test was used to look for correlations and explore the alternative measures used to evaluate QoE in this thesis work. The results from the Cross-Sectional study can be found in Appendix C, Section C.2.

The plots presented in Figure 4.19 and 4.20 visualize the correlations between perceived technical quality and delight, as well as perceived technical quality and annoyance. It was found that delight and quality significantly correlate in the test of *no delay* and *high delay* (Spearman's ρ = 0.516 and 0.721 respectively).

Between quality and annoyance, a significant negative correlation was found when the participants were subjected to a *moderate delay* (Spearman's ρ = -0.575), the higher the quality, the lower is the annoyance. No significant correlations where identified between quality and annoyance for the conditions of *no delay* or *high delay*. The plots presented in the figures clearly show that there are some correlations between quality and delight or annoyance, even though not all correlations are significant.

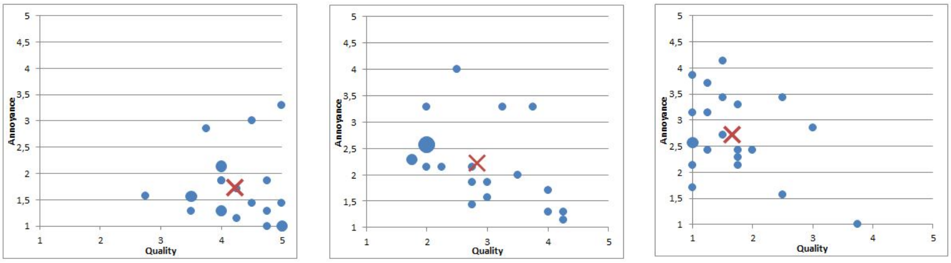


(a) test1, no delay

(b) test2, high delay

(c) test4, no delay

Figure 4.19: Correlation between quality and delight, N=21. 5 representing high experienced quality or high feelings of delight. Cross marking average delight and quality.



(a) no delay

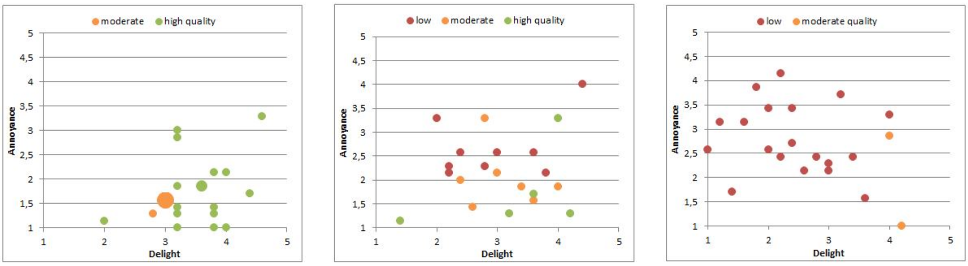
(b) moderate delay

(c) high delay

Figure 4.20: Correlation between quality and annoyance, N=21. 5 representing high experienced quality or high feelings of annoyance. Cross marking average quality and annoyance.

The correlation between "whether results were in line with effort" and "whether the student deserved better score than perceived" is found to be significant for all the three conditions of delay (Spearman's ρ = -0.610, -0.754 and -0.580 respectively). This is a negative correlation, if a student agreed to results being in line with the effort put into the quiz, she would disagree with the statement that she deserved a better score than perceived.

Figure 4.21 presents the correlation between delight and annoyance and is color coded in regards of how the students rated perceived technical quality. As found in the Longitudinal setting, the Spearman's signed-rank correlation analysis did not point to significant correlations between feelings of delight and annoyance within each test, but has been displayed to look for possible visual correlations. Yet again it can be stated that annoyance is more affected than delight by delay. The figure shows that annoyance is somewhat low when technical quality is rated high (Figure 4.21a) and high when technical quality is rated low (Figure 4.21c), while delight is more spread for all tests.



(a) no delay

(b) moderate delay

(c) high delay

Figure 4.21: Correlation between delight and annoyance, N=21. 5 representing high feelings of delight or annoyance, color coded by perceived technical quality.

4.3.7 Effect of Delay Order

The following plots represents how participants were affected by the order of delay settings they were given in the three quizzes of the Cross-Sectional test. The different delay orders test participants were presented with can be found in Table 3.3 on page 26. As an example, one person was given the order *no delay* - *moderate delay* - *high delay* while another person was given *high delay* - *moderate delay* - *no delay* and so on, this was done to be able to investigate a possible order effect, as well as trigger a potential perception of unfairness.

The order of delay conditions can possibly affect how participants grade their experience. If a participant experienced *no delay* in the second quiz, he or she will already have experienced a *moderate* or *high delay.* The following plots, presented in Figure 4.22 - 4.25 visualize the impact of order, meaning whether a delay was presented first second or last. Dotted plots are chosen as the sample size is small (N = 21), ideally the groups could be split up even further displaying how a *moderate* or a *high delay* ahead of *no delay* affects the experience of *no delay.* As the sample used for this lab experiment is small, this has not been done.

In cases where *no delay* and *high delay* are presented in the second quiz, the participant will already have experienced a respectively higher and lower delay creating a similar scenario no matter which of the two other delays that was presented first, on the other hand, when *moderate delay* are given as the second delay, it is the successor of either *no delay* or *high delay*. For this reason, when a *moderate delay* is presented as the second condition, it can be expected that the ratings will be more spread as the experience may have been affected by the previous delay presented. It is important to emphasize that these plots include small sample groups as the total test group only consisted of 21 subjects. Anyway, it is interesting to look into the distribution of the sample to obtain an overview and look for trends.

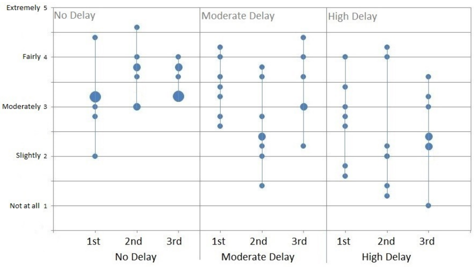


Figure 4.22: Degree of Delight separated by when each participant was introduced to the different delays. Each response is presented in a dot, the smallest dot marks one response and the largest dot marks three responses.

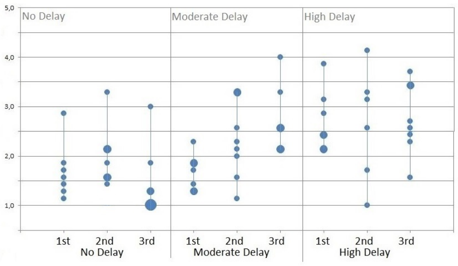


Figure 4.23: Degree of Annoyance separated by when each participant was introduced to the different delays. Each response is presented in a dot, the smallest dots marks one response and the largest dots marks two responses.

Figure 4.22 displays how the test participants experienced Degree of Delight, the higher on the y-axis the higher the experience of delight. In the condition of *no delay*, the plot shows that when users had no previous experience of delay, a larger part of the scale was utilized. When experiencing *no delay* second or last, the users had already experienced delay and it appears as this affected their degree of delight, as the users' responses became more concentrated. In the setting of *moderate delay* it can be found that when this condition was presented second, the responses were highly spread. This is probably due to the different delays presented ahead of this condition, creating expectations for the users. The plot also shows that when *high delay* was experienced, in general the result are more spread, no matter which delay was presented first. It can be assumed that there are different factors influencing Degree of Delight and that these factors are present even though the technical quality is obviously degraded.

As can be seen in Figure 4.23, in the condition of *no delay*, it seems that test participants were not affected by the order of delay in a too great manner when rating annoyance. The main distribution resides between 'not at all' and 'slightly', no matter when the condition of *no delay* was presented. In the condition of *moderate delay*, on the other hand, there is a clear difference in ratings, depending on when the *moderate delay* condition was presented. It is also interesting that when experiencing this delay first, the annoyance was ranked quite similar as for the condition of *no delay*. Perhaps as the *moderate delay* was experienced first, the test participants did not yet have an understanding on how Kahoot! should behave without delay, and therefore did not have strong feelings of annoyance. When looking at the condition of *high delay* it is surprising and interesting to find that the feelings of annoyance are highly spread no matter when the condition is presented. This can be compared with the Degree of Delight presented in Figure 4.22 and it can be assumed that different factors influence both delight and annoyance, and implying that some students were not as emotionally affected as others by the delay.

Figure 4.24 looks into evaluation of quality and how students were affected by the order delay was presented. As expected, the quality of *no delay* was rated higher after experiencing what bad quality looks like. From *no delay* and *high delay* it can be found that the specter of results is larger when the delay was given in the first two quizzes. When a delay was given in the third quiz, it seems that users had been affected by their previous experience of delay, and the majority of the results are more concentrated. In general it can be observed for all conditions that responses are more accumulated when a delay is presented last. This confirms that students seem to take their previous experience into account when rating the experience of a technical quality according to their modified references.

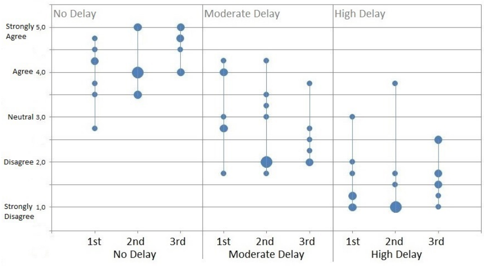


Figure 4.24: Evaluation of Quality separated by when each participant was introduced to the different delays. Each response is presented in a dot, the smallest dots marks one response and the largest dots marks three responses.

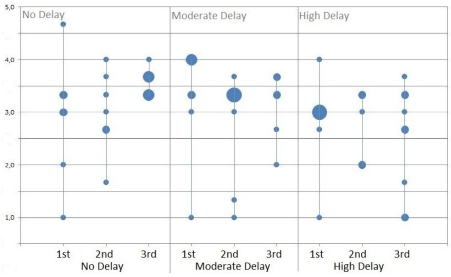


Figure 4.25: Positive Influences separated by when each participant was introduced to the different delays. Each response is presented in a dot, the smallest dots marks one response and the largest dots marks four responses.

The plot in Figure 4.25 shows the Positive Influences experienced due to co-located participants divided by the order of experienced delay. In general, it can be stated that positive influences of others are largely spread. Anyhow, there is one result presented by the figure that is interesting to highlight. When experiencing *no delay* last, the plot shows that the students' responses are all accumulated in a small part of the scale where all agrees to experience positive influences. Perhaps the experience of higher delay led to a greater awareness regarding their co-located participants, the participants were finally experiencing optimal conditions and possibly rejoice on the fact that others were experiencing lower quality.

4.3.8 Fairness

The reason for performing tests in a lab environment was to complement the tests done in the Longitudinal study, gain better control over influencing factors, be able to emulate delay on the devices used by the students and explore how students were affected by an unfair setting when presented with different delay conditions. This was used to look into how unfairness among co-located participants affected perceived QoE.

As mentioned, the test consisted of students from the same class and the majority of the test participants knew each other to some degree. For this reason and the fact that participants were placed close to each other, it was assumed interaction would occur. Surprisingly, a large part of the students were taken by the seriousness of the test and did not pay attention to co-located participants, as can be seen in Figure 4.26, marked by the grey areas of the chart. It can be said that the presence of others did not affect these students in a too great manner. By looking into the results it was found that the students answering 'I do not know' were the same for all conditions of delay.

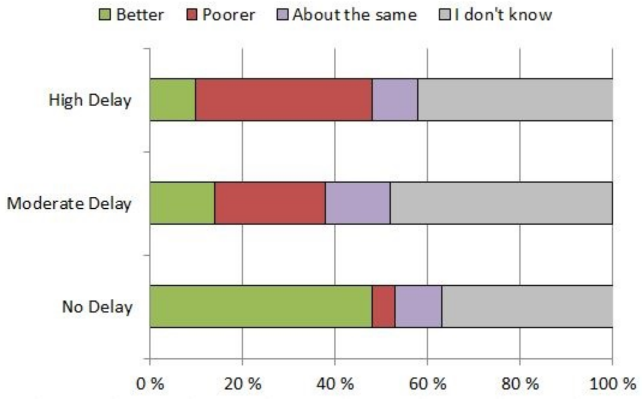


Figure 4.26: How participants believed they scored compared to their neighbours.

As expressed previously, the introduction of delay made it more difficult to gather points in the Kahoot! quiz and the unfair setting was noticed by some of the test participants as can be seen in Figure 4.26, marked by the colored areas. The figure describes how the students believed they scored compared to other co-located participants.

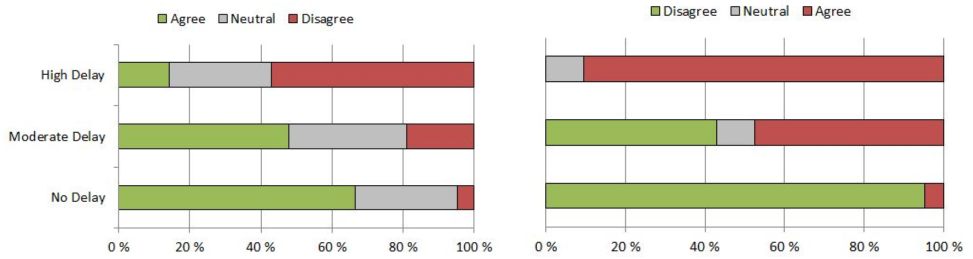
The bar chart clearly shows that the students who did pay attention to others, noticed the unfair setting as they believed they performed poorer than others when presented with the condition of *high delay* and better in the condition where no delay was added to their device. Further investigation showed that the test participants who claimed that they did better than others, had a higher Degree of Delight and a lower Degree of Annoyance than the participants believing they performed poorer than the co-located students. This is confirmed by the Mann-Whitney U test, for which the results can be found in Table 4.4. Significant differences between the students who believed that they performed better and those who believed that they performed poorer than their co-located participants, were identified. This was the case both for delight and annoyance.

Between the students believing they performed better than others and those who did not pay attention to others there was found significant differences in the Degree of Delight. There was also found significant difference in how the students who believed they performed poorer rated annoyance compared to those who did not pay attention to others. This indicates that the perceived performance and delight or annoyance are linked. It can be stated that the unfair setting and the presence of others enhances the most prominent feeling (delight or annoyance) of the user.

How did you perform compared to co-located participants?							
		Degree of Delight			**Degree of Annoyance**		
	N	Mean Rank	U	P	Mean Rank	U	P
Better vs.	15	18.53	52.0	0.020	10.33	35.0	0.002
Poorer	14	11.21			20.00		
Better vs.	15	29.60	96.0	0.004	17.43	141.5	*ns*
Do not know	28	17.93			24.45		
Poorer vs.	14	18.64	156.0	*ns*	29.68	81.5	0.002
Do not know	28	22.93			17.41		

Table 4.4: Mann-Whitney U test results looking for significant differences in delight and annoyance compared by how students believed they performed compared to co-located test participants. A result is significant if p < 0.05, *ns* = not significant.

Figure 4.27a depicts if students felt the score they obtained in the Kahoot! quiz reflected the effort they put into the quiz. As can be seen in the figure more students agreed with this in the condition of *no delay* than in the condition of *high delay*. In Figure 4.27b the students described whether the technical quality of Kahoot! affected their quiz score in a negative manner. Clearly, a majority of the students felt their quiz scores were affected in the condition of *high delay*.



(a) Students responding if the quiz results were in line with their efforts. "Agree" (green) indicates that results were in line with the effort the students put into the quiz.

(b) Participants describing if the technical quality of the Kahoot! session affected their quiz score in an negative manner.

Figure 4.27: Quiz results in relation to effort and technical quality of Kahoot!.

Figure 4.27a shows that in the condition of *high delay*, a great number of students disagreed to the statement that their score was in line with their effort, and claimed that they were not rewarded enough for their effors. The high delay led to students missing out on points as the delay caused technical issues. This is confirmed by looking at the corresponding condition in Figure 4.27b, where 90% of the students agreed that the technical quality reduced their score, and by the average scores presented in Section 4.3.3, where the average score of *high delay* is 3036 while for *no delay* the score is 4700.

The same trend can be observed for the condition of *no delay*. In Figure 4.27a, a great amount of students agreed that their quiz results were in line with their effort, thus only the students' knowledge and efforts impacted the quiz result. In Figure 4.27b this is confirmed as approximately 95% of the students did not believe their score was affected in a negative manner by the technical quality of the session. For the condition of *moderate delay* the amount of students agreeing or disagreeing was about the same.

These results suggest that even though a high percentage of the students were not affected by or did not notice their co-located participants, a perception of unfairness was present.

Chapter 5
Discussion of Results

This chapter will discuss and look further into interesting aspects from the results presented in Chapter 4 and the main focus of this thesis work; to what extent delay affected QoE; investigate how QoE changed over time and how presence of others influenced the experience of fairness.

The results in Chapter 4 show that introduction of delays affects aspects of QoE differently. Delight is somewhat affected, but not in a great manner, while feelings of annoyance are increasing when delay is introduced, and the perception of technical quality is decreasing. It was found that the changes in perceived Evaluation of Quality can be seen as more severe than for the changes of delight and annoyance, and will be further discussed in this chapter. This chapter will also comment on how the impression of Kahoot! changed with the presentation of delays.

5.1 Impact of Influencing Factors

In this thesis work, QoE has been evaluated based on the new definition of QoE proposed by Qualinet [1], which is presented in Section 2.1.2. For this reason QoE has been evaluated through subjective-measures of delight, annoyance and quality using the Likert scale. The new definition moves away from the most used measure of QoE, the MOS scale as used by multiple studies, among these [10], as it has been argued that the scale is insufficiently taking influencing factors into account [13].

Different influencing factors plays a large role in how a test participant makes an evaluation [16, 17]. For the two test setups done in this thesis work, different external influencing factors were present. As previously mentioned, the two tests have been conducted to complement each other, as different factors were controllable and present for the different setups.

First of all, as the Longitudinal study was presented in a real-life environment, influencing factors which can be expected in a natural environment were present for this part of the study. Among these, how students behave among co-located participants and friends. As stated by Zander et al., users of Cloud Gaming might tolerate higher QoS degeneration if they have a strong relationship to other players [18]. Secondly, as it was not feasible to control devices for all test participants in the real-life Longitudinal setting, the Cross-Sectional part of the study focused on fairness. However, the findings indicate that almost 50% of the students participating in the Cross-Sectional study did not pay attention to how they performed compared to other students. It can therefore be argued that the Cross-Sectional lab setting did not create an environment facilitated for interaction in the same manner as it would have been in a real-life environment. It can be assumed that test participants were influenced by the somewhat seriousness of the experiment.

Other uncontrollable influencing factors found present during testing were the state of mind of the test participants as well as their personality. As stated by Blythe [17], user characteristics can influence QoE. If a test participant is feeling fatigue or bright and focused, these are factors that are hard to measure and take into account, but can somewhat affect the reported results. As was found in results from the Cross-Sectional testing, some participants rated annoyance higher than the majority of the students and there was found indications that this may have been due to a more competitive spirit than others. For the Cross-Sectional test, the sample size was small, which made it possible to gain some insight into personality aspects, through the questionnaire. On a larger sample, as can be found for the Longitudinal part of the study, it is close to impossible to separate students based on the cues that may say something about their personality without closely examining each self-reported measure. Anyhow, personality is part of what makes QoE subjective, and as presented in Section 2.1.2 the working definition of QoE states that "Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state"[1]. Thus the personality and current state of the users should be considered when evaluation QoE.

5.2 Impact of Delays

5.2.1 Degree of Delight

By studying the plot in Figure 4.5 on page 45, it can be speculated that there are different reasons for experiencing delight. The plot illustrates the distribution of Degree of Delight in the Longitudinal test where a large part of the scale was utilized for all conditions. This means that while some test participants experienced a high Degree of Delight, others had no feelings of delight during the same conditions. As mentioned previously, a reason for the high Degree of Delight in the conditions with emulated delay could possibly be the influence of factors regarding fun and play associated with participating in the Kahoot!, and is supported by the findings presented in [19] and [20]. The correlations between Positive Influences and Degree of Delight presented in Chapter 4 confirm that the positive attributes of Kahoot, i.e. the social context etc. have an impact on perceived delight, as there is a clear significant positive correlation between the two. This indicates that the emulated delay as such did not have the strongest impact on some of the participants perception of delight. In this context, it can also be referred to the work of Zander et al. who found that players which are captivated by the game may tolerate higher QoS degeneration [18].

The Wilcoxon signed-rank related test results presented in Appendix A, show that there are significant differences between the first test of *no delay* and the two tests of delay, while there are no significant difference between the two tests of *no delay* or between the last test of *no delay* and the two tests of delay. This indicates that delay and previous experiences does not affect delight in a strong manner. The plot in Figure 4.5 on page 45, shows that the self-reported delight remains relatively high, despite the fact that there are some minor indications that delay plays a role. Students who expressed low feelings of delight in the different conditions might have been affected by external influencing factors like their current state of mind or other factors which are difficult to control and measure in this test setup. Another reason for the low rating of delight might be the initial impression some student had of Kahoot!, as it has been found that the test participants who rated Degree of Delight low also characterized Kahoot! as too time consuming or do not see the usefulness of the tool. These students rated delight and impression of Kahoot! similar regardless of the delay condition. If students feel that they are wasting valuable time, not receiving educational benefit, it is perhaps not so surprising that they experienced low delight, regardless of the delay introduced.

Figure 4.14 on page 58 displays the distribution of Degree of Delight from the Cross-Sectional test, indicating a more apparent effect of the introduced delay as the medians show a clearer decrease when delay was introduced. As mentioned, the Wilcoxon signed-rank test found that Degree of Delight was significantly higher in the condition of *no delay* compared to the other two conditions of delay, the test also showed that ratings were significantly higher for *moderate delay* compared to *high delay*. This indicates that delay had a greater impact on delight in the Cross-Sectional test than in the Longitudinal tests, as supported by [40], where it is argued that the specific nature of a social context may significantly influence QoE. A reason for this could be the lack of interactions, as previously discussed, between the students during the Cross-Sectional test. Blythe et al. state that experience is context-related, meaning that the same activity can be delightful in one context, but boring in another [17]. To illustrate this, they use the analogy of playing with friends vs. playing alone. Perhaps the relations between the recruited test panel for the

Cross-Sectional test were not as strong as intended. As the students did not have the opportunity to choose where they could sit during the Cross-Sectional test, they were not immediately close to some of their closest friends, letting the students choose where to sit could have increased interaction, making the context more delightful.

From these results regarding delight, it can be assumed that there were less external factors affecting delight in the Cross-Sectional study due to the unnatural setting of the test and this probably led to test participants being more affected by the delay in this setting compared to the real-life Longitudinal setting. Delight is an emotional state and the absence of influencing factors that would be present in a more natural setting, might lead to reduced feelings of delight. Overall, the discussion above may indicate that Degree of Delight was not affected in a too great manner by the emulated delay, but more by the external influencing factors regarding the context of use, human characteristics and service factors.

5.2.2 Degree of Annoyance

Results presented in Chapter 4 shows that the changes of annoyance between different conditions of delay was greater than the changes of delight, which implies that delay seems to have a greater impact on annoyance than on delight. This also corresponding to the previous argumentation that feelings of happiness, amusement, satisfaction and so on, may be highly present regardless of the occurence of delay. This corresponds with previous findings by Sunde in [15].

The plots in Figure 4.6 on page 46 and 4.15 on page 60, presents the distribution of Degree of Annoyance for the Longitudinal and Cross-Sectional study. The plots show that for both test setups and considering the self-reported annoyance, the lower part of the scale is also used in all the conditions of delay, thus some participants were not as affected by the emulated delay. As discussed above, different influence factors plays an important role, and the personality of some users might affect them in a positive manner preventing feelings of annoyance which could have been provoked by the delay. As QoE is context-related it can as well be speculated whether the students experiencing no or low feelings of annoyance would have felt differently in another context where the quiz had been of more importance to the users (if the results had an impact on their grade etc.). If technical issues in the Kahoot! session would have had a direct effect on the students grades, it might have caused stronger emotions and possibly increased feelings of annoyance among a larger share of students, than what has been found in this study. Even though some students were not particularly annoyed by the introduction of delay, the plots from both test settings show a clear change in rated annoyance between the conditions with and without emulated delay.

As mentioned in Section 4.3.3, in the condition of *high delay* for the Cross-

Sectional setting, multiple students reported on the short timeframe during which they were able to answer a question, which made them stressed and frustrated. It is interesting to see that the frustration does not seem to originate from the delay as such, but more from the fact that delay hampers with the process of completing the task successfully. In this situation, delay can be regarded as an indirect cause to the frustration.

Regardless of the fact that effect of delight may be mediated by other factors, the plots visualizing the correlation between delight and annoyance (Figure 4.11 on page 51 and 4.21 on page 65), do confirm that annoyance in some matter is affected by delay. As can be seen in the figures, there are some correlations between annoyance and perceived technical quality, while delight seems to be more spread and not as affected by the perceived technical quality, confirming the discussion above.

5.2.3 Evaluation of Quality

As mentioned, the Evaluation of Quality can be seen as a less subjective measurement than the other subjective measures of QoE focused on in this thesis work. Results have shown that delay has an impact on all self-report measures of QoE, but to a clearer extent on the quality evaluation aspect than on the perception of delight or annoyance. As mentioned previously, most studies on QoE focus specifically on the technical quality that users experience while using cloud-based services [4, 5, 10, 11]. The findings obtained and presented in this thesis work have shown that delay has a greater impact on the perceived technical quality than on delight and annoyance, and that the rating of technical quality does not necessarily correspond to the feelings of delight and annoyance, it is therefore clear that it is not sufficient to only take the technical quality into account when evaluating QoE.

It can be argued that students might have different preferences in regards of good and bad quality. Students with a higher demand for quality might be more aware of quality degradation and may rate the technical quality lower than the regular user with a "normal" demand for quality. This is also the case when looking at experience, as users with more experience in regards of quality and/or the application have an opinion on how the application should perform at its best and when evaluating technical quality and performance, these preferences and previous experiences are likely to play a role. The effect of experience and high demand for quality has been discussed by Underdal in [56], where it was found a difference in rating of perceived technical quality between these and "normal" user, as they were stricter in their judgment of quality. From this it can be stated that preferences are important external factors.

The trend of rating quality in relation to demands and experience can be part of the explanation why the results from the Longitudinal testing are more spread for the first condition of *no delay* as well as *high delay* while in the result from the last test of *no delay*, quality is rated higher and the ratings are more concentrated. This is supported by [25] saying that QoE is related to fulfillment or violation of expectations, as previous tests create a reference point for the users, affecting expectations. Furthermore, it can be seen for the condition of *moderate delay* for the Cross-Sectional testing that Evaluation of Quality is more spread, and it can be argued that these results are in strong relation to the previous experience of the students. Again it can be stated that experience matters.

5.2.4 Positive and Negative Influences

The variables of Positive and Negative Influences were added to the questionnaire to investigate how the students were affected by co-located participants. As shown by the results presented in Chapter 4, the Friedman ANOVA test (Appendix B) run on the subjective measures from the Cross-Sectional testing, did not yield any significant differences between the variables of Negative Influences and these results could not be utilized further. For the Cross-Sectional study, Positive Influences were investigated and it was found that these results were highly spread, did not give a very clear indication of the effect of co-located others on the participants, and were therefore hard to interpret. As a result, it is difficult to draw conclusions based on the evaluation of Positive Influences. It can be discussed that this is because of the setting of the study, not giving students the ability to communicate in the same manner as during lecture and that the setting was experienced as to serious for interaction among the participants.

When looking into the subjective measures of Positive and Negative Influences from the Longitudinal test, the analysis using the Wilcoxon signed-rank and Mann-Whitney U test found that there were few differences between the ratings. As can be seen in Figure 4.8 on page 49, the medians within the variables do not change between the conditions of delay. Again it can be said that delay did not affect the Positive and Negative Influences caused by co-located participants. On the other hand, there has been found a strong correlation between delight and positive feelings influenced by others, thus indicating that the influence of co-located players is an influencing factor affecting QoE.

5.2.5 Impression of Kahoot!

As mentioned in Chapter 4 the overall impression of Kahoot! is similar for all conditions of delay. While there are minor differences in impressions from the Longitudinal study, the results from the Cross-Sectional test showed that all students described their impression equal for all conditions of delay. The results for both the Longitudinal and the Cross-Sectional experiment is presented in Figure 5.1.

As can be seen in the figure a high percentage of the students rated the application as a "Useful learning tool" creating a "Nice break/diversion during lecture". This might be due to students finding the application valuable, creating increased attention and improved learning. This is supported by [35] and [33], saying that students need to interact in order to make sense of new information and that a CRS can increase the confidence level of the students. Findings from other studies show that use of CRS increases students' attendance, attentiveness, enthusiasm and in-class participation [36, 37, 38]. In addition, Kahoot! is a known tool among the students as it has been utilized in multiple lectures before being presented with a delay. Meaning students could have already had made up their mind, not taking the latest experience of the application into account when giving their answer. This is in particularly clear from the Cross-Sectional testing where none of the participants changed their mind in regards of the application through the three tests.

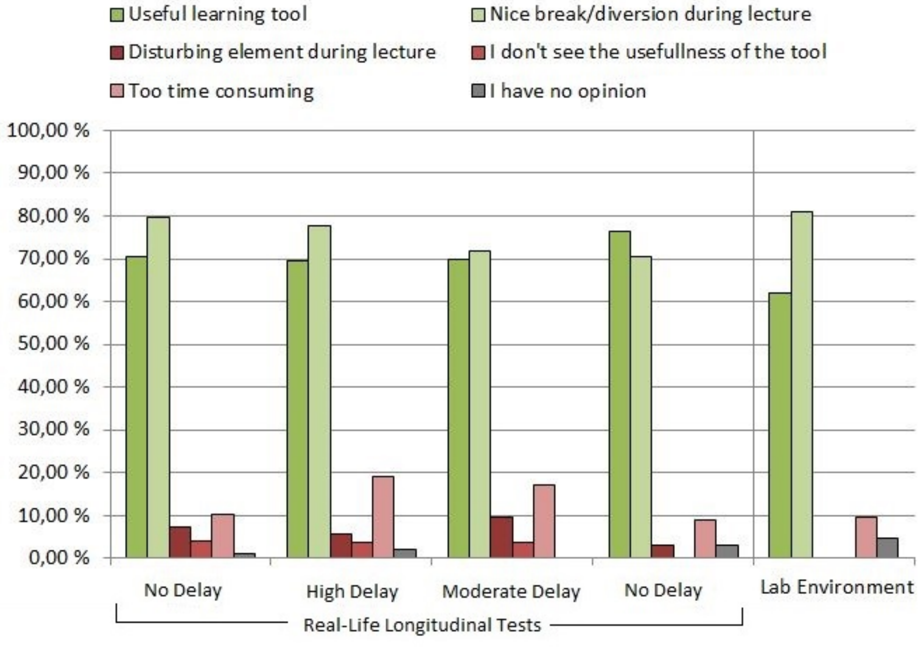


Figure 5.1: Impression of Kahoot!. Comparing different test setups. Multiple descriptions of the application could be chosen for this question.

From the figure it is mostly interesting to look at the column indicating the percentage of students rating the application as "too time consuming". For the conditions with emulated delay, an increase in the respective ratings could be observed and this could be due to the delays that were experienced. What is interesting is that the increase would have been expected to be larger. Again it can be said that students had already made up their mind in regards of Kahoot!, and that the introduction of delay only impacted the students impression of Kahoot! to a small extent.

5.3 Longitudinal Effect

Results regarding the Longitudinal test shows that experience matters and that there are significant differences in how the participants evaluate QoE-indicators when presented with *no delay* in the fourth test, after experiencing conditions of delay.

It has been found that feelings of annoyance did decrease from the first to the last test. In addition, the results show that less participants stated that they experienced delay between the monitor and their device in the last test, as many as 34% agreed to this in the first test while 12% agreed to this in the last test. The findings also indicate that in the first test without delay, 14% believed that they would have obtained a higher score if the technical quality of Kahoot! had been better, while no students believed this for the last test. From the first test, all students who believed that their score was affected, also claimed they experienced delay between their device and the monitor during this test.

First it is interesting that such a large percentage believed to experience delay in the first test as no delay was emulated. As the test was conducted in a real-life setting, it is impossible to confirm or deny if delay actually was experienced and it is difficult to speculate on why such a high percentage believed to experience delay. One student commented: "I noticed significant delay between the lecture screen and my PC". Further investigation shows that this student managed to answer all the questions of the Kahoot! and received an about average score. Other comments suggest that the students "need more time to think before answering". As mentioned, the students have under normal conditions 30 seconds to answer a question. This could indicate that the users misinterpret the concept of delay. If the experiences of students were in fact not influenced by delay, the responses indicating that they experienced delay are probably influenced by their demand for quality and their incorrect/mistaken expectation of how the application should perform at its best.

As mentioned, the percentage of students believing to experience delay did decrease from the first to the last test. 12% can still be seen as a somewhat high percentage, but as the number of students for this test was small, 12% includes only four students. One of these students commented on technical difficulties and had a negative impression of the application and a negative response to all questions regarding quality. As for the remaining three students', the evaluation of technical quality is relatively neutral, this might suggest that these students have misinterpreted the question. This could as well be the case for some of the students (34%) in the

Longitudinal setting, but this is hard to speculate on.

Both statements above indicate that the test participants were more satisfied with the quality in the last test. One indication for this could be that the previous tests with emulated delay changed the test participants reference point regarding quality, leading to a more satisfactory quality when *no delay* was presented after conditions of delay. Another reason could be the fact that the number of participants drastically decreased from the first to the last test.

The Longitudinal study indicates that previous experience matters. As presented in Chapter 2, the definition of QoE [1] states that QoE results from the fulfilment of the users' expectations. Based on the results of this thesis work, it can be stated that when given a new reference point regarding quality, students' expectations with respect to Kahoot! were better met in the last test.

5.4 Fairness

The main goal of the Cross-Sectional test was to complement the real-life Longitudinal test, look into fairness and how users react when students are playing a Kahoot! on different conditions of delay, this creating an unfair setting between the players.

From a closer investigation of the results for the students who did pay attention to co-located participants, it was found that students who believed that they had performed better then the others, also reported on a higher Degree of Delight and a lower Degree of Annoyance than the students who believed they performed poorer than their co-located students. As performing better than others goes hand in hand with an increase in delight, it can be said that students were happy with their performance or felt a higher sense of achievement when performing better than others. However, no conclusions in terms of causality can be drawn and a question remains whether students feel delighted of being "the best" or "better" or delighted as everything is working as expected.

Most indications of a better performance than the co-located others, can be found in the no delay condition. However, such indications were given in the other conditions as well. When confronted with the condition of *high delay*, the majority of students who had an opinion on their performance compared to others indicated that they performed poorer. Moreover, the students who indicated that they performed worse than the others around them, also agreed that they would have received a higher score if the quality of the session would have been better. This indicates that they may have experienced the setting as unfair.

As mentioned, multiple students were taken by the serious of the setting for the Cross-Sectional test. By looking at the students who did not pay attention to the performance of co-located participants, it is interesting to find that for *no delay* these students report a lower delight and a higher annoyance than the students believing they performed better than others. For *high delay* the opposite occurs and the students who did not pay attention to how they others were performing, felt a higher delight and a lower annoyance than the students believing they performed poorer than their co-located participants. This confirms that the awareness of presence of others probably affected the participants' feelings of delight and annoyance.

As mentioned, the Mann-Whitney U test (results can be found in Table 4.4 on page 69) resulted in significant differences for Degree of Delight between the students believing they performed better and poorer than co-located participants as well as between students believing they performed better and those who did not pay attention to others. This confirms that for this setting, Degree of Delight is affected by (the awareness of) the unfair setting. When looking at the results for Degree of Annoyance, a similar result was found and again, there is found a significant difference in feelings of annoyance between the students performing better and poorer than co-located participants as well as between the students believing they performed poorer and did not pay attention to others. It can be stated that feelings of delight and annoyance are enhanced when students are aware of how others perform compared to themselves. As mentioned in Chapter 2, Poels et al. have found that players put more effort into a game when they play against co-located participants [41]. It can be argued that students that are aware of their co-located participants may have been more engaged in the game, and therefore delight was high when they performed well and feelings of annoyance occurred when they performed poorer then they would have expected.

The statements above are supported by what is mentioned in Chapter 2, "a game can be interpreted as fun, delightful, challenging and victorious before a friend effortlessly makes a better score. Then the experience may be reinterpreted more as a waste of time" [16]. This was observed in the Cross-Sectional test, where students lost interest in the quiz when they were unable to answer due to the high delay emulated. Thus the Degree of Delight decreased and the Degree of Annoyance increased. A similar finding has been presented by Zander et al. in [18] and by Chen et al. in [28], these papers states that as the network conditions deteriorates, users experience an increased desire to leave the game.

The discussion above shows that delay, performance and awareness of co-located participants are different factors influencing fairness and the feelings of delight and annoyance, thus the QoE.

5.5 Comparing Longitudinal and Cross-Sectional Setup

In the real-life Longitudinal test the ecological validity was high but control over possible influencing factors was low. In the Cross-Sectional controlled lab environment, different factors regarding delay were more controllable, but the test environment was artificial and did not reflect the natural user context. Regarding this, the two different tests conducted during this thesis work somewhat complement each other by fulfilling each other's limitations. Although the two test setups had different characteristics it is interesting to compare the results.

When comparing medians for the same variables, i.e. delight, annoyance and so on, within the two different test setups, the initial impression is that the medians are somewhat similar and that the findings from the two tests show equal trends. This reinforces the robustness of the findings from the individual studies. Medians for the Longitudinal and Cross-Sectional test is listed in Table 5.1.

	No Delay			**Moderate Delay**		**High Delay**	
	RL 1	RL 4	CS	RL	CS	RL	CS
Delight	3.40	3.40	3.20	3.00	3.00	2.80	2.40
Annoyance	2.00	1.71	1.57	2.57	2.14	2.57	2.57
Quality	3.75	4.25	4.25	-	2.75	2.00	1.50
Positive	3.00	3.00	3.33	3.00	3.33	-	3.00
Negative	2.00	2.00	1.50	2.00	2.00	2.00	2.00

Table 5.1: Comparing medians from Longitudinal and Cross-Sectional setup. RL refers to real-life Longitudinal tests and CS to the Cross-Sectional tests conducted in a lab environment. In the column of *no delay* RL 1 refers to the first test conducted, and RL 4 to the fourth test.

The table show that the medians from the Cross-Sectional testing are similar to the medians resulting from the Longitudinal study. The largest differences that can be observed, resides between the ratings of Evaluation of Quality. There are some differences between the medians for rated annoyance, while the evaluations of delight, positive and negative emotions are close to similar. The medians from the setups show that in the condition of *no delay*, quality was rated higher in the Cross-Sectional setup than in the first test of the Longitudinal setup. This outcome can be a result of the order of delay given in the Cross-Sectional test. In the condition of *no delay* in the Longitudinal test, the participants had not yet experienced a Kahoot! where delay was emulated, but in the Cross-Sectional test, test participants experienced *no delay* in different orders, thus the participants had previously experienced a higher delay prior to this condition. This assumption is strengthened by looking at the last test conducted with *no delay* in the Longitudinal test, where students had experienced delays previously, similar to the same condition in the Cross-Sectional test. As can be seen in the table, the medians for these two tests are equal.

It is also worth mentioning that in the Cross-Sectional setting, the medians for delight, annoyance and quality, are all equal or less than those in the Longitudinal setting. This could indicate that students' opinions in the Cross-Sectional setting were affected by delay in a higher degree, which might be due to the lack of other (non-technical) influencing factors in this setting. This claim supports the necessity of conducting tests evaluating QoE in a real-life setting as well as in a controlled lab environment.

In addition, results have shown that a *moderate delay* is considered too high by the test participants. When a *moderate delay* was emulated in both the Longitudinal and the Cross-Sectional tests, the majority of the students rated the quality low, saying Kahoot! did not function optimally and that the overall quality was not as it should be. This indicating that a delay of 5000 ± 1000 is not acceptable regarding technical quality and a possible acceptable level of delay will reside somewhere below this threshold.

Chapter 6
Limitations and Future Work

The purpose of this chapter is to present limitations that possibly had an impact on test execution and results, as well as reflect on the influence of choices made during the study. In addition this chapter presents possible improvements, which can guide directions for future research.

6.1 Limitations

As mentioned previously, QoE is subjective, causing measurement of QoE to be challenging due to considerations of multiple factors. Human and context related factors like age, gender, personality and cultural background, may influence the results. In the real-life Longitudinal test, the distribution of gender was somewhat skewed, with approximately 20% females and 80% males, somewhat reflecting the actual gender distribution of TTM4100. An option to avoid this skewness could be to conduct the tests in a course with equal distribution of genders. It was attempted to address this skewness and to improve the gender distribution in the Cross-Sectional test: 43% females and 57% males participated. The scope of this thesis work made it difficult to consider all human and context related factors that may impact the results. To ensure reliable results, influencing factors should have been taken into further consideration, as well through questions in regards of personality and relations to co-located participants.

As mentioned in Chapter 3, questionnaires are a widely used data collection method in educational and evaluation research. Obtaining data using scales where test participants indicates to which degree they agree or disagree with a statement, is often used to collect subjective measures. An issue occurring by using these scales is that participants may have dissimilar interpretations about the scales [57]. What one participant defines as excellent, another might define differently, making it difficult to interpret, increasing the value of a large test group.

The test participants were all students and can be seen as a similar group of people. By using a larger and more diverse test panel the results could be found different. Students of technology can be said to be of high demand or expert users, as they have grown up with the newer technologies and everyone in the group had experience within the subject studied.

Another limitation is the age group chosen for this study. A CRS is created for use in classrooms with students of all ages. To test a CRS and its wide spread of use, it could be interesting to look into different age groups to find if these results would differ. On the other hand, by testing on older students, it is safe to say that the students that participated in studies in this thesis work are most likely the more demanding group.

Another factor to be mentioned is the task complexity. Completing quizzes in Kahoot! can be regarded as a simple, repetitive task that requires little problem solving. In cases like this, it has been shown that users want to perform these types of tasks rapidly, and becomes more annoyed by delays than they would if they were working on more complex problems [58]. As a result, the findings presented here are not automatically applicable to other types of applications and services that contains more complex tasks.

6.1.1 Limitations for the Real-Life Longitudinal Testing

As mentioned, in the class of TTM4100 used for testing, attendance was voluntarily. The effect of this was a fast decrease of the number of students participating in class. By the last test, participation in Kahoot! had decreased by about 75%. A limitation influenced by this decreasing participation is the small sample from the last test as well as the fact that it was not possible not follow a larger group of test participants over time and thus create a dataset based on related samples, as was initially the goal. Due to this, it was more complex to run the data through statistical analysis as parts of the data set were related and the other half unrelated. In future work, tests should be done where participation is mandatory.

Finally, a test in a real-world environment is never guaranteed to be executed as intended even though multiple pilot tests have been carried out ahead of testing. In both tests where delay was emulated, the execution did not happen according to plan. A limitation of the setting is therefore that the results may have been affected by the lost connections to the server and how this makes these tests fairly similar. In the third test, a new delay was set after losing connection to the server, would the results have been different if the delay of 5000ms ± 1000ms was continued throughout the quiz and without losing connection? Did the lost connection happened as a result of the large user group or caused by bad luck?

6.1.2 Limitations for the Cross-Sectional Lab Testing

The fact that the Cross-Sectional test was conducted in a controlled lab environment puts limitations to the study, as users were aware of being in an unnatural test setting. Even though most contestants knew each other of some degree, they were taken by the seriousness of the condition and little talk or attention to others occurred. As previously discussed, the lack of interaction between the students in this test possibly impacted the result to some degree. For future work it is important to encourage participants of such a lab study to communicate and interact with the other participants, as they would have in a real-life setting.

The sample size used was small and for this reason results apply to this study and cannot be generalized or draw conclusions for a big population. It is on the other hand interesting to look at the results and the resemblance to the real-life Longitudinal testing.

A different limitation in regards of this part of the study might be the questionnaire. It was decided to use the same questionnaire in both the real-life Longitudinal test and the Cross-Sectional test. For the Longitudinal test the questionnaire worked as intended, and procured the desired responses. For the Cross-Sectional test it was intentional to look further into fairness. In addition to the observation that fairness is a concept that is very difficult to grasp and to create an adequate measurement setting for, it was found that the questions regarding influences by other participants did not produce results in a close connection to fairness (as intended). Not including a wider diversity of questions in regards of fairness somewhat limits the results presented for this topic. The intended items representing fairness in the questionnaire, i.e. the positive and negative influence regarding co-located participants, yielded some interesting results, but have been emphasized less than intended. Instead other items concerning performance compared to co-located participants have been used as a measure for fairness.

6.2 Future Work

The field of Cloud Gaming is a newer paradigm where still little research has been done. As mentioned, to investigate QoE in relation to Cloud Gaming and the impact of a range of influencing factors, a large number of user studies should be conducted in different application domains and research environments (e.g., in the lab, online, in the natural context of use). This thesis work can be seen as a contribution to the field, but more research still needs to be done.

Studies similar to the one conducted here should be done to investigate QoE of cloud-based gaming when considering other game genres. Morover, more research should be done in regards of CRS and how QoE is affected by different service parameters like jitter and packet loss, in addition to other technical parameters such as dependability and security attributes. It would be interesting to look at how different age groups and user characteristics reacts to degradation of these service parameters, including delay.

In future work, studies should be done including quizzes with questions of different difficulty levels. In addition, it would be interesting to compare results from tests between users who know each other and users who do not, to explore how these relationships and their intensity affect the results. Zander et al. states that users of Cloud Gaming may tolerate higher QoS degeneration if they have strong relationships to other players, as they become more captivated by the game [18].

In this thesis work subjective measures have been collected through alternative measures in questionnaires. As previously mentioned, MOS is a widely used scale when attempting to measure QoE, but as the definition of QoE is changing so should the tools used to measure QoE. For this reason a range of alternative measures of QoE, measured on 5-point Likert scales, were used in this thesis work and items with same construct were computed into new subjective measures of delight, annoyance, quality, and positive and negative influences of co-located participants. The findings in this thesis work are linked to the chosen measures of QoE in the questionnaires, meaning that the results are based on items included in the questionnaires. If other items and measures had been used, there is a possibility that this would have affected the results.

In an attempt to strengthen the reliability of the chosen items and states of the subjective measures of QoE, these comprised of several items of similar construct. To strengthen the measures of QoE even further it would be interesting for future work to add other measures, such as behavioral measures (e.g., measures of people's facial expressions, mouse or keyboard pressure), physiological measures (e.g., heart beat, galvanic skin response) and other self-report measures. Another alternative measure to consider is the Self-Assessment Manikin (SAM) scale which is a pictorial rating system to measure pleasure, arousal and dominance. Due to the use of pictures and its non-verbal design, this rating system is usable regardless of age, educational or cultural background of the test participants [59] and has proven to give more correct interpretations. It could be interesting to utilize the method used by Nacke et al. [60] and collect data from facial expression and pressure on different devices (key-board, computer mouse etc.), as well as the self-report questionnaire. According to Nacke et al. [60] a multi-measure approach would enable better characterization of player experience.

It could be interesting to take the real-life Longitudinal experiment further. A longitudinal study is described as tests done over a longer period of time. The tests done in this study were done with only a week apart. An idea could be to conduct more tests further apart and again repeating some of the conditions to investigate whether perceived QoE changes and whether test participants still learn from their experience as tests are done further apart. Future studies should include a larger user group, for example a class with a large student group registered to take the course as well as mandatory attendance, to ensure a large sample of related data.

As mentioned previously, during the Cross-Sectional test students had to leave the room between the quizzes for new levels of delays to be emulated on their answering devices. This procedure was somewhat time-consuming and a better approach would be to create a script in advance that would change delays automatically.

To avoid having to report on experiments where tests were not carried out as intended, multiple tests should be conducted for each condition tested. This to make sure that the results represent the conditions that were intended to be tested, as well as to be able to look at changes in experience and changes of opinion over time.

The overall QoE is influenced by delight and annoyance, but the relative importance between the influencing factors is not well understood, might vary over time and depends on the application. It can be said that degradation of technical quality influences annoyance in a stronger degree than delight. Results from this thesis work indicate that annoyance is in some manner affected by degradation of technical quality, while delight is not affected as much, meaning that a single measure focused on perception of (technical) quality is not sufficient when attempting to evaluate delight. There are several factors influencing the Degree of Delight, among other the context with co-located participants, as shown by the results. It would be interesting for future work to attempt to identify more of the factors influencing delight.

Chapter 7

Conclusion

In this thesis work the cloud-based CRS Kahoot! has been studied through a real-life and a controlled lab environment. This was done to investigate to which degree delay impacts user-perceived QoE, how QoE changes over time and how the presence of co-located participants influences the experience of fairness. QoE has been evaluated based on the new definition of QoE proposed by Qualinet [1], and for this reason QoE has been evaluated through subjective-measures of delight, annoyance and quality. The study has been conducted by introducing different conditions of delay in a real-life longitudinal setup (N=175) as well as in a controlled lab environment using a cross-sectional setup (N=21). The two different test setups were used to complement each other regarding limitations and influencing factors.

Based on the self-reported measures, results have shown that delays emulated on a Kahoot! session affects the users' perception of delight, annoyance and quality and that the introduction of delay in previous sessions affects students ratings of these variables when given a new reference point. The evaluation of quality is the most sensitive to delay, and this is probably due to the less subjective characteristics of quality, while Degree of Delight is impacted the least, this probably due to the fact that external factors may influence feelings of delight. Moreover, it can be stated that emulation of delay does not impact the players' overall perception of the application to a very large extent. Despite the fact that the conditions of delay changed, Kahoot! was still characterized by most players as a useful learning tool creating a nice diversion during class.

Because of the impact external factors have on users, measurement of QoE is difficult. QoE is multidimensional and thus incorporates the influence of non-technical aspects such as user characteristics and the context of use. It is hard to analyze all the external factors as they are subjective and complex to document. As no significant correlation has been found between the self-reported Degree of Delight and Annoyance, it can be stated that QoE is affected by multiple factors and based on the results it can be stated that delay is one of them. Among these, the result indicate that feelings of how a student performs in comparison to co-located participants and when experiencing an unfair setting, enhance the feelings of either delight or annoyance depending on which side of the unfair setting the student is located.

It was found that a large range of influencing factors were highly present in the real-life experiment, while in the lab experiment, students were taken by the seriousness of the setting excluding some of these factors, and for this reason it is highly important for future work to further investigate QoE in a real-life setting.

In regards of quality, results from this study showed that the *moderate delay* has been characterized as not acceptable by the users of the application Kahoot!. There might still be a degree of delay acceptable to users, but this resides below a delay of 5000ms ± 1000ms.

To sum up, an unfair setting where some students experience delay while others do not, enhances the feelings of annoyance among the affected users while it increases the feelings of delight among the users who are not affected by delay. When given a reference point, students are affected by this, changing their QoE in relation to a certain delay condition.

This thesis work is an attempt to make a contribution to the literature on QoE in the context of Cloud Gaming and CRS, but the findings need to be further explored and validated in follow-up research. Moreover, new questions have been raised, which can guide directions for future research.

References

[1] P. Le Callet, S. Möller, and A. Perkis, "Qualinet white paper on definitions of quality of experience," *European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version*, vol. 1, 2012.

[2] S. Flæten, "Teknologibragdens stolte vinner," *Teknisk Ukeblad*, vol. 161, no. 3, pp. 65–67, 2014.

[3] N. Parker, "Distribution and monetization strategies to increase revenues from cloud gaming," *Cloud Gaming Report*, 2012.

[4] T. Hoßfeld, R. Schatz, M. Varela, and C. Timmerer, "Challenges of qoe management for cloud applications," *Communications Magazine, IEEE*, vol. 50, no. 4, pp. 28–36, 2012.

[5] K. Vandenbroucke, K. De Moor, and L. De Marez, "Use-and qoe-related aspects of personal cloud applications: An exploratory survey," in *Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on*, pp. 36–37, IEEE, 2013.

[6] R. Schatz, T. Hoßfeld, L. Janowski, and S. Egger, "From packets to people: Quality of experience as a new measurement challenge," in *Data Traffic Monitoring and Analysis*, pp. 219–263, Springer, 2013.

[7] Microsoft, "What is qos?." http://technet.microsoft.com/en-us/library/cc757120(v=ws.10).aspx, March 2003. Accessed: Feb. 5th 2014.

[8] ITU-T Rec. E. 800, "*Terms and definitions related to quality of service*," Int. Telecomm. Union, Genova, 2008.

[9] M. Siller and J. Woods, "Improving quality of experience for multimedia services by qos arbitration on a qoe framework," in *in Proc. of the 13th Packed Video Workshop 2003*, Citeseer, 2003.

[10] P. Amrehn, K. Vandenbroucke, T. Hoßfeld, K. De Moor, M. Hirth, R. Schatz, and P. Casas, "Need for speed? on quality of experience for cloud-based file storage services," International Workshop on Perceptual Quality of Systems Vienna, Austria, 2013.

[11] S. Möller, D. Pommer, J. Beyer, and J. Rake-Revelant, "Factors influencing gaming qoe: Lessons learned from the evaluation of cloud gaming services," 2013.

[12] I. Rec, "P. 10 (2007) vocabulary for performance and quality of service," *International Telecommunication Union, Geneva.*

[13] 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 qoe," pp. 90140U–90140U, 2014.

[14] J. McCarthy and P. Wright, "Technology as experience," *interactions*, vol. 11, no. 5, pp. 42–43, 2004.

[15] M. Sunde, "Evaluation of qoe in cloud gaming," NTNU, 2013.

[16] L. Ermi and F. Mäyrä, "Fundamental components of the gameplay experience: Analysing immersion," *Worlds in Play: International Perspectives on Digital Games Research. New York: Peter Lang Publishers*, pp. 37–53, 2007.

[17] M. Blythe and M. Hassenzahl, "The semantics of fun: Differentiating enjoyable eeperiences," in *Funology*, pp. 91–100, Springer, 2005.

[18] S. Zander and G. Armitage, "Empirically measuring the qos sensitivity of interactive online game players," in *Australian Telecommunications Networks & Applications Conference*, pp. 8–10, 2004.

[19] B. Gajadhar, Y. de Kort, and W. IJsselsteijn, "Influence of social setting on player experience of digital games," in *CHI'08 extended abstracts on Human factors in computing systems*, pp. 3099–3104, ACM, 2008.

[20] N. Ravaja, T. Saari, M. Turpeinen, J. Laarni, M. Salminen, and M. Kivikangas, "Spatial presence and emotions during video game playing: Does it matter with whom you play?," *Presence: Teleoperators and Virtual Environments*, vol. 15, no. 4, pp. 381–392, 2006.

[21] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hoßfeld, "An evaluation of qoe in cloud gaming based on subjective tests," in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2011 Fifth International Conference on*, pp. 330–335, IEEE, 2011.

[22] M. Claypool and K. Claypool, "Latency and player actions in online games," *Communications of the ACM*, vol. 49, no. 11, pp. 40–45, 2006.

[23] M. Claypool and K. Claypool, "Latency can kill: precision and deadline in online games," in *Proceedings of the first annual ACM SIGMM conference on Multimedia systems*, pp. 215–222, ACM, 2010.

[24] R. Shea, J. Liu, E.-H. Ngai, and Y. Cui, "Cloud gaming: architecture and performance," *Network, IEEE*, vol. 27, no. 4, 2013.

[25] P. Chen and M. El Zarki, “Perceptual view inconsistency: an objective evaluation framework for online game quality of experience (qoe),” in *Proceedings of the 10th Annual Workshop on Network and Systems Support for Games*, p. 2, IEEE Press, 2011.

[26] P. Mell and T. Grance, “The nist definition of cloud computing (draft),” *NIST special publication*, vol. 800, no. 145, p. 7, 2011.

[27] S. Wang and S. Dey, “Modeling and characterizing user experience in a cloud server based mobile gaming approach,” in *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, pp. 1–7, IEEE, 2009.

[28] K.-T. Chen, P. Huang, and C.-L. Lei, “How sensitive are online gamers to network quality?,” *Communications of the ACM*, vol. 49, no. 11, pp. 34–38, 2006.

[29] J. Jansz and L. Martens, “Gaming at a lan event: the social context of playing video games,” *New media & society*, vol. 7, no. 3, pp. 333–355, 2005.

[30] C. P. Fulford and S. Zhang, “Perceptions of interaction: The critical predictor in distance education,” *American Journal of Distance Education*, vol. 7, no. 3, pp. 8–21, 1993.

[31] C. Chou, “Interactivity and interactive functions in web-based learning systems: a technical framework for designers,” *British Journal of Educational Technology*, vol. 34, no. 3, pp. 265–279, 2003.

[32] K. Siau, H. Sheng, and F.-H. Nah, “Use of a classroom response system to enhance classroom interactivity,” *Education, IEEE Transactions on*, vol. 49, no. 3, pp. 398–403, 2006.

[33] E. Suchman, K. Uchiyama, R. Smith, and K. Bender, “Evaluating the impact of a classroom response system in a microbiology course,” *Microbiology Education*, vol. 7, p. 3, 2006.

[34] V. M. Borden and K. L. Burton, *The impact of class size on student performance in introductory courses.* 1999.

[35] L. S. Vygotsky, *Mind in society: The development of higher psychological processes.* Harvard university press, 1980.

[36] D. Bullock, V. LaBella, T. Clingan, Z. Ding, G. Stewart, and P. Thibado, “Enhancing the student-instructor interaction frequency,” *The Physics Teacher*, vol. 40, no. 9, pp. 535–541, 2003.

[37] J. Roschelle, W. R. Penuel, and L. Abrahamson, “Classroom response and communication systems: Research review and theory,” in *Annual Meeting of the American Educational Research Association, San Diego, CA*, pp. 1–8, 2004.

[38] E. Wit, “Who wants to be... the use of a personal response system in statistics teaching,” *MSOR Connections*, vol. 3, no. 2, pp. 14–20, 2003.

[39] D. Duncan, “Clickers in the classroom,” *Addison: San Francisco, CA*, 2005.

[40] C. Bracken, R. Lange, and J. Denny, “Online video games and gamers' sensations of spatial, social, and co-presence',” in *Proceedings of the 2005 FuturePlay Conference*, 2005.

[41] K. Poels, Y. de Kort, and W. Ijsselsteijn, “It is always a lot of fun!: exploring dimensions of digital game experience using focus group methodology,” in *Proceedings of the 2007 conference on Future Play*, pp. 83–89, ACM, 2007.

[42] Mobitroll, “Products by mobitroll.” http://www.mobitroll.no/. Accessed: February 25th 2014.

[43] C. J. Bonk and C. R. Graham, *The handbook of blended learning: Global perspectives, local designs.* John Wiley & Sons, 2012.

[44] S. Øygarden Flæten, “Sjokkert og glad vinner av teknologibragden.” http://www.tu.no/it/2014/02/06/sjokkert-og-glad-vinner-av-teknologibragden, Jan 2014. Accessed: Feb 11th 2014.

[45] G. Anderson, “This norwegian edtech startup is growing 150,000 users a week.” http://www.arcticstartup.com/2014/03/18/this-norwegian-edtech-startup-is-growing-100000-users-a-week, March 2014. Accessed: May 3rd 2014.

[46] E. L. Solbu, “Gameshow i forelesningssalen.” http://www.nrk.no/viten/gameshow-i-forelesningssalen-1.11516268, Feb 2014. Accessed: Feb 5th 2014.

[47] A. Juergen Haas, “Linux/unix command: ping.” http://linux.about.com/od/commands/l/blcmdl8_ping.htm. Accessed: Feb 14th 2014.

[48] Linux-Foundation, “Netem.” http://www.linuxfoundation.org/collaborate/workgroups/networking/netem, Nov. 2009. Accessed: January 23rd 2014.

[49] A. Jurgelionis, J.-P. Laulajainen, M. Hirvonen, and A. I. Wang, “An empirical study of netem network emulation functionalities,” in *Computer Communications and Networks (ICCCN), 2011 Proceedings of 20th International Conference on*, pp. 1–6, IEEE, 2011.

[50] P. Dash, “Bandwidth throttling with netem network emulation.” http://www.linuxforu.com/2012/06/bandwidth-throttling-netem-network-emulation/, June 2012. Accessed: January 23rd 2014.

[51] D. C. Robinson, Y. Jutras, and V. Craciun, “Subjective video quality assessment of http adaptive streaming technologies,” *Bell Labs Technical Journal*, vol. 16, no. 4, pp. 5–23, 2012.

[52] R. Likert, “A technique for the measurement of attitudes.,” *Archives of psychology*, 1932.

[53] SurveyMonkey, "The likert scale explained." https://www.surveymonkey.com/mp/likert-scale/. Accessed: Feb. 11th 2014.

[54] LærdStatistics, "The ultimate ibm spss guides." https://statistics.laerd.com/. Accessed: Feb. 13th 2014.

[55] A. Field, *Discovering statistics using SPSS*, vol. 3. London: Sage publications, 2009.

[56] A. Underdal, "Qoe in adaptive video streaming - research of upper and lower bit rate boundaries," NTNU, 2013.

[57] P. E. Rossi, Z. Gilula, and G. M. Allenby, "Overcoming scale usage heterogeneity: A bayesian hierarchical approach," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 20–31, 2001.

[58] S. B. Shneiderman and C. Plaisant, "Designing the user interface 4 th edition," *ed: Pearson Addison Wesley, USA*, 2005.

[59] S. Reichert, "Self-assessment manikin (sam)." http://www.qu.tu-berlin.de/menue/forschung/laufende_projekte/joyofuse/joy_of_use/joy_of_use/measurement_methods/sam/. Accessed: May 06th 2009.

[60] L. E. Nacke, A. Drachen, K. Kuikkaniemi, J. Niesenhaus, H. J. Korhonen, v. d. W. Hoogen, K. Poels, W. IJsselsteijn, and Y. Kort, "Playability and player experience research," in *Proceedings of DiGRA*, 2009.

Appendix A

Real-Life Longitudinal Wilcoxon Signed-Rank and Mann-Whitney Test Results

The Wilcoxon signed-rank and Mann-Whitney U test has been run on the different variables of Degree of Delight, Degree of Annoyance, Evaluation of Quality, Positive and Negative Influences. The tests are looking into the subjective measures of QoE, consisting of multiple items of self reported measures, and compare the resulting variable of each delay with each other. This is to investigate whether there are significant differences in the ratings of delight, annoyance, quality, positive and negative influences, when considering the different delays.

In Section 4.2.4 on page 43 it is explained that the test results had to be split up between unrelated and related measures to run statistical analysis and look for significant differences. Wilcoxon signed-rank test was run on the related measures while the Mann-Whitney U test was run on the unrelated measures, both to find where significant differences are located. This Appendix presents the result from these tests. The results are lined up presenting the results of the unrelated and related measures of the same delays next to each other for easier comparison and understanding of the results.

A difference between two tests are significant from a statistical point of view whenever the p-value is below 0.05. Any p-value above 0.05 is characterized as not significant, when this event occurs it has been marked in the table as *ns*, indicating not significant. The T-value from the Wilcoxon signed-rank test is the test statistic while the median value is the median value of the values calculated by the Cronbach's alpha test for each group being compared. From the Mann-Whitney U test the U-value is the test statistic. As the test groups produce dissimilar distributions, the mean rank is presented as well as the median value for the unrelated measures. The group with the highest mean rank has a greater number of higher scores. Meaning if the p-value indicates a significant difference, there should be a greater difference between the two mean ranks in the comparison. If the mean ranks are close to similar, this confirms that the differences between the variables being tested are not significant.

Degree of Delight									
	Unrelated Measures					**Related Measures**			
Delay	N	Mean Rank	Median	U	P	N	Median	T	P
No(1) vs.	53	63.03	3.2	1429.5	*ns*	43	3.4	7	0.000
High(2)	63	54.69	3.0				2.4		
No(1) vs.	68	50.59	3.2	538.0	0.013	29	3.4	5	0.004
Moderate(3)	24	34.92	3.0				2.8		
No(1) vs.	76	44.80	3.3	479.0	*ns*	20	3.5	7	*ns*
No(4)	13	46.15	3.2				3.6		
High(2) vs.	70	46.65	3.0	409.5	0.046	35	2.4	12	0.050
Moderate(3)	17	33.09	2.2				3.0		
High(2) vs.	84	44.68	2.8	183.0	0.001	22	2.8	9	*ns*
No(4)	11	73.36	3.6				3.2		
Moderate(3) vs.	32	18.23	2.9	55.5	0.000	21	3.0	7	*ns*
No(4)	12	33.88	3.6				3.0		

Table A.1: Unrelated Mann-Whitney U and related Wilcoxon signed-rank test results for Degree of Delight. The value inside parenthesis denotes which test it is referred to, i.e. (1): First test, no delay. (2) Second test, high delay. (3): Third test, moderate delay. (4):Fourth test, no delay

Degree of Annoyance									
	Unrelated Measures					**Related Measures**			
Delay	N	Mean Rank	Median	U	P	N	Median	T	P
No(1) vs.	53	48.67	2.00				1.86		
High(2)	63	66.77	2.43	1148.5	0.004	43	2.71	5	0.000
No(1) vs.	68	41.35	1.71				2.14		
Moderate(3)	24	61.10	2.36	465.5	0.002	29	2.71	5	0.001
No(1) vs.	76	46.22	2.00				1.93		
No(4)	13	37.88	1.43	401.5	*ns*	20	1.785	8	*ns*
High(2) vs.	70	42.91	2.43				2.86		
Moderate(3)	17	48.47	2.57	519.0	*ns*	35	2.71	11	*ns*
High(2) vs.	84	49.74	2.57				2.715		
No(4)	11	34.68	2.29	315.5	*ns*	22	1.57	4	0.001
Moderate(3) vs.	32	25.45	2.57				2.86		
No(4)	12	14.63	1.86	97.5	0.011	21	1.57	2	0.001

Table A.2: Unrelated Mann-Whitney U and related Wilcoxon signed-rank test results for Degree of Annoyance. The value inside parenthesis denotes which test it is referred to, i.e. (1): First test, no delay. (2) Second test, high delay. (3): Third test, moderate delay. (4):Fourth test, no delay

Evaluation of Quality									
	Unrelated Measures					**Related Measures**			
Delay	N	Mean Rank	Median	U	P	N	Median	T	P
No(1) vs. High(2)	53	83.15	3.75	363.0	0.0	43	3.6	1	0.0
	63	37.76	2.25				1.75		
No(1) vs. No(4)	76	43.2	3.75	357.0	*ns*	20	3.5	1	0.0
	13	55.54	3.75				4.375		
High(2) vs. No(4)	84	42.76	2.00	21.5	0.0	22	1.75	1	0.0
	11	88.05	4.25				4.0		

Table A.3: Unrelated Mann-Whitney U and related Wilcoxon signed-rank test results for Evaluation of Quality. The value inside parenthesis denotes which test it is reffered to, i.e. (1): First test, no delay. (2) Second test, high delay. (4):Fourth test, no delay

Positive Influences									
	Unrelated Measures					**Related Measures**			
Delay	N	Mean Rank	Median	U	P	N	Median	T	P
No(1) vs.	68	47.03	3.0				3.33		
Moderate(3)	24	45.00	3.0	780.0	*ns*	29	3.00	6	0.011
No(1) vs.	76	44.91	3.0				3.33		
No(4)	13	45.54	3.0	487.0	*ns*	20	3.165	7	*ns*
Moderate(3) vs.	32	20.13	3.0				3.0		
No(4)	12	28.83	3.5	116.0	0.046	21	3.0	7	*ns*

Table A.4: Unrelated Mann-Whitney U and related Wilcoxon signed-rank test results for Positive Influences. The value inside parenthesis denotes which test it is referred to, i.e. (1): First test, no delay. (2) Second test, high delay. (3): Third test, moderate delay. (4):Fourth test, no delay

Negative Influences									
	Unrelated Measures					Related Measures			
Delay	N	Mean Rank	Median	U	P	N	Median	T	P
No(1) vs.	53	57.56	2.00	1619.5	*ns*	43	2.00	11	0.034
High(2)	63	59.29	2.00				2.00		
No(1) vs.	68	46.63	2.00	807.0	*ns*	29	2.25	10	*ns*
Moderate(3)	24	46.13	2.00				2.25		
No(1) vs.	76	45.35	2.00	467.5	*ns*	20	2.25	6	*ns*
No(4)	13	42.96	2.00				2.125		
High(2) vs.	70	45.15	2.00	514.5	*ns*	35	2.00	11	*ns*
Moderate(3)	17	39.26	1.75				2.25		
High(2) vs.	84	48.18	2.00	446.5	*ns*	22	2.00	6	*ns*
No(4)	11	46.59	2.50				2.00		
Moderate(3) vs.	32	23.27	2.00	167.5	*ns*	21	2.25	4	*ns*
No(4)	12	20.46	1.38				2.00		

Table A.5: Unrelated Mann-Whitney U and related Wilcoxon signed-rank test results for Negative Influences. The value inside parenthesis denotes which test it is referred to, i.e. (1): First test, no delay. (2) Second test, high delay. (3): Third test, moderate delay. (4):Fourth test, no delay

Appendix B
Cross-Sectional Friedman ANOVA Test Results

For the Cross-Sectional testing, the Friedman ANOVA was run on the new subjective measures of QoE, consisting of multiple items from the questionnaire, to find how delay affects QoE. These items can be found in Table 3.5 on page 30 and has been cleared as consistent items by the Cronbach's alpha test, as can be seen in Table 4.2 on page 56, before computed into new variables.

The Friedman ANOVA was then run on the subjective measures of QoE, to check for significant differences between the different delay conditions. As can be seen in Table B.1, presented below, significant differences was found within the conditions for every variable except from Negative Influences. This meaning that the Negative Influence of co-located participants is not significantly affected by the different delay settings.

As the test has pointed to significant differences among Degree of Delight, Degree of Annoyance, Evaluation of Quality and Positive Influences, these self-reported measures have been affected by delay, and have been further evaluated by the Wilcoxon signed-rank test, and these results can be found in Table 4.3a on page 57.

		Degree of Delight	**Degree of Annoyance**	**Evaluation of Quality**	**Positive Influence**	**Negative Influence**
	P-value	0.002	0.000	0.000	0.001	ns
	Chi-square	12.382	16.481	37.544	14.394	5.320
	df	2	2	2	2	2
Median	*no delay*	3.2000	1.5714	4.2500	3.3333	1.5000
	moderate delay	3.0000	2.1429	2.7500	3.3333	2.0000
	high delay	2.4000	2.5714	1.5000	3.0000	2.0000

Table B.1: Results from Friedman ANOVA test comparing new variables for significant differences in the three conditions of delay, N=21. A p-value less than 0.05 indicates that there exist a significant difference between two or more variables in the sample. The Chi-square is the test statistic. df referrers to the degree of freedom.

Appendix C Correlations

The different variables measuring QoE, Degree of Delight, Degree of Annoyance, perceived technical quality and positive influences has been tested to find if there exist a correlation in how they were rated on each delay. This means that the different resulting variables have been checked for correlation to each other for the different delay setting. The mentioned variables above have been tested for correlations with two questions from the questionnaire: whether the student believed obtained score was in line with the effort put into the quiz and whether the student believed she should have received a higher score, as her score was affected by delay. All variables were to be answered on a five point Likert scale, 1 being strongly disagree or not at all and 5 being strongly agree or Extremely.

To do this, the Spearman's rank-order correlation was used. A perfect correlation is found when correlation = 1, and 0 marks no correlation. A correlation of 0.3-0.6 is considered a low correlation, 0.6-0.8 a clear correlation and a correlation of 0.8-1 as very high. A correlation can be seen as significant when the significance value is equal or lower than 0.05, significant correlations has been marked in the following tables by one or two asterisks (*). If a correlation is negative this means that one variable increases while the other variable decreases. An example to this could be a negative correlation between delight and annoyance saying that when delight is high, annoyance is low and opposite.

C.1 Correlation Longitudinal Testing

No Delay Test 1							
		Degree of Delight	Degree of Annoyance	Evaluation of Quality	Positive Influences	Results in line with effort	Deserved Better Score
Degree of Delight	Correlation	1,000	,051	,193	,279**	-,219*	,034
	Sig.(2-tailed)	.	,624	,060	,006	,032	,743
Degree of Annoyance	Correlation	,051	1,000	-,223*	,075	,147	-,142
	Sig.(2-tailed)	,624	.	,029	,470	,154	,168
Evaluation of Quality	Correlation	,193	-,223*	1,000	,026	-,043	,005
	Sig.(2-tailed)	,060	,029	.	,804	,679	,959
Positive Influences	Correlation	,279**	,075	,026	1,000	-,111	-,143
	Sig.(2-tailed)	,006	,470	,804	.	,281	,164
Results in line with Effort	Correlation	-,219*	,147	-,043	-,111	1,000	-,252*
	Sig.(2-tailed)	,032	,154	,679	,281	.	,012
Deserved Better score	Correlation	,034	-,142	,005	-,143	-,252*	1,000
	Sig.(2-tailed)	,743	,168	,959	,164	,012	.

Table C.1: (N = 96). * = Correlation is significant at the 0.05 level (2-tailed). ** = Correlation is significant at the 0.01 level (2-tailed).

*Correlation coefficient

High Delay Test 2						
		Degree of Delight	**Degree of Annoyance**	**Evaluation of Quality**	**Results in line with effort**	**Deserved Better Score**
Degree of Delight	Correlation	1,000	-,107	,374**	,096	-,085
	Sig.(2-tailed)	.	,273	,000	,327	,388
Degree of Annoyance	Correlation	-,107	1,000	-,381**	-,085	,206*
	Sig.(2-tailed)	,273	.	,000	,384	,034
Evaluation of Quality	Correlation	,374**	-,381**	1,000	,010	-,080
	Sig.(2-tailed)	,000	,000	.	,916	,416
Results in line with Effort	Correlation	,096	-,085	,010	1,000	-,404**
	Sig.(2-tailed)	,327	,384	,916	.	,000
Deserved Better score	Correlation	-,085	,206*	-,080	-,404**	1,000
	Sig.(2-tailed)	,388	,034	,416	,000	.

Table C.2: (N = 106). * = Correlation is significant at the 0.05 level (2-tailed). ** = Correlation is significant at the 0.01 level (2-tailed).

Moderate Delay Test 3						
		Degree of Delight	**Degree of Annoyance**	**Positive Influences**	**Results in line with effort**	**Deserved Better Score**
Degree of Delight	Correlation	1,000	-,160	,598**	-,117	-,026
	Sig. (2-tailed)	.	,251	,000	,402	,856
Degree of Annoyance	Correlation	-,160	1,000	,042	,060	-,153
	Sig. (2-tailed)	,251	.	,767	,668	,275
Positive Influences	Correlation	,598**	,042	1,000	-,197	,090
	Sig. (2-tailed)	,000	,767	.	,157	,522
Results in line with Effort	Correlation	-,117	,060	-,197	1,000	-,402**
	Sig. (2-tailed)	,402	,668	,157	.	,003
Deserved Better score	Correlation	-,026	-,153	,090	-,402**	1,000
	Sig. (2-tailed)	,856	,275	,522	,003	.

Table C.3: (N =53). * = Correlation is significant at the 0.05 level (2-tailed). ** = Correlation is significant at the 0.01 level (2-tailed).

Evaluation of Quality has not been tested for correlation for this test, as the Cronbach's alpha test did not find a high enough alpha value to create a new variable. Instead each items that should have been part of the new quality variable has been tested for correlation with the variables presented in Table C.3. As there was not found any correlations among the items of Evaluation of Quality and the variables presented in the table, these results are not included in the table.

No Delay Test 4							
		Degree of Delight	**Degree of Annoyance**	**Evaluation of Quality**	**Positive Influences**	**Results in line with effort**	**Deserved Better Score**
Degree of Delight	Correlation	1,000	,057	,405*	,760**	,369*	,006
	Sig. (2-tailed)	.	,754	,020	,000	,035	,974
Degree of Annoyance	Correlation	,057	1,000	-,036	,102	-,130	,019
	Sig. (2-tailed)	,754	.	,842	,574	,471	,917
Evaluation of Quality	Correlation	,405*	-,036	1,000	,551**	,137	,092
	Sig. (2-tailed)	,020	,842	.	,001	,446	,612
Positive Influences	Correlation	,760**	,102	,551**	1,000	,358*	,047
	Sig. (2-tailed)	,000	,574	,001	.	,041	,797
Results in line with Effort	Correlation	,369*	-,130	,137	,358*	1,000	-,024
	Sig. (2-tailed)	,035	,471	,446	,041	.	,893
Deserved Better score	Correlation	,006	,019	,092	,047	-,024	1,000
	Sig. (2-tailed)	,974	,917	,612	,7976	,893	.

Table C.4: (N = 33). * = Correlation is significant at the 0.05 level (2-tailed). ** = Correlation is significant at the 0.01 level (2-tailed).

C.2 Correlations Cross-Sectional Testing

No Delay							
		Degree of Delight	**Degree of Annoyance**	**Evaluation of Quality**	**Positive Influences**	**Results in line with effort**	**Deserved Better Score**
Degree of Delight	Correlation	1,000	,226	,516*	,336	,313	-,090
	Sig. (2-tailed)	.	,325	,017	,137	,167	,698
Degree of Annoyance	Correlation	,226	1,000	-,240	-,154	-,304	,341
	Sig. (2-tailed)	,325	.	,294	,505	,181	,131
Evaluation of Quality	Correlation	,516*	-,240	1,000	,356	,475*	-,584**
	Sig. (2-tailed)	,017	,294	.	,113	,030	,005
Positive Influences	Correlation	,336	-,154	,356	1,000	,037	,042
	Sig. (2-tailed)	,137	,505	,113	.	,874	,856
Results in line with Effort	Correlation	,313	-,304	,475*	,037	1,000	-,610**
	Sig. (2-tailed)	,167	,181	,030	,874	.	,003
Deserved Better score	Correlation	-,090	,341	-,584**	,042	-,610**	1,000
	Sig. (2-tailed)	,698	,131	,005	,856	,003	.

Table C.5: (N = 21). * = Correlation is significant at the 0.05 level (2-tailed). ** = Correlation is significant at the 0.01 level (2-tailed).

*Correlation coefficient

Moderate Delay							
		Degree of Delight	**Degree of Annoyance**	**Evaluation of Quality**	**Positive Influences**	**Results in line with effort**	**Deserved Better Score**
Degree of Delight	Correlation	1,000	,013	,231	,529*	,102	,028
	Sig. (2-tailed)	.	,956	,314	,014	,659	,905
Degree of Annoyance	Correlation	,013	1,000	-,575**	-,043	-,630**	,609**
	Sig. (2-tailed)	,956	.	,006	,854	,002	,003
Evaluation of Quality	Correlation	,231	-,575**	1,000	,336	,675**	-,657**
	Sig. (2-tailed)	,314	,006	.	,136	,001	,001
Positive Influences	Correlation	,529*	-,043	,336	1,000	,042	,051
	Sig. (2-tailed)	,014	,854	,136	.	,858	,827
Results in line with Effort	Correlation	,102	-,630**	,675**	,042	1,000	-,754**
	Sig. (2-tailed)	,659	,002	,001	,858	.	,000
Deserved Better score	Correlation	,028	,609**	-,657**	,051	-,754**	1,000
	Sig. (2-tailed)	,905	,003	,001	,827	,000	.

Table C.6: (N = 21). * = Correlation is significant at the 0.05 level (2-tailed). ** = Correlation is significant at the 0.01 level (2-tailed).

High Delay							
		Degree of Delight	**Degree of Annoyance**	**Evaluation of Quality**	**Positive Influences**	**Results in line with effort**	**Deserved Better Score**
Degree of Delight	Correlation	1,000	-,291	,721**	,343	,583**	-,401
	Sig. (2-tailed)	.	,200	,000	,128	,006	,072
Degree of Annoyance	Correlation	-,291	1,000	-,205	,102	-,251	,365
	Sig. (2-tailed)	,200	.	,372	,661	,273	,103
Evaluation of Quality	Correlation	,721**	-,205	1,000	,288	,519*	-,385
	Sig. (2-tailed)	,000	,372	.	,205	,016	,085
Positive Influences	Correlation	,343	,102	,288	1,000	,007	,083
	Sig. (2-tailed)	,128	,661	,205	.	,977	,722
Results in line with Effort	Correlation	,583**	-,251	,519*	,007	1,000	-,580**
	Sig. (2-tailed)	,006	,273	,016	,977	.	,006
Deserved Better score	Correlation	-,401	,365	-,385	,083	-,580**	1,000
	Sig. (2-tailed)	,072	,103	,085	,722	,006	.

Table C.7: (N = 21). * = Correlation is significant at the 0.05 level (2-tailed). ** = Correlation is significant at the 0.01 level (2-tailed).

Appendix D
Questionnaire

Following is the questionnaire given to the test participants after answering the Kahoot! quiz. The same questionnaire was given after the Longitudinal and the Cross-Sectional tests. All questions were given in English with a Norwegian translation below due to all test participants being Norwegian and to limit possible misunderstandings.

Questionnaire

* Required

1. **Username.** *
   Brukernavn. (Samme brukernavn som under Kahoot! quizzen)

   ....

2. **Age** *
   Alder
   Mark only one oval.

   ○ 19
   ○ 20
   ○ 21
   ○ 22
   ○ 23
   ○ 24
   ○ 25
   ○ 26
   ○ 27
   ○ Older

3. **Gender** *
   Kjønn
   Mark only one oval.

   ○ Female
   ○ Male

4. **Field of study** *
Studieretning
Mark only one oval.

○ Datateknikk
○ Elektronisk systemdesign og innovasjon
○ Energi og miljø
○ Energiforbruk og energiplanlegging
○ Industriell økonomi og teknologiledelse
○ Informatikk
○ Kommunikasjonsteknologi
○ Kybernitikk og robotikk
○ Other

5. **What kind of connection did you use during the Kahoot! quiz?** *
Hvilken netttverkstilkobling benyttet du under Kahoot! quizzen?
Mark only one oval.

○ WiFi (eduroam)
○ WiFi (NTNU)
○ 3G
○ 4G
○ Edge
○ Don't know

6. **What device were you using during the Kahoot! quiz?** *
Hvilken device benyttet du deg av under Kahoot! quizzen?
Mark only one oval.

○ iPhone
○ Android phone
○ Mac
○ PC
○ iPad
○ Android tablet
○ Other

7. **What is your impression of Kahoot!?** *

Hva er din oppfatning av Kahoot!? (Minst ett svar)

Check all that apply.

□ Useful learning tool

□ Nice break/diversion during lecture

□ Disturbing element during lecture

□ I don't see the usefullness of the tool

□ Too time consuming

□ I have no opinion

8. **Indicate to which degree you experienced the following feelings during the test.** *

Indiker i hvilken grad du opplevde følgende følelser under testen. Ett svar per linje.

Mark only one oval per row.

	Not at all	Slightly	Moderately	Fairly	Extremely
Frustrated	○	○	○	○	○
Entertained	○	○	○	○	○
Happy	○	○	○	○	○
Bored	○	○	○	○	○
Annoyed	○	○	○	○	○
Delighted	○	○	○	○	○
Satisfied	○	○	○	○	○
Irritated	○	○	○	○	○
Amused	○	○	○	○	○
Tense	○	○	○	○	○
Concentrated	○	○	○	○	○
Disappointed	○	○	○	○	○
Competitive	○	○	○	○	○
Nervous	○	○	○	○	○

9. **Did you answer all the questions in the Kahoot! quiz?** *
Svarte du på alle spørsmålene i Kahoot! quizzen?
Mark only one oval.

○ Yes
○ No, I ran out of time
○ No, I did not know the answer
○ No, I did not care to answer
○ No, I left the quiz because of techincal difficulties
○ No, I was thrown out of the quiz
○ Other: ....

10. **Indicate to which degree you agree/disagree with the following statements related to technical quality aspects while using Kahoot! (not regarding the content of the quiz)**
Indiker i hvilken grad følgende utsagn passer. Ett svar per linje.
Mark only one oval per row.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
The overall quality of Kahoot! was good	○	○	○	○	○
Kahoot! functioned optimally	○	○	○	○	○
The overall (technical) quality of Kahoot! was NOT acceptable	○	○	○	○	○
The overall technical quality of Kahoot! was as it should be	○	○	○	○	○
I did NOT experience delay between the monitor and my device	○	○	○	○	○

11. **How did you perform compared to your neighbors?** *
Hvordan gjorde du det i forhold til de som sitter ved siden av deg?
Mark only one oval.

○ Better
○ Poorer
○ About the same
○ I don't know

12. **Indicate to which degree you agree/disagree with the following statements.** *

Indiker i hvilken grad følgende utsagn passer. Ett svar per linje.

Mark only one oval per row.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
I peeked/cooperated with my neighbor(s)	○	○	○	○	○
I deserved better scores than I got	○	○	○	○	○
My result was in line with the efforts that I put into the quiz	○	○	○	○	○
I would have had a higher score if the technical quality of the Kahoot! session would have been better	○	○	○	○	○

13. **How did the people sitting close to you influence you while participating in the Kahoot!? Please indicate to which extent you (dis)agree with the following statements: They...** *

Indiker i hvilken grad du er enig med følgende utsagn. Ett svar per linje.

Mark only one oval per row.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Stressed me	○	○	○	○	○
Made me more competetive	○	○	○	○	○
Made me feel good	○	○	○	○	○
Made me laugh	○	○	○	○	○
Made me feel embarrased	○	○	○	○	○
Made me feel scared to answer	○	○	○	○	○
Made me answer quicker than I normally would	○	○	○	○	○
Had no impact at all	○	○	○	○	○
Distracted me	○	○	○	○	○

14. **Comments**

Eventuelle kommentarer

....