How to measure and model QoE for networked games? A case study of World of Warcraft

Sužnjević, Mirko; Skorin-Kapov, Lea; Čereković, Aleksandra; Matijašević, Maja

Noname manuscript No.
(will be inserted by the editor)

How to Measure and Model QoE for Networked Games?

The case study of World of Warcraft

Mirko Suznjevic · Lea Skorin-Kapov · Aleksandra Cerekovic · Maja Matijasevic

Received: date / Accepted: date

Abstract In this paper we investigate methodologies for modelling Quality of Experience (QoE) for networked video games, focusing on Massively Multiplayer Online Role-Playing Games (MMORPGs), and using Blizzard Entertainment's World of Warcraft (WoW) as a case study. In two user studies, involving a total of 104 players, we investigate system, user, and context parameters and evaluate their impact on QoE and related quality features. We also discuss some methodological questions with relation to measuring gaming QoE, which can be used as guidelines for future gaming QoE studies. We further analyse a set of quality metrics "beyond MOS". Having evaluated different modelling techniques, we present and evaluate four linear statistical models and three (non-linear) machine learning models for MMORPG QoE. Finally, we make our datasets available to the research community to foster further analysis and reproducibility of results.

Keywords Quality of experience · networked games · QoE assessment and modeling · MMORPGs · machine learning

1 Introduction

The general term *Quality of Experience* (QoE) has been introduced in early 2000s, as referring to, and emphasising, the subjective aspect of user-perceived service quality, as opposed to the technical performance based *Quality of Service* (QoS) [26]. From the user's perspective, QoE is seen as resulting from "the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in light of the user's personality and current state" [34]. The growth of the global games market value [38], especially in the mobile segment, has created a need for QoE models and assessment methodologies suitable for networked video games, as well as future standards which would (eventually) reach the maturity level of those now in place for "classical" multimedia services, such as conversational voice, streaming video, and audio-visual services [23, 25, 24].

Until fairly recently, QoE models and assessment methodologies for networked video games have received less attention, due to a multitude of factors and inherent complexity involved in understanding the player experience and satisfaction. Playing games is a human-machine interaction based activity, oriented towards achieving game-related outcomes (generally linked to entertainment), as opposed to typical media delivery services such as audiovisual streaming. To this end, video game enjoyment has been described as a *flow experience* [51], and the psychological aspects of gameplay have been studied within the wider field of user experience (UX) [1, 47]. These findings have provided significant insight into gaming QoE assessment, which may be further applied for developing of QoE models and methodologies. Related standards are just emerging, with ongoing efforts at ITU-T aimed at specifying recommendations for QoE assessment methodologies of networked games [36] [1].

Mirko Suznjevic, Lea Skorin-Kapov, Aleksandra Cerekovic, Maja Matijasevic
University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, HR-10000, Zagreb
Tel.: +385-16129-755
Fax: +385-16129-832
E-mail: {mirko.suznjevic, lea.skorin-kapov, maja.matijasevic}@fer.hr E-mail: aleksandra.cerekovic@gmail.com

[1] An interested reader is referred to the ITU-T work item P.GAME in the scope of Study Group 12: Subjective test

Three main contributions of this paper are related to: 1) analysis and modeling of QoE for MMORPGs based on multiple system, user, and context influence parameters, 2) providing insights for QoE assessment methodologies, and 3) an annotated dataset comprising subjective player scores, which is made available[2] to the research community. We further clarify these contributions in detail.

This paper presents the results based on two subjective user studies involving a total of 104 players. As a use case for testing purposes, we take World of Warcraft (WoW), as the most popular subscription-based MMORPG (while subscriber data was still published) in 2014, according to MMOdata [15]. Preliminary results of the first study were published in [46] which this paper expands in many aspects. The methodology of the measurement of both studies is thoroughly described including details about the initial screening survey, laboratory setup, test procedure, and design of the test scenarios. We draw from the generic classification of factors as proposed in [34], and further make reference to the taxonomy of gaming QoE aspects proposed in [37] when deriving our empirical test methodology. We take careful considerations of the methodology applied and discuss four general "research questions" related to measuring and modeling QoE for games, and apply the derived insights when designing our user studies. We hope this example may be valuable to researchers in similar future studies.

Based on the generic classification of influence factors [34] we measure, analyze, and model the impact of multiple system, user, and context influence parameters on QoE. In the first user study, we focused on determining the impact of the degradation of specific parameters on QoE, while in the second study we complement this work by studying the impact on QoE of simultaneous factor degradations. In addition to reporting the Mean Opinion Scores (MOS) for various scenarios testing the impact of different influence factors on QoE, we also examine other metrics beyond MOS (as suggested in [18]), such as the percentages of users judging the gameplay scenario as *Good or Better* or *Poor or Worse*, as well as acceptance measures. Besides collecting ratings of *overall QoE* in user studies, we also collect data about various gaming QoE features (dimensions) and investigate their impact on QoE. A QoE feature has been defined as: "*A perceivable, recognized and nameable characteristic of the individual's experience of a service which contributes to its quality*" [34]. Based on previous work [5], we have identified the following quality features for which we collect subjective player ratings under different test conditions: perceived *immersion*, perceived *interactivity* (defined by the network quality), and perceived *fluidity* (defined by the power of the client device), as well as perceived *challenge* in the game itself.

Based on the performed measurements and analysis, we create QoE models which simultaneously address the impact of multiple factors including system, context, and user factors. In this way, we attempt to derive novel multidimensional QoE models. For this purpose we rely on both linear statistical models and (non-linear) machine learning models.

The remainder of the paper is organized as follows. Section 2 gives an overview of related work addressing the QoE of networked games. In Section 3 we address four research questions related to methodologies for measuring and modeling QoE for games. Section 4 presents the specific test methodology we used in our studies, including an initial screening survey and two laboratory studies. The results, including influence parameters and quality metrics, are analyzed in detail in Section 5, while Section 6 presents the obtained QoE models. Conclusions and the outlook for future work are given in Section 7.

2 Related work on gaming QoE

A "traditional" client-server gaming architecture consist of three main components: the game server(s), the communication network, and the client device(s) (including the player input/output hardware and software). The server sends game specific updates over the network to the client, and the client renders the game content locally. We limit our study in this paper to this type of architecture, and a personal computer (PC) as the gaming platform. We have also tackled cloud gaming, in which the game content is delivered from a server to a client as a video stream (while game controls are sent from the client to the server) in a different study [42].

Many papers studying the user perception of game quality have focused on the impact of "traditional" network QoS parameters (e.g., delay, jitter, loss, throughput) on subjective user-perceived quality [7,11,33,50]. Reported acceptable network impairment thresholds clearly differ for different game genres. Specifically for MMORPGs, acceptable latency values have been shown to be under 120 ms [41]. This finding is strengthened by another study, which showed that MMORPG play session durations decline sharply (i.e., users quite the game) for latencies between 150 ms and 200 ms [4].

methodology for gaming based applications (`https://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=13773`)

[2] The dataset is available upon request, please visit `http://www.fer.unizg.hr/qmanic/data_sets` for details.

A number of QoE studies focusing on MMORPGs have used WoW as a case study, in certain cases with contradictory findings. For example, network delay variation (jitter) has been shown to have a significant effect on QoE [41], while other studies were not able to confirm these findings [45].

The impact of network factors on online game quality has also been studied based on objective performance metrics. The Ubicon Inc. whitepaper [48] proposes an impairment factor metric, which is mapped to an *Online Playability Score* (OPScore), according to the following equations:

$$R = (W_L \cdot L \cdot W_J \cdot J) \cdot (1 + E) \quad (1)$$

$$OPScore = Lookup(R) \quad (2)$$

where W_L and W_J are the weighting factors for latency (L) and jitter (J) respectively, E is the packet loss rate, and Lookup() is a piece-wise linear lookup table mapping the impairment factor R to the *OPScore*. Other approaches involving objective game QoE assessment include the proposal of a Game Outcome Score [12], and using synthetic players to assess gameplay [30].

With regards to studying factors related to the end user device, the impact of frame rate on the performance of players and their perceived playability and quality in First Person Shooter (FPS) games has been investigated [8]. Results have shown that decreased frame rate lowers playability and player performance, especially when frame rate drops under 15 frames per second (fps). Another study focuses on evaluation of frame rate [35] and its impact on perceived end-to-end delay through in-depth examination of interaction of client-side frame rate and server-side tick rate.

A number of studies further focus on the impacts of various types of client devices and novel game delivery architectures on QoE. For example, the authors in [2] asses the impact of system parameters on QoE for mobile games.

Following the QUALINET white paper [34], in recent work Möller *et al.* [37] have proposed a detailed taxonomy of gaming QoE aspects. Aimed at providing a generic evaluation framework, they identify the following three layers: (1) QoE influence factors (related to the user, system, and context); (2) user and system interaction performance aspects; and finally (3) QoE features related to the end user quality perception and judgement processes. The authors classify QoE **influence factors** as being the following:

- *user factors*: experience, playing style, intrinsic motivation, static factors (e.g., age, gender), and dynamic factors (e.g., emotional status);
- *system factors*: game genre, structure, game mechanics and rules, technical set-up (including server, network, network delay, interface software, and device characteristics); and
- *context factors*: physical environment, social context (e.g., relation to other players involved), extrinsic motivation, and service factors (e.g., access restrictions, gaming cost).

The given influence factors impact system and user performance resulting from player interaction with the system, and are finally linked to the following **quality features** (dimensions): interaction quality (also linked to *playability*), playing quality (addressing game learnability and intuitivity), aesthetic aspects, and overall player experience. As previously proposed by Poels *et al.* [39], player experience may be considered in terms of sub-aspects flow, challenge, control, tension, immersion, positive and negative affect. Finally, understanding the psychological complexities of gameplay needs to be taken into account. In particular, it has been proven that different factors impact user motivation for playing online games in general [21] and MMORPGs in particular [53,54].

Focusing on QoE, we conclude that the previous research has mainly focused on evaluation of the impact of system factors. User factors, such as user skill [12, 55] or psychological motivators for playing MMORPGs [54], are just beginning to be understood. As for various context factors, user physical effort and playing context (i.e., interaction with other players) are examined in [49]. Moreover, the cascading impact of network delay is examined in [20], whereby the authors examine the context of a cooperative game and how network delay of some players affects the QoE of players without additional network delay.

Motivated by the fact that QoE is a multidimensional concept, and that the majority of previous research addressing gaming QoE has addressed the impact of limited and isolated influence factors on players' QoE (mostly focusing on network QoS), we aim to simultaneously address the impact of multiple factors including system, context, and user factors. In this work we aim to model QoE for an MMORPG, considering all three defined types of influence factors: system, user, and context factors. Furthermore, we consider QoE in terms of multiple quality features. Stemming from the generic taxonomy proposed by Möller *et al.* [37], we propose a test methodology to study player QoE in the case of MMOPRGs, addressing a chosen number of influence factors and quality features as described further in Section 4.

3 Research questions related to QoE assessment methodologies for games

Ongoing standardization work in the scope of the ITU-T Study Group 12 is focusing on proposing a new recommendation for subjective evaluation of gaming quality referred to as P.GAME [27]. The aims of this Recommendation and open issues are summarized by Möller et al. [36]. We highlight some of the key research questions (RQ) derived from that work, which we specifically address with respect to the studies presented in this paper. Where relevant, we also refer to relevant gaming QoE research found in literature.

RQ1: *Are gaming QoE studies conducted in a laboratory environment "ecologically valid"?*

QoE studies are typically conducted in a controlled lab environment so as to provide a common setting for all test participants, to control various external influence factors, and to enable the repeatability of tests. While the "ecological validity" (i.e., whether the testing environment has significant impact on the testing results) of this approach has been questioned in the context of QoE research [10], in the case of gaming this may be especially pronounced, as "gamers" commonly use their own personalized game setup (referring to screen, interaction devices such as keyboard and headphones, etc.). As discussed in [36], a relatively new paradigm for conducting gaming QoE evaluation tests is that of using crowdsourcing as a means to collect a large amount of ratings from geographically distributed participants [6]. While we focus only on lab studies in this work, we report on the results of a questionnaire-based survey conducted among 105 participants to obtain their opinions on whether or not they feel they can fully enjoy games outside of their usual game setup.

RQ2: *Does previous player experience (with the game genre in general, or with the game in question) have a significant impact on reported subjective QoE scores?*

One of the key challenges when conducting subjective laboratory tests is recruiting test participants based on the target population. Game systems are complex and a certain amount of time is required for players to learn and adapt. Player experience and skill have thus been shown to be potential key influence factors to consider when conducting gaming QoE studies [36]. The player experience can be divided into: general game experience (i.e., how much time the test participant spends weekly playing digital games), experience related to game type (i.e., how much experience does the test participant have related to the type of the game under test), and experience playing the exact game under test. In our studies, we include both novice and experienced players, and perform statistical analysis of results to determine the effect of player experience on QoE. We also note that a number of previous QoE studies do not explicitly consider player experience [3, 41].

RQ3: *Do different contexts and player actions within the same game impact QoE?*

There are hundreds of thousands of games available on the market, characterized by different mechanics, architectural implementations, genres, target player devices, etc. This clearly indicates the need for a wide range of QoE modeling approaches, and imposes challenges when deciding which games to test for QoE model development. The resulting question is: do different action categories within the same game also require different QoE models (or not)? Since our focus is on MMORPGs, and specifically, WoW, for the purposes of test design, we rely on our previous work where have proposed a classification of player actions in WoW [44], based on the number of players involved in the activity, and the distinct attributes of that activity (cooperation level, dynamics, mobility, number of non-player characters (NPCs), combat or no-combat situation, and communication aspect).

RQ4: *In multiplayer scenarios, is there an impact of the skill of other players (and skill difference between the players) on QoE?*

In multiplayer scenarios, multiple players play either with or against each other. Based on the positive psychology theory pioneered by Csikszentmihalyi [9] to enter the state of *flow*, also known as *the zone*, the challenge presented to a player must be in balance with the player skill. While there has been a lot of research regarding balancing player teams in games which are based on players competing against players (or groups of players), from first rating systems for chess, the Elo system [13], to advanced rating systems for teams of players [16], there has been little research on the impact of the characteristics of players within the same team/group playing against computer controlled NPCs. One such study focused on how the network performance experienced by a single player in the team affects the perceived quality of other players [20]. In this work we further focus on this *social context* and investigate the impact of team player skill level on QoE in both homogeneous and heterogeneous teams (in terms of player experience).

We take into account the considerations posed by the four research questions above when designing the user experiments with WoW so as to find out about the user perception of the impact onto QoE for the lab

environment (RQ1), previous player experience (RQ2), context and player actions (RQ3), and skill of other players (RQ4).

4 QoE assessment methodology

The results presented in this paper are derived from two separate studies, referred to in the remainder of the text as *Study 1* and *Study 2*. The methodology applied in both studies was very similar and it is described in detail below. Both studies included two phases:

(1) initial screening survey, involving all participants, aiming to collect data on self-reported skill level and expectations of players, and
(2) lab testing, in which participants took part in three-hour gaming sessions.

The participants in both studies were recruited from Masters level students at the University of Zagreb, Faculty of Electrical Engineering and Computing. In Study 1, a total of 69 participants took part in the initial survey, while 55 of them took part in the lab testing. In Study 2, a total of 35 participants took part in both the initial survey and lab testing. Participant demographics and experience level for the players taking part in the lab testing phase are given in Table 1.

4.1 Initial screening survey

In both studies an initial survey was performed several weeks before the lab testing using an online questionnaire. The primary aim of the survey was to gather data about the participants' previous gaming experience (games in general, multiplayer games, MMORPGs, and specifically WoW as a game which was to be tested). Moreover, participants were asked to provide their opinions regrding the impact of various gaming QoE influence factors. Survey questions for both Study 1 and Study 2 were exactly the same. Participants were asked to rate their skill level with regards to playing online games as being: novice, intermediate, or advanced. We opted for three levels, given that there were participants who had no previous experience in game play. In addition to their self-evaluated experience, we asked for information on how much time they spent playing games (per week: less then an hour, between 1 and 3 hours, between 3 and 10 hours, and more then 10 hours). We used this data as an indicator of the validity of the self-reported experience. Following the survey, obtained data was used as input for designing the lab tests (in terms of grouping players according to skill) and to help in answering the research questions discussed in Section 3.

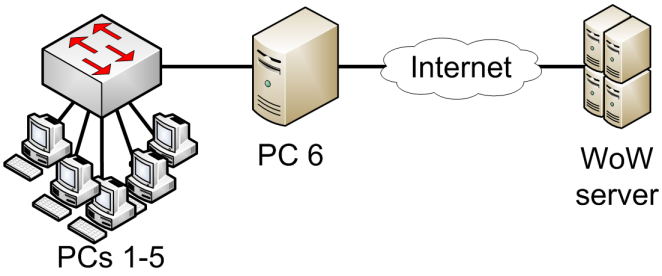


Fig. 1: Laboratory setup

4.2 Laboratory set-up

The lab set-up that was used in both studies is shown in Figure 1. The game being played was WoW version 5.3. (expansion Mists of Pandaria[3]). The game was played on five PCs (marked PC1 through PC5), all running the WoW client on Windows 7 with the following configuration: Dell Optiplex 390, i3@3,3 GHz, 4GB RAM, ATI Radeon HD 6450. The graphical settings on each WoW client were set to *fair*, resulting in the frame rate being stable between 50 and 60 fps. PC 6 was used as a gateway to the Internet used to manipulate network transmission parameters (delay, packet loss). We acknowledge the fact that we could not control the transmission quality of the Internet connection to our lab, but note that our lab is connected to the Internet via a 100 Mbps link, and that round trip time (RTT) between PC 6 and the WoW server was continuously measured and established to be in the range of 30-40 ms during the entire experiment.

4.3 Influence factors

A summary of QoE influence factors (IF) and values considered in our studies is given in Table 2. Drawing from the taxonomy proposed by Möller et al. [37], factors are classified as being either system-, user-, or context-related.

System IFs. As previously mentioned, we consider the impact of four system factors: delay, packet loss, jerkiness, and framerate. We chose these system factors to be able to differentiate the impact of client device computational power and impact of network quality. Moreover, we aim to examine the impact of degradation *severity* versus *frequency*. Network packet loss for TCP based games (such as WoW) translates into bursts of increased network latency (i.e., high severity, but low frequency of the degradation), while adding delay increases network latency for a fixed amount for all pack-

[3] http://us.blizzard.com/en-us/games/mists/

Table 1: Data about the participants in the lab testing

	Study 1			Study 2		
No. of participants	55			35		
Age	21-26 (average 23)			22-28 (average 23)		
Gender	male		female	Male		Female
	37 (67%)		18 (33%)	21 (60%)		14 (40%)
Experience	novice	intermediate	experienced	novice	intermediate	experienced
	14 (25%)	23 (42%)	18 (33%)	11 (31%)	17 (49%)	7 (20%)

Table 2: Influence factors (IFs) and corresponding values in both studies

Factor	Values		IF category
	Study 1	Study 2	
Delay (RTT)	0 ms, 200 ms, 400 ms	0 ms, 150 ms, 200 ms, 300 ms, 400 ms	System
Packet loss (probability)	0, 0.05, 0.1	0, 0.02, 0.04, 0.06, 0.08	System
Jerkiness	0, 0.033 (1 s every 30 s), 0.133 (2 s every 15 s)	0, 0.033 (1 s every 30 s), 0.066 (1 s every 15 s), 0.1 (2 s every 20 s), 0.133 (2 s every 15 s)	System
Frame rate (FPS)	60, 25, 15	60, 45, 35, 25, 15	System
Game genre	MMORPG		System
Game	World of Warcraft		System
Transport protocol	TCP		System
Age	21 - 26 (average 23)	22-28 (average 23)	User
Gender	male, female		User
Player experience	novice, intermediate, experienced		User
Player experience in MMORPGs	Yes, No		User
Action category	Questing, Dungeons		Context
Social context	homogeneous (2 novice, 3 intermediate, 2 experienced groups) and mixed (4 groups)	mixed (7 groups)(	Context
Physical environment	University lab		Context
Extrinsic motivation	Obtaining credits for the course		Context

ets (i.e., low severity, but high frequency of the degradation). The same applies for frame rate and jerkiness as related to the client device power. Jerkiness translates into short bursts of very low frame rate (i.e., high severity but low frequency). Further explanations are given as follows:

- *Delay*: Delay was introduced on PC 6 using an Integrated Multiprotocol Network Emulator/Simulator tool (IMUNES) [56].
- *Packet loss*: Packet loss was controlled through a FreeBSD firewall, and set to an equal value for both (uplink and downlink) directions.
- *Jerkiness*: We introduced the jerkiness effect using a special script (run on PCs 1 – 5) which created a large number of processor jobs, effectively shortly freezing the game.
- *Frame rate*: Frame rate was limited using the graphical settings on each WoW client (PCs 1-5).

User IFs. With regards to user factors, gender, age, and experience were recorded. We make an assumption that more experienced players have a higher skill level. This assumption was partly confirmed through initial screening study as shown later in Section 5. Self-reported player experience was taken into account when forming test groups for multi-player test scenarios.

Context IFs. Further regarding context factors, we manipulated the *social context* by forming different types of player groups, where a *group* corresponds to five players simultaneously taking part in the lab testing. The group composition was either homogeneous (meaning all group members were of the same skill level), or mixed (meaning the five players involved were of mixed skill levels, including novice, intermediate, and expert). The group composition was an important factor as players were also requested to take part in collaborative group efforts. This is related to the tasks the players were requested to take part in, whereby we refer to these tasks as action categories.

A total of 34 IF combinations, referred to as "scenarios" were tested in Study 1, and 142 scenarios we tested in Study 2.

4.4 Measured parameters: QoE, quality features, and other objective metrics

Following the identified IFs, we summarize the parameters which we measured in Table 3. Subjective ratings were collected using a standardized 5-point Absolute Category Rating (ACR) scale to obtain Mean Opinion Scores (MOS) for QoE, and the following additional quality features: *perceived immersion*, *perceived responsiveness* (in terms of the system reacting to user commands in real-time), and *perceived fluidity* (referring to the perception of the smoothness in the rendering of the virtual scene). Following a given test scenario, players were also requested to rate the *level of challenge* they experienced in the given scenario (5-point scale, from 1 being "very simple" to 5 being "very challenging"). This parameter is not a vector-model feature in which "higher is better", but an ideal point feature as defined in [34]. Making once again reference to the taxonomy proposed in [37], we can consider this metric as being related to the user performance in terms of perceptual effort.

Finally, two objective metrics we collected include: (1) overall game play success "score" achieved by a given player and (2) the number of "disruptive" events (i.e., player deaths, player getting lost). Score corresponds to the level reached while questing, and the number of bosses (i.e., very strong enemies, with only a few of them in each dungeon) slain while passing through dungeons). It should be noted that a question related to a player's willingness to continue playing (i.e., question: "Would you continue playing under these conditions?") was added in Study 2 based on the feedback obtained after Study 1 results were published.

When conducting QoE studies, the reporting of only MOS values has previously been criticized, as MOS values average out variations between users [43, 52]. Thus, considering different stakeholder perspective, additional QoE-related metrics derived from subjective studies may be of interest when managing and optimizing service quality, such as score distributions, or the percentages of satisfied and dissatisfied users [18]. Such metrics can provide a clearer picture of end user satisfaction and potential causes for user churn [31]. Therefore, after obtaining QoE score for experiments we have calculated additional sets of metrics: *Good or Better* (GoB) ratio, *Poor or Worse* (PoW) ratio, *acceptance measures*, and *Standard deviation of Opinion Scores* (SOS).

4.5 Test procedure in the lab

The test procedure for both studies is illustrated in Figure 2. After the set-up, which included introductory remarks, learning the game basics and playing the best and the worst case scenarios, each test scenario (referring to a given combination of tested IFs) lasted for 5 minutes. The scenarios were set-up and coordinated by a test administrator, who requested players to pause after 5 minutes of game play, and to provide subjective ratings of overall QoE, immersion, fluidity, responsiveness, and perceived challenge. Following these ratings, players continued to play the game (at the point in the game where they had left off), but under the (changed) conditions of a new scenario. In all tests, an experienced WoW player "consultant" was available on site to provide advice to inexperienced players (if and when needed, e.g., if someone got disconnected from the server, or needed a quick tip on how to proceed or perform a task).

In both studies, we opted for first testing the Questing action category, followed by Dungeons. This decision follows also the game design rules as new players first have to perform simple tasks and fight simple NPCs during Questing, and only later on enter a team-based combat in dungeons. Each player played the same class/race combination in questing and dungeons, so inexperienced players could more easily familiarize themselves with the game play during Questing, in preparation for more complex activities corresponding to Dungeons.

Table 3: Measured parameters

Name	Metrics
Overall QoE	5 pt. ACR scale (1-bad, 5-excellent)
Perceived Immersion	5 pt. ACR scale (1-bad, 5-excellent)
Perceived Responsiveness	5 pt. ACR scale (1-bad, 5-excellent)
Perceived Fluidity	5 pt. ACR scale (1-bad, 5-excellent)
Perceived Challenge	5 pt. scale (1 - very simple, 5 - very challenging)
Score	Level reached (Questing), bosses slain (Dungeons)
Disruptive events	Death count, players getting lost, disconnects, etc.
Service acceptability	Willingness to continue playing (Yes/No)

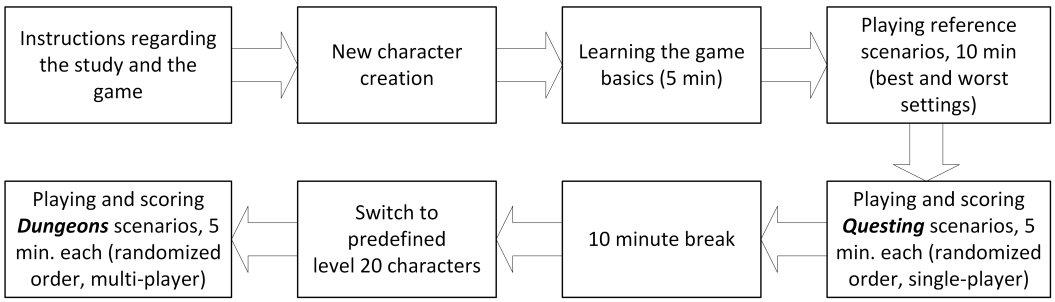


Fig. 2: Test procedure used in both studies for lab testing

4.6 Scenario design

Due to the large number of IF manipulations and corresponding test configurations that we wanted to evaluate, the studies were organized as follows. The goal of Study 1 was to quantify the influence of each specific factor, i.e., we only degraded *one system factor at a time*, with only several testing scenarios involving the degradation of multiple system IFs. Each tested system IF was set to three values corresponding to: *no degradation*, *noticeable degradation*, and *severe degradation*. The values of the manipulated factors were chosen based on empirical testing with the goal of finding two limits: noticeable degradation and severe degradation. Two players played the game, with each listed parameter being slowly degraded. The first value was chosen when the players reported they first noticed the degradation, while the second value was chosen at the point where players reported degradation as severe. Only the values for latency were taken from previous studies of QoE for MMORPGs [41]. In Study 1, we first focused on test scenarios in which only one system factor was degraded while other factors had nominal values (e.g., frame rate of 60 FPS, no packet loss, 200 ms of latency, and no jerkiness). Sixteen (16) such scenarios were tested (8 for questing and 8 for dungeons) **by all 55 players**. Scenarios were randomized to avoid ordering effects. We then further designed 16 additional test scenarios that involved multiple factor degradations. However, given that additionally testing all of these scenarios would have resulted in overly lengthy test sessions for participants, we assigned to each participant group an additional 4 out of 16 scenarios (chosen differently for each group) with multiple factor degradations. Thus, each group tested 16 (single-factor degradation) + 4 (multi-factor degradation) + 2 (best and worst case) scenarios. Best case scenario (i.e., none of the parameters degraded) and worst case scenario (i.e., all of the parameters degraded to the maximum degradation level tested in the study) were always shown on the start of the study, as it is common practice in QoE studies.

Following Study 1, the goal of Study 2 was to investigate how the **simultaneous degradation of multiple influence factors** affects QoE. Therefore, in Study 2, we further tested multiple simultaneous factor degradations with the aim being to explore their interactions and model QoE in terms of multiple factors. IFs were manipulated between 5 values (please refer to Table 2) so as to obtain more accurate models. After a certain amount of filtering, a total of 142 possible test scenarios were identified. Subsets of 20 scenarios were assigned to each group for testing. In the end, we made sure that each of the 142 scenarios was tested by 10 or more participants[4].

[4] Further details on the scenario design are provided together with the dataset, please visit `http://www.fer.unizg.hr/qmanic/data_sets`

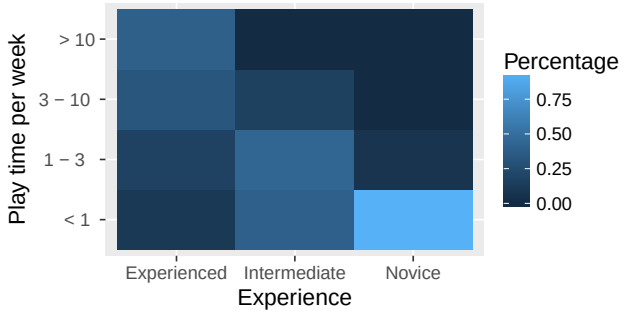


Fig. 3: Relation between players' self reported experience and self reported time played (hours per week)

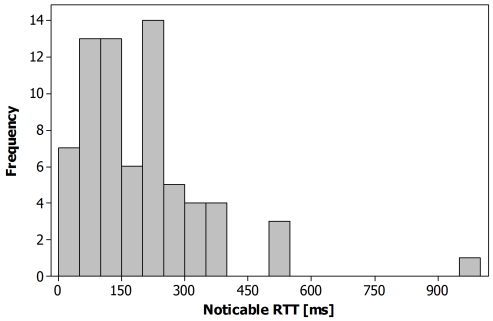


Fig. 4: Players' expectations regarding noticeable RTT values

5 Analysis of obtained results

5.1 Initial screening survey

In this section we briefly present an analysis of the results of both screening surveys, conducted prior to Study 1 and Study 2. Given that the survey was the same in both cases, for the most part we aggregate scores (across 104 participants) to draw interesting general conclusions with respect to player opinions on IFs, and correlations between self-reported experience and play time.

Figure 3 presents a heat map indicating for specific self-reported skill ratings, the amount of time spent playing per week as reported by participants. This plot confirms that players' self-reported experience ratings are in line with game play time, as none of the novice players reported playing more than 1-3 hours a week, and none of the intermediate players spent more than 10 hours a week playing. We can also see that all players who claimed to be experienced in fact play over 3 hours per week, and in most case more than 10 hours per week. We therefore conclude that our self-reported experience ratings can be treated as valid, and can be used as a differentiator when looking at reported QoE.

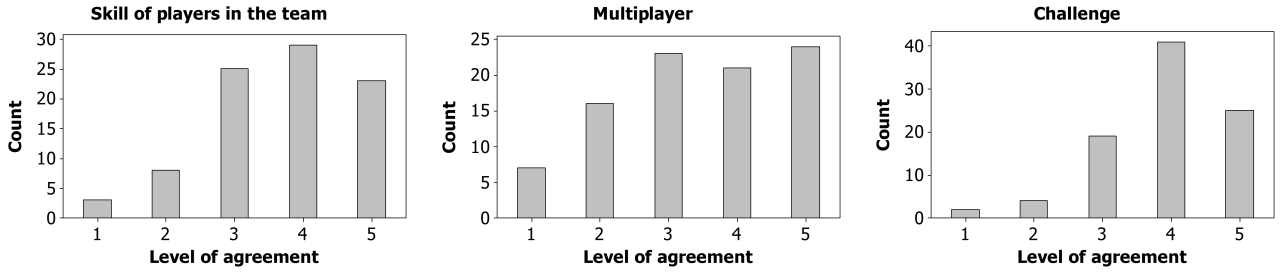
We further asked participants about their expectations in terms of what minimum network delay (in terms of round trip time) they would consider as being noticeable and a cause of game play degradation for online role-playing games. The histogram in Figure 4 shows that the majority of players stated values below 200 ms, and only 3 players stated values larger then 400 ms. Therefore, for our laboratory tests we opted to test values of RTT up to 400 ms based on these results, but also on previous work done on WoW [41].

In terms of context factors, we were interested in user opinions with respect to whether or not they expect the following factors to have an impact on player QoE: 1) the skill of other players in their groups 2) the multiplayer aspect of gameplay, and 3) perceived challenge of the task at hand. The second factor was tested to see whether players expect to enjoy a cooperative multiplayer activity (Dungeons) more than a single player activity within the game (Questing). Results are presented in Figure 5 and show that participants expect the skill of other players in their team (in multiplayer scenarios) to have an impact on QoE, while there was only a slight overall preference for multiplayer scenarios. Based on these results we can hypothesize that there will be an influence of team composition (in terms of skill) on QoE. Also, we found that players mostly concur with the statement that the level of challenge has an impact on QoE, which is another reason why we tested both the Questing action category (simpler tasks) and the Dungeons category (more complex and challenging tasks).

To address RQ1 (outlined in Section 3) of whether or not it is ecologically valid to conduct gaming QoE tests in a lab, we considered the major components which differ between a lab environment and a "normal" gameplay environment. We asked the participants to what extent they agree (1 – do not agree, 5 – fully agree) with the following statements regarding their gaming preferences:

- I can fully enjoy games only in my usual gaming space and on my usual gaming configuration.
- I can fully enjoy games not depending on the time of the day.
- I can fully enjoy games regardless of input/output devices such as mouse, keyboard, controller and monitor (under the condition that they are functioning properly and that all of their performance parameters are acceptable).

Distributions of responses to these questions are depicted in Figure 6. Regarding the space used for gaming, there is no consensus among the players as they are more or less evenly distributed across the levels of



(a) Skill of team members (b) Multiplayer (c) Challenge

Fig. 5: Level of agreement that factors related to social context have an impact on QoE (1 – do not agree, 5 – fully agree)

agreement, with only slightly more players not concurring with the statement that they can fully enjoy games only in their own space. Participants for the most part felt they could enjoy gameplay regardless of the time of day. The same can be applied to input and output devices, as only a small portion of the players did not concur with the given statement. Based on these results, we can conclude that laboratory testing probably affects the player's QoE to some degree, mostly due to the space (physical environment) in which the testing is conducted. While clearly a good practice would be to test players in their customary real world environments as unobtrusively as possible, employing such an approach depends on the research question being addressed, and generally poses the problem of controlling test scenarios and applying the same conditions to multiple users.

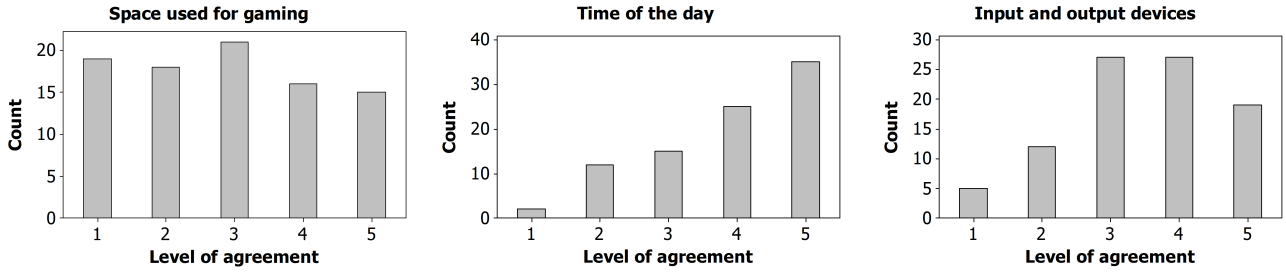
5.2 Influence factor analysis

Following the analysis of the screening survey, in this section we focus on analyzing the results obtained from our laboratory studies. The analysis presented in this section mostly uses 18 scenarios from Study 1 (i.e., one scenario with no system parameters degraded, another scenario with all parameters degraded, and the remaining 16 scenarios in which only one of the parameters was degraded while others were kept constant). We choose these scenarios as they were tested by all 55 participants taking part in Study 1, and can thus be used to draw reliable conclusions. We present not only MOS values, but also distributions of scores, as well as other metrics as suggested in [18] that provide better insight into the user ratings. Where relevant, we also refer to results obtained in Study 2.

First step was to analyse and filter the data to remove any unreliable user ratings. This procedure is often needed in crowdsourcing QoE studies to remove the participants which give their scores arbitrarily just to finish with the survey. In previous work [19, 40] some of the methods for filtering such data are listed. Here we use a simple method from [40] in which the sample correlation coefficient between the average user rating of a user and the global average rating is used to identify unreliable users. The user ratings are averaged for the same test conditions. A user is rejected, if the correlation coefficient is below a certain threshold, e.g. 0.25. To evaluate performance of individual participants and to remove any subjects with invalid responses, we followed this procedure and calculated the correlation coefficient between individual subject ratings and MOS values for all experiments. The procedure was as follows. We first calculated the MOS ratings for 16 scenarios (8 Questing scenarios and 8 Dungeons scenarios with only one system parameter degraded). We then calculated the Pearson and Spearman correlation coefficient between the MOS of a scenario and the rating given by a particular user for that scenario for all users in the dataset across 16 scenarios. The results of the preliminary analysis and filtering are: (1) User 5 scores for Questing were removed as all scores were 1 (probably due to an error); (2) Depending on the obtained values we pruned the dataset by removing User 52, who had both correlation coefficients below 0.25, as it is illustrated in Figure 7 (user 52 is marked with a red circle). It can be noted that the dispersion of coefficients is quite high amongst the users with average value being 0.65 for Pearson's and 0.61 for Spearman's correlation coefficient.

5.3 User factors

We report findings regarding the following user factors: gender, previous MMORPG experience, and previous gaming experience in general. We did not look at the effect of age, as all test participants could be placed in the same age group of young adults.



(a) Gameplay space (b) Time of the day (c) Input and output devices

Fig. 6: Level of agreement that context related to playing environment have an impact on QoE (1 – do not agree, 5 – fully agree)

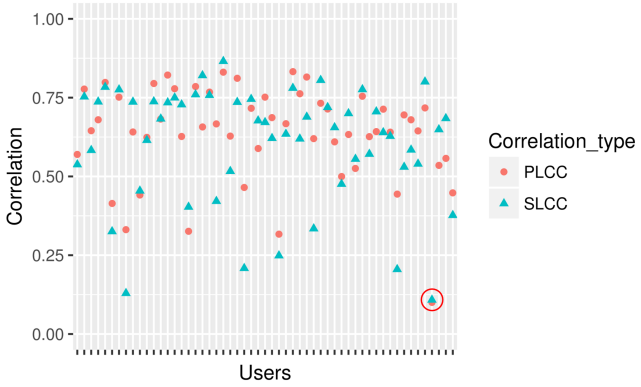
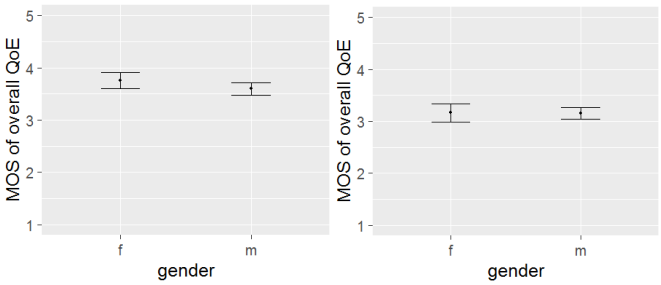


Fig. 7: Pearson and Spearman correlation coefficients per user



(a) Questing scenarios (b) Dungeons scenarios

Fig. 8: QoE MOS per gender with 95% confidence intervals

5.3.1 Gender

With respect to gender, we perform analysis separately for the Questing and Dungeons scenarios. Beside discovering potential links between QoE perception and gender, we were also interested in differences between results in the two scenarios, considering that the context and other user parameters might influence the reported QoE. Figure 8 presents means with 95% confidence intervals (CI) for QoE separately for Questing (Figure 8a) and for Dungeons (Figure 8b) for male and female players. The overlapping is very evident for the Dungeons scenario, while in Questing it is also present, but to a lesser degree. Questing scenarios seem a better way to test the impact of gender, because in that case we can eliminate the impact of other participants in the group. To further evaluate the data regarding Questing, we did a type 3 ANOVA tests (for unequal sample sizes) for all system variables and gender. The p-value for gender was 0.05112 which is considered statistically insignificant. We also did a pairwise t-test which yielded a score of 0.13. When repeating these procedures for Dungeons and taking into account mixed and homogenous groups separately, we obtained clear results from both ANOVA and paired t-test that **gender has no statistically significant impact on QoE for the Dungeons category**.

When looking at gender-based QoE responses across single play, mixed groups, and homogeneous groups in Figure 9 we see that there is no clear difference between any of the groups except for the case of females in mixed(f.m) and females in homogeneous groups (f.h). Additionally, to test whether there is a correlation with other user parameters, we perform a Chi-squared test of independence between gender and user's self-reported general gaming experience. The test statistics show with very small p-value that these two variables are not independent. Therefore, it is important to notice that differences, even though they might not be statistically significant, might be attributed to the difference between playing experiences and not gender. Differences between homogeneous and mixed groups phenomenon can be attributed to the majority of female players being novice so when playing in mixed groups they usually played with players better than themselves, but when playing in homogeneous groups they usually played with novice players.

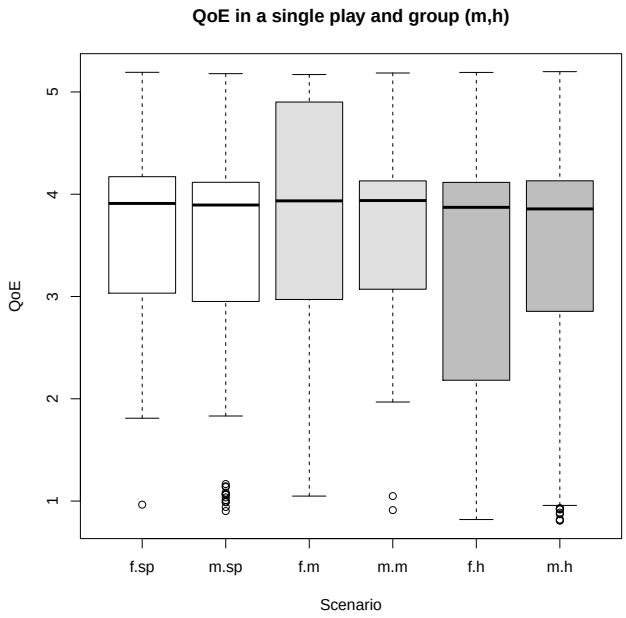


Fig. 9: QoE for female (f) and male (m) players during single play scenarios (sp), mixed group scenarios (m) and homogeneous group scenarios (h)

Therefore, based on all presented data we can conclude that **gender did not have a significant impact on game QoE** for Study 1.

5.3.2 Previous experience in games

Focusing on RQ2 (outlined in Section 3, we investigated QoE collected from three groups based on previous experience in games: *novice*, *intermediate*, and *experienced*. Additionally we looked at experience regarding playing MMORPGs divided into two categories: *some previous experience* and *no previous experience*.

As in the previous subsection, we present the analysis for the scenarios for which all participants had the same system parameters (i.e., scenarios 1-18 in Study 1) and divide that data based on whether the gameplay was Questing or Dungeons. Questing data may be taken as the most relevant for extraction of parameters related to previous experience, as other context parameters are not affecting the data. Figure 10 shows the means with 95% confidence intervals per experience category and per action category. For Questing, a declining trend, negatively related to the level of experience, can be observed. A statistically significant difference between QoE values was observed under the same system parameters in the Questing scenarios between novice and experienced players. When running ANOVA analysis with previous experience level and other system parameters for Questing, previous experience is identified as a significant factor (p-value <0.005), while for Dungeons this was not the case (p-value $= 0.645$).

To further inspect this conflicting issue and validate the results, we look at the data from Study 2. We once again note that Study 2 used a between-subjects design, and that not all participants rated the same conditions. In Study 2, skill level had a significant effect according to ANOVA analysis for both Questing and Dungeons. Therefore, we conclude that **previous experience does have a significant impact on game QoE** for our tests studies.

In addition to general previous experience we also divided participants into those who had some experience with the MMORPG genre and those without any experience with the MMORPG genre. Figure 10c shows the MOS scores for two groups with and without any MMORPG experience. As it can be seen, there is an overlap of confidence intervals, and t-tests showed no significant difference at any level (both Questing and Dungeons scenarios considered).

To conclude, data shows that **general game experience is a significant factor when assessing game QoE**, as we found significant statistical differences between experienced and novice groups. Therefore, it should be left to UX community and game designers to choose which group to target when conducting future studies. For example, designers could aim to increase QoE of novice players so as to have better retention of new players.

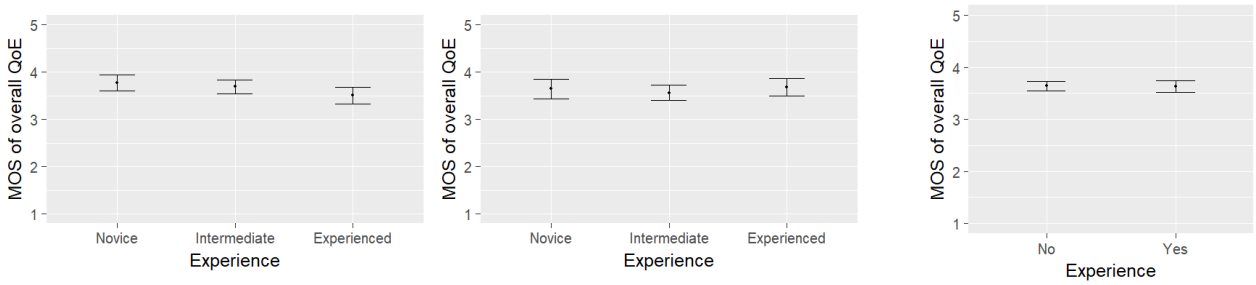
5.4 Context factors

Three context IFs were tested, namely: action category played, social context (i.e., the composition of the group in terms of players' experience), and in-game events which we dubbed "disruptive events". Disruptive events comprised the events of a character dying, getting lost, game crashes, etc.

5.4.1 Action categories

When considering research question RQ3, i.e., the impact of action category, we note that Questing is an action category in which players perform relatively simple tasks and usually do not require high player skill (especially true for the starting quests performed in our scenarios). On the other hand, Dungeons are a much more demanding action category which requires cooperation between players, enemies are much more dangerous, and players can easily be killed.

We compare the results for Questing scenarios 3-10 and Dungeons scenarios 11-18 from Study 1 and present



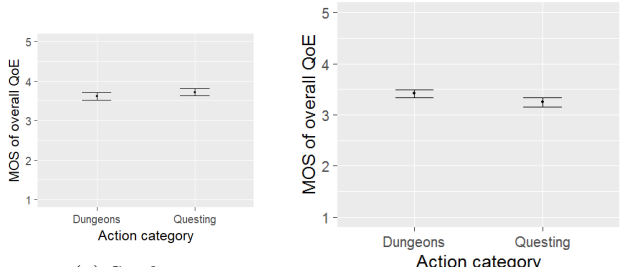
(a) Questing scenarios (b) Dungeons scenarios (c) Previous MMORPGs exp.

Fig. 10: QoE MOS per (MMORPG-reported) experience level with 95% CIs

the data in Figure 11a. It is evident that the error bars are overlapping. We also perform a t-test between these two groups of scenarios which confirms that there is **no statistically significant difference between the Questing and Dungeons experiments**.

QoE scores for each testing scenario in Study 1 are depicted in Figure 12. We found a difference between Questing and Dungeons only in the case of very high packet loss, which may be attributed to a high latency spike resulting from loss and leading to unresponsiveness of the controls for a longer period of time (which can result in "group wipe," i.e., all players in the group dying).

We also perform the same procedure in Study 2, although it should be noted that in Study 2 the conditions differed between players, but the conditions within the groups for Questing and Dungeons were the same. Therefore it is valid to test the difference in Questing and Dungeons even in Study 2. The results for Study 2 differ from those obtained in Study 1. Figure 11b indicates that there is a significant difference, as well as the t-test performed on the data. What is interesting is that in Study 1, Questing has higher scores on average, while in Study 2 Dungeons have higher scores on average. Based on the results of both studies **no definite conclusion regarding impact of action category as a context parameter on the perceived QoE can be drawn**.



(a) Study 1 (b) Study 2

Fig. 11: The impact of action category on QoE

5.4.2 Social context

Related to research question RQ4, the *social context* refers to the composition of the 5-member group that a player was involved in. We have compared how the reported experience of a player's teammates affects their perceived QoE on scenarios 11-18 in Study 1 (i.e., Dungeons scenarios performed by all test subjects). From 11 groups, 2 groups were composed of only novice players, 2 groups from only experienced players, and 3 groups from only intermediate players. The remaining 4 groups were "mixed" groups which always comprised at least

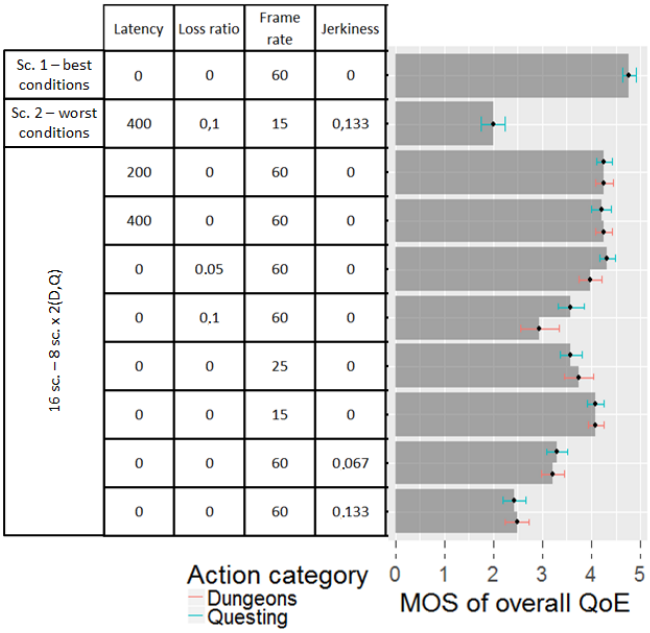


Fig. 12: The QoE scores of best case scenario, worst case scenario, and 16 scenarios with only one parameter degraded: avg. values and 95% CI.

one novice player, one experienced player and two intermediate players. Figure 13a indicates that there is a difference between the tested groups, and, surprisingly, average QoE was higher in mixed groups. We additionally run a t-test which confirms that there is a difference between two groups (p-value = 0.031), and an ANOVA test with system parameters and group composition which confirms that group composition has an impact (p-value=0.011).

We perform further analysis taking into account the relationship of players' previous gaming experience and group composition. In Figure 13b we compare average QoE scores for players in "homogeneous" groups with the scores of players of the same skill level but in mixed groups. Interestingly, results have shown that both novice and intermediate players improve their QoE when playing in mixed groups, most likely due to improved group game performance resulting from the involvement of more experienced players. The majority of the difference between mixed and homogeneous players comes from the novice player group. On the other hand, experienced players reported slightly lower average QoE when novice and intermediate players were involved. We presume that this may be because players of higher skill notice that their team members are playing worse then expected. The highest observed difference between mean values is for novice players (0.3). Nevertheless, the confidence intervals are overlapping for each of the categories and the t-tests show no significant difference between scores except for the intermediate skill level in which there is a statistically significant difference between the mixed and homogeneous groups.

Based on the observed results, we can conclude that **social context in terms of the skill of team members does have a significant effect on QoE**, but it should be noted that this effect significantly depends on which players are taken into account. In particular, for only experienced players, data did not support this finding.

5.5 System factors

The results of the QoE scores obtained across the first 18 scenarios of Study 1 are shown in Figure 12. All of the degradation combinations have been done for both the Questing (Q) and Dungeons (D) action categories, except for the first two (reference) scenarios.

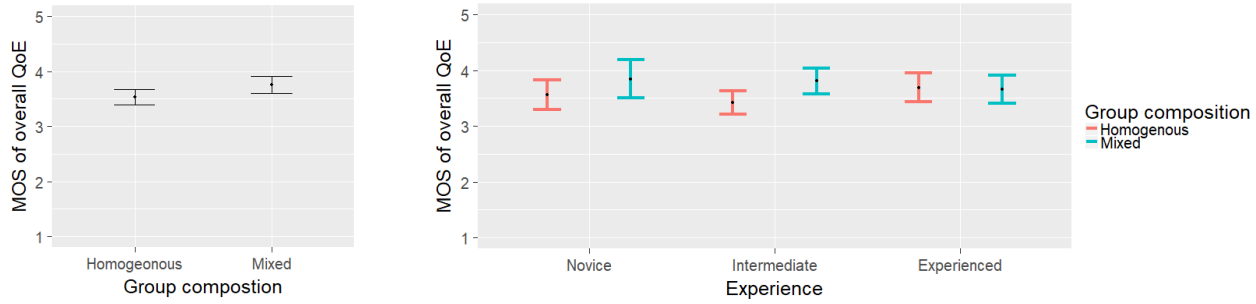
Results indicate that introducing what we have referred to as jerkiness (or freezing) is the factor which has the strongest impact on QoE, resulting in an average QoE score of 2.4, which is slightly more then the average of 2.0 reported in scenario 1 which is the worst case. The second most influential factor proved to be packet loss, followed by frame rate degradation, and in the end latency. While it has been reported in literature that some games, e.g., Quake 3 can tolerate up to 30% packet loss rates (with MOS scores over 4), for other games, such as Halo, loss rates of 2% already resulted in MOS scores dropping below 4 [55]. Our studies have shown that for WoW (Dungeons action category), packet loss of 10% resulted in average scores of 2.56, while for 5% packet loss average MOS score was 3.88. The impact of loss may be attributed to the TCP transport protocol being used. Another indicator of how packet loss affects the gameplay is based on the in-game latency indicator[5]. Introducing 1% packet loss (through PC6 in laboratory setup) resulted in reported latency estimations of hundreds of milliseconds (due to TCP retransmission mechanisms), although no delays were actually present on the transmission link.

What we found peculiar was the issue of latency, whereby we introduced latencies of 200 ms and 400 ms, which resulted in RTTs being up to 240 ms and 440 ms. Contrary to previous measurements and QoE models in which introducing this much latency resulted in significant lowering of the reported QoE, e.g., a MOS of 2.6 for 400 ms latency reported in [41], the latency degradation proved to be barely noticeable to our test players. This phenomenon might be attributed to the degradation of other parameters which resulted in more easily observable degradation (e.g., jerkiness), in-game mechanisms for hiding/combating latency such as ability queueing, or unfamiliarity of tested player group with the game under test (WoW). To shed more light on this issue we aim to further test this finding in future experiments.

We performed an ANOVA for different forms of the dataset. We tested for scenarios in which only one parameter was changed separately for Questing (i.e., scenarios 3-10) and for Dungeons (i.e., scenarios 11-18), and we also tested for all scenarios (i.e., scenarios 1-34). Results yielded that latency is not a relevant factor when considered separately for Questing and for Dungeons, but when performed on the whole dataset latency was a significant factor, but only with a p-value of 0.005. Other system factors were found significant with very small p-values in any case.

Further, we wanted to inspect how combinations of different simultaneous degradation parameters affect the QoE. We inspected this through an additional 16 scenarios (i.e., scenarios 19 to 34) which were only performed in the Dungeons action category (referring to Study 1). In each of these scenarios all parameters were degraded (i.e., frame rate, jerkiness, latency, and loss)

[5] Hovering over a computer icon in the main menu of WoW results in a pop-up window showing the estimated latency by the WoW client.



(a) Per group composition

(b) Per composition and experience level

Fig. 13: The impact of group composition on QoE

Table 4: Results of scenarios with multiple degradation

	Loss rate	0.05		0.1	
Jerkiness	Delay / Frame rate	200ms	400ms	200ms	400ms
0.067	25	2.7	2.8	1.9	2.19
	15	2.73	3.1	1.93	1.6
0.133	25	1.9	2.53	1.5	1.7
	15	3.25	2	1.5	1.7

= 5

= 1

to two different levels. Each of the scenarios was performed by two different player groups, and for some even three players. The results are presented in Table 4. In the table, darker fields correspond to lower QoE score. Results of these scenarios confirm the findings in the first set of scenarios that jerkiness and loss rate are the factors that affected QoE the most. The highest QoE degradations (i.e., the lowest scores) are noted when both of these parameters are severely degraded (level 2 degradation). It is interesting that the lowest reported values are not reported in cases involving increased latency.

5.6 Additional QoE metrics beyond MOS

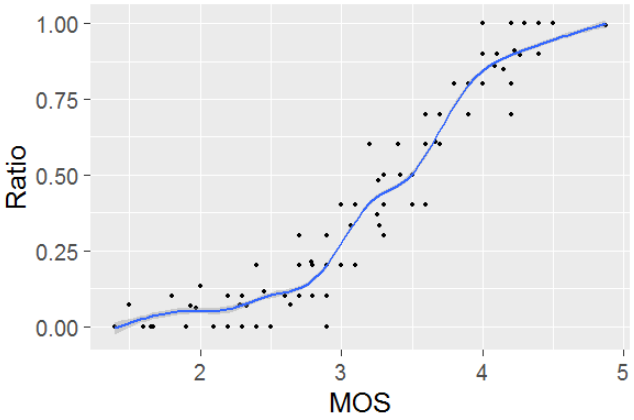
In addition to the QoE MOS values we have reported so far, we go beyond these metrics and look at the percentages of users judging the gameplay scenario as *Good or Better* (%GoB, referring to the ratio of users scoring 4 or 5 on a 5 pt. ACR scale) or *Poor or Worse* (%PoW, referring to the ratio of users scoring 1 or 2 on a 5 pt. ACR scale), as well as acceptance measures (an interested reader is referred to [18] for an in-depth discussion of QoE metrics beyond MOS). GoB and PoW measures are calculated based on the results of both Study 1 and Study 2, while acceptance is calculated based only on the results of Study 2.

Figure 14a shows GoB scores for all of the experiments. It can be seen that for very small MOS differences from 3.5 to 3, the ratio of users judging the service as good or better falls by approx. 25%. According to our data, to maintain at least 80% of users satisfied (rating 4 or 5), we need to consider those scenarios where MOS was calculated at or above 4. Figure 14b further shows PoW scores for all of the experiments. It can be seen that already at MOS of 3.2, 20% of users are dissatisfied with the service, potentially resulting in users abandoning the service in real-world scenarios.

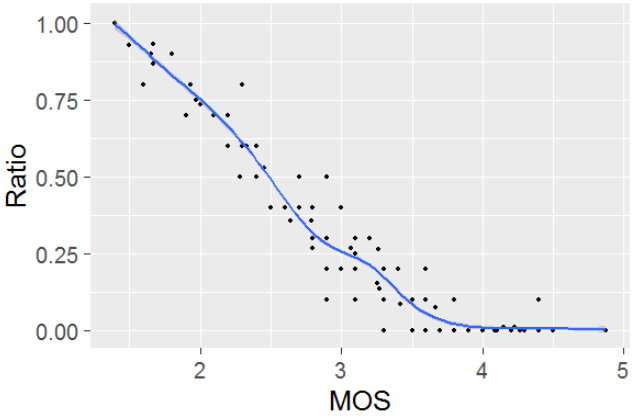
The relationship between MOS and acceptability level (i.e., willingness of players to continue using the service under the given conditions) can be seen in Figure 15a (only experiments from Study 2 are depicted). It can be seen that acceptability of the service is at 90% or 100% for MOS scores above 4, leading to the conclusion that players would continue playing under those corresponding test conditions. The majority of scenarios in Study 2 were tested by only 10 players, therefore acceptability levels are shown in steps of 10%. A significant drop in acceptability occurs when MOS drops below 3.5, at which level approx. 40% of the users indicated they would not continue playing. It is noticeable that the acceptability ratio rises and falls faster then MOS (i.e., for higher values of MOS, acceptability is on average higher, while for lower MOS values acceptability is lower), so even at MOS values slightly below 3, the acceptability ratio falls to 30% and less. This leads us to the conclusion that players in general have strict quality requirements that need to be met for them to continue playing the game.

As related to user score distributions, we finally look at the diversity of user ratings. We refer to the SOS (Standard deviation of Opinion Scores) hypothesis as defined in [17], which relates SOS values to MOS values, by postulating a square relationship between the variance SOS and MOS parametrized by the SOS parameter a. Taken from [17], for a 5-point rating scale the SOS hypothesis is described as follows:

$$SOS(x) = \sqrt{a} \cdot \sqrt{-x^2 + 6x - 5} \quad (3)$$

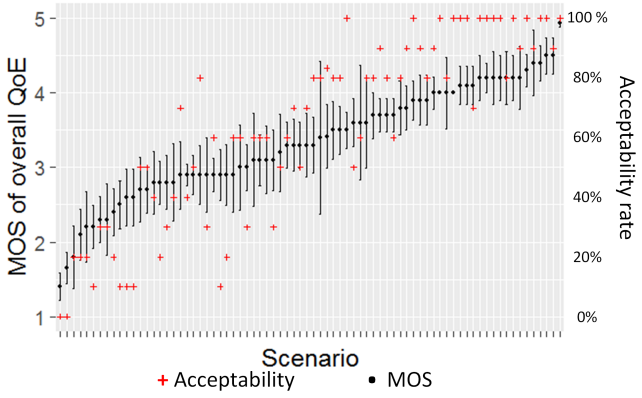


(a) Gameplay scored as good or better

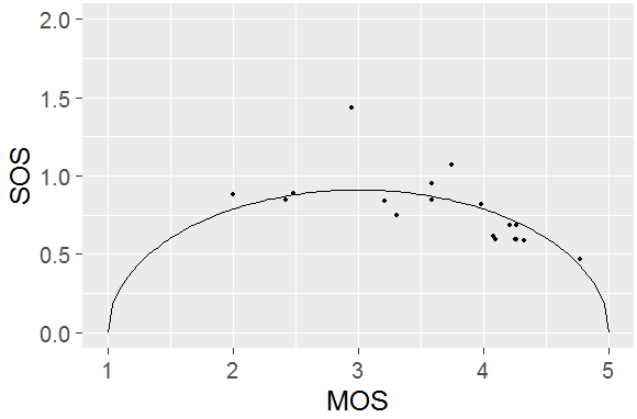


(b) Gameplay scored poor or worse

Fig. 14: Ratio of scores per experiment



(a) Acceptability ratios and MOS for QoE (with 95% CIs) per experiment (Study 2)



(b) Distribution of Standard deviation of Opinion Scores (SOS)

Fig. 15: Acceptability and SOS metrics

One SOS parameter a is computed for an entire subjective test campaign. The lower the a, the less diverse are the user ratings for a given test condition, i.e., subjects are presumably more confident in their ratings. As it can be seen in Figure 15b, the values of $SOS(x)$ for all test scenarios are pretty low, except for one scenario where the value reaches almost 1.5 and corresponding to a MOS of approx 3. We obtained a value of $a = 0.20696$ by curve fitting and minimization of least squared error with residual standard error being 0.1611. For comparison, our obtained a value is lower than for example values obtained for cloud gaming for different games (fast paced $a = 0.2718$, medium paced $a = 0.3287$, and slow paced games $a = 0.3466$) which were presented in [17] based on the QoE studies done in [29].

5.7 Evaluation of quality features

In addition to overall QoE, we asked participants to rate other quality features, including: perceived immersion, perceived responsiveness, perceived fluidity, and perceived challenge. These measures have been discussed in Section 4.4 and are summarized in Table 5.

When looking at the relationship between reported QoE and the aforementioned quality metrics, we found that for all features but challenge there is a very significant correlation. The relationship between QoE and other quality features for Study 1 are illustrated on a heatmap in Figure 16. The results for Study 2 are very similar. This indicates a high correlation level between the quality features and QoE, except for challenge level. This result was expected, as challenge level is determined not only by the system parameters being enforced in each experiment, but also (and much more) by the situation in the game which is dynamic and can not be exactly controlled.

Pearson's correlation coefficients between all the quality features for all data are given in Table 5. It can be seen that besides having very high correlation with QoE (around 0.8), the quality features are themselves

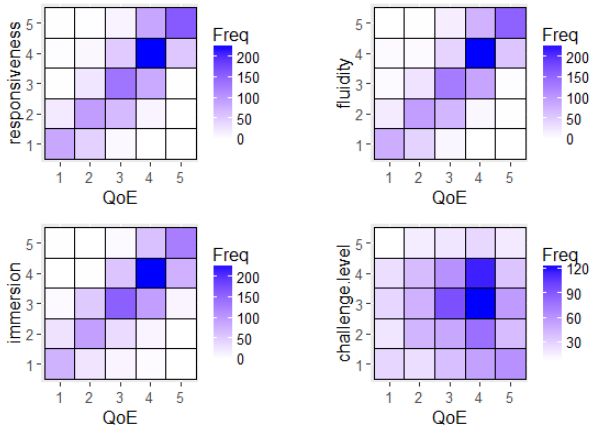


Fig. 16: Relationship of QoE and other quality features in Study 1

highly correlated (above 0.7). To confirm this finding we also calculated correlation coefficients for scenarios from Study 1 in which only one parameter was degraded (i.e., scenarios 3-18). Even for only those scenarios, correlation coefficients are very high for fluidity and responsiveness, which in these cases should not be so strongly correlated because they are determined by different system influence factors with only one common. Fluidity is related to frame rate and jerkiness while responsiveness is related to latency, packet loss, and in smaller measure jerkiness. This leads us to the conclusion that users hardly differentiate between separate quality features, and that being asked to rate one feature may have a large impact on the ratings of other features (even though the other features may not be manipulated with IFs). Therefore, we conclude that measuring multiple quality features may provide misleading results if each feature is considered separately without the notion of the other features measured during the same time.

5.8 Disruptive events during gameplay

Additionally, we wanted to inspect the influence of in-game performance parameters such character deaths, players getting lost in the virtual world, and unexpected disconnects from the game server. These events were noted for each of the performed scenarios (all 34 scenarios). Most of the noted events (over 90%) were character deaths. Our hypothesis was that if a player's character would die, or if something else disruptive occurred, that would result in lower QoE. We inspected the mean QoE of all the players who reported disruptive events in a particular scenario versus the QoE of players with no disruptive events. We found that there was no significant correlation between in-game performance and QoE in our specific case. Our results indicate that in 13 scenarios the average reported QoE of players who faced disruptive events was even higher than average QoE of players who did not face any disruptive events. In 15 scenarios we found the opposite case, while in the 5 remaining scenarios no disruptive events were reported.

	Resp.	Fluid.	Immer.	Chall. lvl	QoE
Resp.	1.00	0.77	0.70	-0.04	0.83
Fluid.	0.77	1.00	0.73	0.00	0.82
Immer.	0.70	0.73	1.00	0.07	0.78
Chall. lvl	-0.04	0.00	0.07	1.00	0.00
QoE	0.83	0.82	0.78	0.00	1.00

Table 5: Correlation between quality features

6 QoE modeling

We now focus on investigating the effect of dependent and independent variables on QoE with predictive modeling techniques. We use two approaches:

(1) modeling the "raw" QoE results, meaning that each entry in the dataset represents a QoE score from a single person (i.e., a discrete value from the set $\{1, 2, 3, 4, 5\}$), and
(2) modeling the MOS values (in the continuous range $[1, 5]$) for every single combination of system, user, and context factors, and use them as an input to the modeling procedure.

For the second approach we use additional preprocessing on the data. Within each train/test set we aggregate samples which have the same values of input features to machine learning models, such that final QoE is computed as mean of QoEs for those samples (this procedure is done on training and test set distinctly).

A summary of data preprocessing and modeling techniques is given in Table 6 and an interested reader can find more about the applied algorithms in [28]. In addition to linear regression, we introduce the following linear models: partial least squares which can deal with highly correlated features, ridge regression, a penalized model which can deal with overfitting, and Support Vector Regression (SVRs) with linear kernel. For non-linear models we use SVRs with Radial Basis Function kernel, random forests, and also gradient boosting machines [14]. For modeling we use the statistical package Caret in R [32]. The categorical variable gender is encoded using one-hot encoding (i.e., representation of categorical variables as binary vectors) and skill level

Type	Model	Data preprocessing	Training and tuning
Linear	Linear regression	centering, scaling	10-fold cross-validation
	Partial least squares	centering, scaling	10-fold cross-validation
	Ridge regression	centering, scaling	10-fold cross-validation, grid search: λ = [0 -0 .1]
	SVRs (linear)	centering, scaling	10-fold cross-validation, grid search: C = [0.25 - 32]
Non-linear	SVRs (RBF)	centering, scaling	10-fold cross-validation, grid search: C = [0.25 – 32], sigma = [0.001 – 0.01]
	Random forests	-	number of trees = 1000, k = [6 : 10]
	Gradient boosting	-	leave-group-out cross validation, grid search: number of trees = [100 : 1000], tree depth = [1 – 9], shrinkage = [0.01 – 0.1]

Table 6: Summary of used QoE modeling techniques

is coded as values 1, 2, 3, for novice, intermediate, and experienced players, respectively.

First, we perform QoE modeling using data collected in a single play mode (Questing). Then, we perform QoE modeling using data collected in a group play mode, but only in a mixed group scenario (Dungeons). We build these models on data aggregated from Study 1 and Study 2, as training sets produced for each case separately would contain a small number of samples (e.g., in a mixed group scenario in Study 1). Finally, we model QoE by aggregating all data collected in Study 1 and Study 2.

The models are trained on single play datasets from the Study 1 and Study 2. Each dataset is first cleaned from missing values, then categorical variables are coded, and then randomly shuffled. After that, 75% of the randomly selected players from Study 1 dataset are assigned to training set, and 25% of players to test (evaluation) set. The same is done with data from Study 2. We performed the sampling of the dataset to training and testing set 3 times with three different seeds (500, 998, 3800). The results between different sampling techniques vary very slightly (e.g., 0.01 of the value of R^2). In the paper we report the best results obtained with seed values. For Questing and for the whole dataset, the seed value is 998, and in case of Dungeons the seed value is 3800.

Table 7 (left column) shows results, measured on the test set with Root Mean Squared Error (RMSE) and R^2 value. R^2 is a statistical measure for goodness of fit. Is shows how close the data is to the fitted line on a scale from 0 to 1. In general the higher R^2 the better the model explains the data. It is important to note that R^2 does not prove that the model is adequate for the data. In our model R^2 values are not high, which is manifested through rather large variance of some of the predicted values (e.g., ground truth for QoE score 3), and weaker prediction of QoE score 1. Nevertheless, R^2 values below 50% are used in some disciplines (e.g., psychology – human behavior).

The data suggests that there is little difference between linear and non-linear models. We observe that the relationship between input features, or variables is mostly captured equally good by non-linear models and linear models, as evidenced by the R^2 value. The best performing model is random forests (RF) for the complete dateset in the MOS modelling approach, with RMSE of 0.62 and R^2 of 0.63 whereas differences between models in other cases is almost non-existent. In general, due to the averaging of the values when calculating MOS and loss of variance the modelling procedure based on MOS when compared to raw QoE ratings yields models with higher R^2 and lower RMSE except in the Questing scenarios when R^2 was higher for QoE modeling approach.

6.1 Questing

Training samples for qesting are collected from 75% of players from both studies, and their distribution across QoE scores is as follows: 66(1), 159(2), 237(3), 361(4), and 193(5). The rest of the data, collected from 25% of players, is assigned to the test set, which contains 28, 50, 75, 89, and 52 samples of scores equal to 1, 2, 3, 4, and 5, respectively. Models in Table 6 were used.

When it comes to variable, or feature importance we in detail report the findings from the first modeling approach (i.e., the one based on raw QoE values). The model quality of the second modeling approach (i.e., based on MOS values) is depicted in Table 7 while the fit is illustrated in Figure 17. In linear regression the lowest p-value for t-statistic is obtained for jerkiness (p-value $< 2 \cdot 10^{-16}$), which is then followed by framerate (p-value $= 4.5 \cdot 10^{-11}$). Other two system impact factors have lower p-values, but still significant: loss ratio (p-value $= 2 \cdot 10^{-03}$) and latency (p-value $= 4 \cdot 10^{-03}$). Skill level (p-value = 0.022) was found to have only a minor influence on QoE when compared to the other variables. Similarly to our previous conclusion, the gender of the players was not found to have an influence on QoE in a single play (Questing) scenario (p-value = 0.28). On the other hand, SVR (RBF)

Type		Questing				Dungeons				Whole dataset			
	Model	**RMSE**		R^2		**RMSE**		R^2		**RMSE**		R^2	
	Input	QoE	MOS	QoE	MOS	QoE	MOS	QoE	MOS	QoE	MOS	QoE	MOS
Linear	Linear regression	0.82	0.81	0.54	0.48	0.80	0.74	0.38	0.48	0.79	0.66	0.48	0.57
	Partial least squares	0.82	0.81	0.54	0.47	0.80	0.74	0.38	0.47	0.79	0.66	0.48	0.57
	Ridge regression	0.82	0.81	0.55	0.47	0.80	0.74	0.38	0.48	0.79	0.66	0.48	0.57
	SVR (linear)	0.82	0.82	0.56	0.48	0.80	0.75	0.37	0.47	0.79	0.67	0.48	0.56
Nonln.	SVR (RBF)	0.81	0.82	0.57	0.47	0.80	0.74	0.38	0.47	0.87	0.66	0.48	0.56
	Random forests	0.84	0.82	0.53	0.47	0.83	0.74	0.37	0.48	0.81	**0.62**	0.48	**0.63**
	Gradient boosting machines	0.85	0.82	0.53	0.47	0.85	0.74	0.36	0.48	0.80	0.64	0.48	0.60

Table 7: Performance of the QoE models in the Questing scenario (left), Dungeons scenario (center), and QoE models based on the whole dataset (right).

ranked the following variables with respect to their effect on QoE: jerkiness (100), framerate (22), loss ratio (17), latency (17), gender (0.5) and skill level (0). The effect on QoE termed "variable importance" in Caret, is computed using a loss function (mean squared error, MSE), measured on trees built by permuting input variables, and is normalized in the range [0 - 100] [32]. In our case, jerkiness was found to have the highest possible influence. Comparing these results to linear regression we see that both models identify the same variables as being important, and in the same order, except for last two least influential parameters.

Figure 17 depicts the performance of Linear regression and SVR (RBF), as the best performed model of the tested models in case of testing. It can be observed that machine learning models better handle the edge values of 1 and 5. An interesting point which can be observed is that the second approach has a large number of "rounded" values of 1, 2, 3, 4, and 5. It is understandable when there have been 120 unique combinations of all investigated parameters so only few of them have more than one value in our testing set - resulting in only few of cases (i.e., unique combination of the values of parameters taken into account in the modeling procedures) having multiple scores with different QoE values.

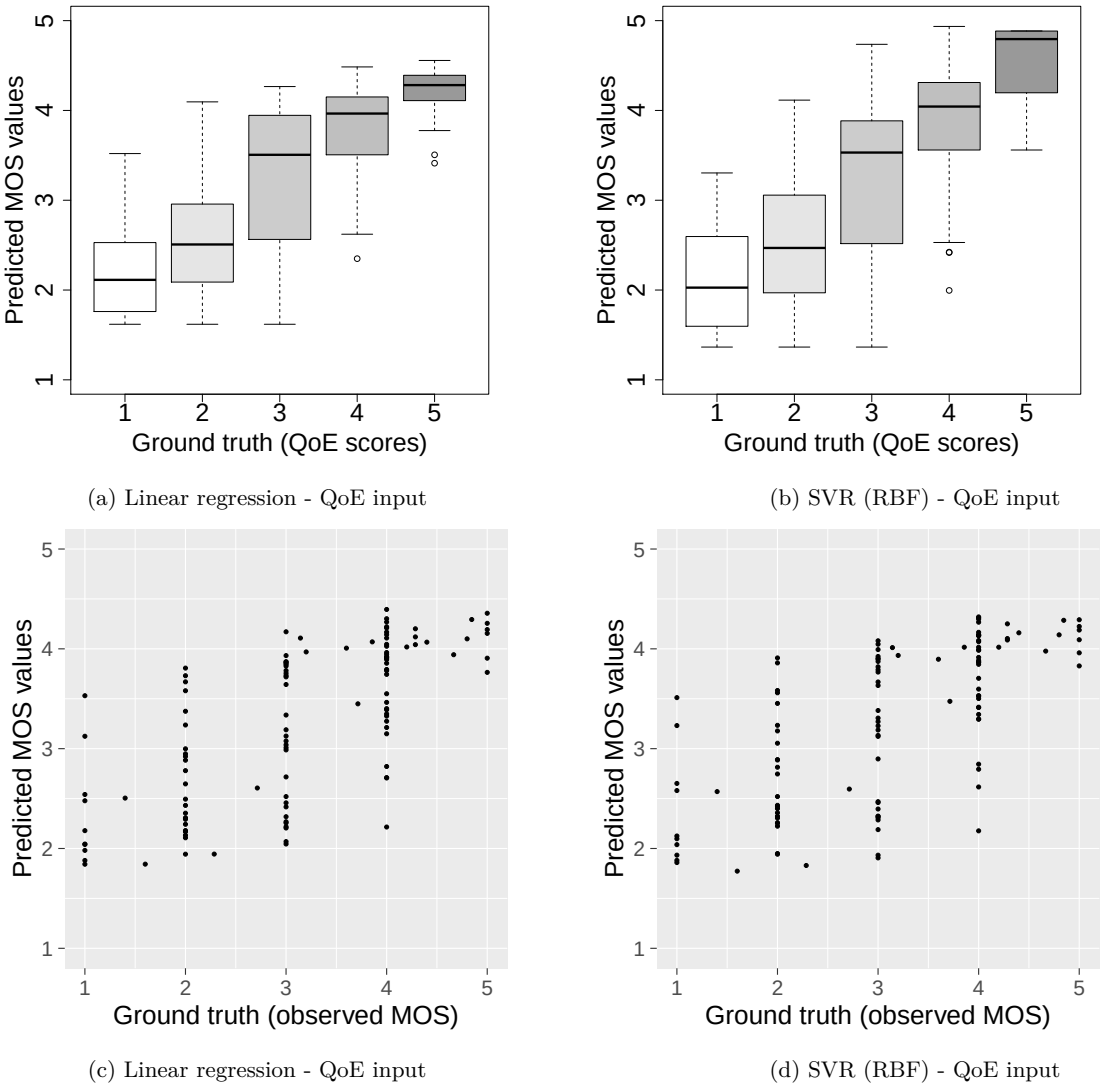
6.2 Dungeons

In Study 1, a subset of players (36 out of 55) participated in a mixed group play in the Dungeons scenario (meaning they played in groups composed of players with comparable experience levels). A response from this subset is concatenated to the Dungeons scenario samples collected in Study 2. For the training set, samples from 75% of randomly selected players are obtained from Study 1 and Study 2 (n samples = 695; QoE(n): 1(34), 2(96), 3(226), 4(255), 5(84)). The rest is assigned to the test set (n samples = 192; QoE(n): 1(4), 2(25), 3(67), 4(68), 5(28))). The same modeling methods as in subsection 6.1 are applied. Table 7 (middle column) shows results, measured on the test set with Root Mean Squared Error (RMSE) and R^2 value.

Results for data collected in the Dungeons scenarios show almost no differences between non-linear and linear models. Numerical values suggest similar performance to the models from the Questing scenarios. However, the distribution of predicted values on the test set data show slightly different behavior. Figure 18 shows predictions of the linear regression, predictions of the best non-linear model, which is SVR (RBF), on the test set. In both cases, predicted labels 1, 2 are more shifted towards label 3, when compared to prediction of the Questing models (Figure 17). On the other hand, variance of predicted labels is less evident in this case. Also, the results for the modeling based on MOS values are pretty similar with both models overestimating lower values and especially very low scores. We can see that most issues related to prediction for both approaches are on the edges of the data, especially for predicting very bad quality.

We list the p-values for the linear regression and importance values from the best performing non-linear model SVR (RBF). In linear regression the lowest p-value for t-statistic is obtained for jerkiness (p-value $< 2{\cdot}10^{-16}$), which is then followed by loss ratio (p-value $= 7.2{\cdot}10^{-12}$). For this model only skill level has significant p-value, although lower (p-value $= 1.92 \cdot 10^{-03}$). Other parameters latency (p-value = 0.29), gender (p-value = 0.49), and surprisingly frame rate (p-value = 0.92). The highest importance by SVR (RBF) is obtained for jerkiness (100), which is then followed by loss ratio (24). Skill level (3.7), gender (0.9), latency (0.5) and, surprisingly framerate (0), were found to have no influence on QoE when compared to other two variables. Again the same sequence of importance of variables are in both models.

Comparing these findings to the Questing scenario, we conclude that the same variables have a similar effect on QoE. In both cases jerkiness was found to be the most important, loss ratio somewhat important, and latency less important. The only exception is framerate,



(a) Linear regression - QoE input

(b) SVR (RBF) - QoE input

(c) Linear regression - QoE input

(d) SVR (RBF) - QoE input

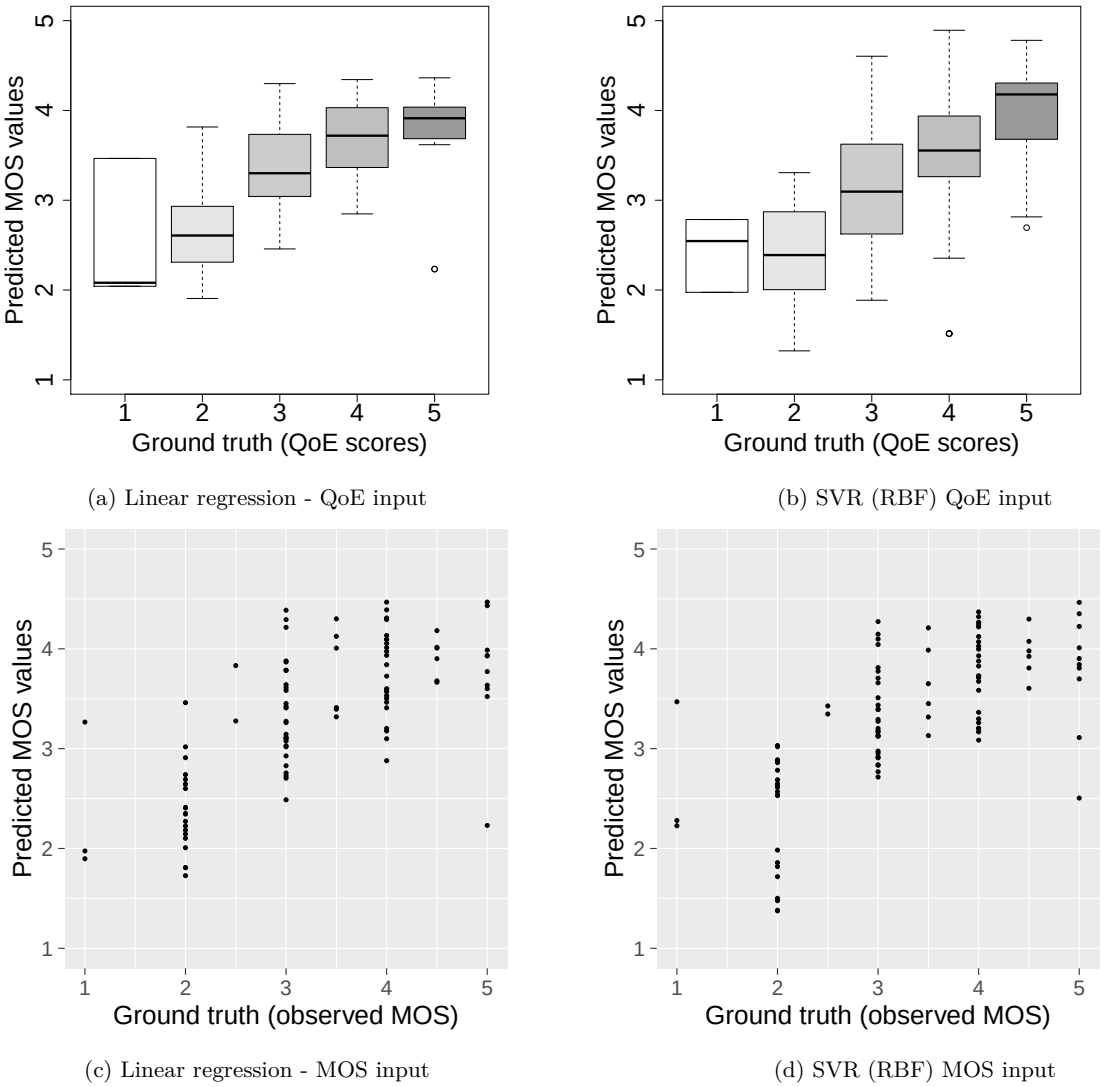
Fig. 17: Prediction results of the linear regression (left) and SVR(RBF) (right) in the Questing scenario

which has a bigger influence in the Questing scenario than in the Dungeons scenario.

6.3 Aggregated data

Finally, we aggregate all data collected in Study 1 and Study 2, and model QoE, leaving group out as a dependent variable. We consider this model as well because the analysis showed non conclusive data regarding the impact of different action categories, as well as non conclusive data regarding the impact of particular group composition, meaning that these parameters might not have statistically significant impact on the QoE. If the models perform similarly to those of a particular group, this is also an indicator that group can be ignored. The training set is generated as previously explained in subsections 6.1 and 6.2) - samples collected from 75% of players are selected and assigned to the training set (n samples = 1668; QoE(n): 1(112), 2(254), 3(464), 4(598), 5(260)). The rest of the data is assigned to a test set (n samples = 509; QoE(n): 1(20), 2(76), 3(141), 4(175), 5(97)).

Evaluation results, measured on the test set are given in Table 7 (right column). The results suggest that using an aggregation type of the modeling for this large dataset results in significantly better model (0.1 improvement in both RMSE and R^2 values). This is again logical, because a lot of the variance is lost in averaging the values to MOS per specific case, and in the case of aggregation of all of the data, there are more entries



(a) Linear regression - QoE input

(b) SVR (RBF) QoE input

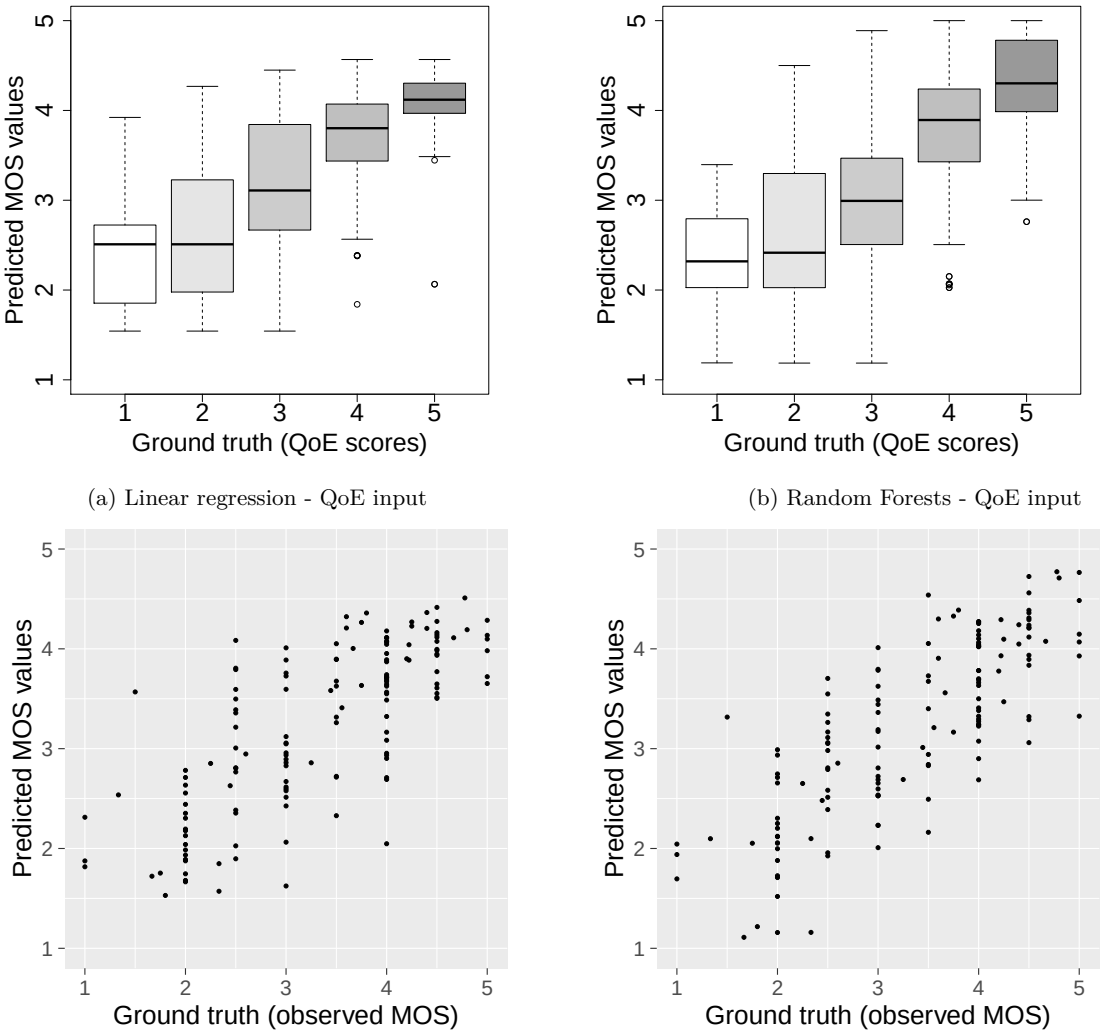
(c) Linear regression - MOS input

(d) SVR (RBF) MOS input

Fig. 18: Prediction results of the linear regression (left) and SVR(RBF) (right) in the Dungeons scenario

per specific case. The best performing model is random forests (RF), whereas differences between linear models are almost non-existent. Box plots of predictions on the test set for linear regression are shown in Figure 19a, and for random forests (RF) in Figure 19b. The distance between predicted values is shorter in case of linear regression, whereas the variance of predictions is slightly higher in case of RF. The models trained on whole dataset "suffer" from the same effect we have previously noticed – median of predicted values of ground truth label 1 is positioned around labels 2, and of label 5 around label 4. Figure 19d presents the best model fit - Random forests for MOS input values and it can be seen that this model has almost no overestimation on the middle part of the scale while still slightly overestimates lower values (1–2) and underestimates higher ones (4–5).

The importance of variables captured by our models is different in case of all the data taken into account. For linear regression as shown in Table 8, only gender of the players is found not important. When it comes to variable importance measured by RF, the following variables are ranked with respect to their effect on QoE: jerkiness (100), framerate (40), loss ratio (38), latency (36), skill level (22) and gender (0). In addition to presenting the RF model for the QoE based modeling approach we also list the importance of values for RF model for the model with highest R^2 value RF for the MOS modeling approach: jerkiness (100), loss ratio (41), framerate (25), latency (18), skill level (14) and



(a) Linear regression - QoE input

(b) Random Forests - QoE input

(c) Linear regression - MOS input

(d) Random Forests - MOS input

Fig. 19: Prediction results of the linear regression (left) and Random Forests (right) for aggregated data

gender (0). The importance of values is the same except that in the MOS modeling approachh loss ratio is found to be more important then frame rate.

Variable	**Estimate**	**Std. Error**	**t value**	**p-value**
(Intercept)	3.39865	0.01993	170.556	< 2e-16 ***
skill level	-0.13353	0.02368	-5.639	2.02e-08 ***
latency	-0.11441	0.02070	-5.527	3.79e-08 ***
loss ratio	-0.17925	0.02060	-8.701	< 2e-16 ***
framerate	0.12220	0.02033	6.010	2.29e-09 ***
jerkiness	-0.64188	0.02051	-31.301	< 2e-16 ***
gender	-0.02830	0.02367	-1.196	0.232

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Table 8: Summary of linear regression QoE model built on the whole dataset

Table 8 gives a summary of the linear regression model, as output by Caret. Gender is coded as a binary variable, set to 1 for male and 0 for female and skill is coded as 1, 2, 3, for novice, intermediate, and experienced players respectively. Estimate parameters give an approximation of the mathematical model for the final QoE:

$$QoE = 3.4 - 0.13 \cdot S - 0.11 \cdot L - 0.17 \cdot LR + 0.12 \cdot FR - 0.64 \cdot J \quad (4)$$

where S is the skill level, L is the latency, LR is the packet loss rate, FR is the framerate, and J denotes the jerkiness values. Note that in this formula we left the out variables that our linear regression model found insignificant (gender). It should be noted that the model values are based on ranges we manipulated within our studies (see Table 2 for numerical values).

Compared to linear regression, RF provides slightly better results, and it suggests that there might be a more complex relationship between variables affecting QoE than a simple linearity (e.g. when it comes to framerate, QoE might have a constant value until a certain threshold, and then it can degrade linearly/exponentially – as indicated in related work [8]). However, RF does not offer any possibility of deriving a mathematical formula for QoE.

7 Conclusions and future work

Assessing and modeling player QoE for networked games is clearly a complex task, with a number of open research issues. Our key findings may be summarized as follows:

- We found no significant impact of gender on QoE.
- We did not find previous experience to have a significant impact on game QoE in terms of our three defined categories. There was a significant difference only when comparing novice and experienced players. More experienced players were found to be more critical than non-skilled players, which can be concluded based on lower average QoE scores for skilled players.
- For the given case of WoW, there was no definitive conclusion as to whether action category affects QoE.
- We found that there was no significant correlation between in-game performance and QoE.
- Out of four manipulated system factors (delay, packet loss, jerkiness, and frame rate), we found jerkiness to have the greatest impact on players' QoE scores, followed by packet loss and frame rate. This may of course be related to the strength of factor manipulations.
- Increasing network delay resulting in RTTs from 40 ms up to 440 ms – surprisingly – did not result in significantly lower QoE scores.
- The impact of packet loss on QoE was greater in the case of the action category *Dungeons* as opposed to *Questing*, indicating also that the impact of system factors differs depending on the concrete actions being performed by the players. It should be taken into account that the impact of actions taken was not confirmed for all system factors.
- Players' social context (in terms of the experience level of other players in a joint team) in certain cases had an impact on QoE.
- There is a very significant correlation between QoE and other quality metrics (immersion, fluidity, and responsiveness).
- In terms of modeling, better results were found with non-linear than with linear models.

We highlight a number of study limitations which need to be considered. First, our studies were performed in a laboratory environment, and while the majority of players stated that they did not mind playing in a laboratory space and using equipment other than their own, this set-up may still have an impact on QoE. Hence, studies such as those we reported may be complemented with field and/or crowdsourced studies to provide more ecologically relevant results. Moreover, while we focused on lab tests, more valid results might be obtained from long-term longitudinal studies. Secondly, our test population was comprised of electrical engineering and computer science students, and future studies should address a wider population base. With respect to the actual test procedures, it should be noted that not all combinations of values for system parameters were tested due to time constraints. Approaches using crowdsourcing might alleviate these issues. We also opted to use a simple rating questionnaire to assess QoE and various QoE dimensions, given time constraints and the large number of test scenarios. It should however be noted that more complex and comprehensive questionnaires have been used in previous studies, such as the Game Experience Questionnaire (GEQ) [22], which makes use of 42 items to assess the features immersion, tension, competence, flow, negative affect, positive affect, and challenge. For our test purposes, using the GEQ was not feasible in the time alloted to test participants.

With regards to further work, additional studies are needed to consider more fine-grained factor manipulations and their impact on QoE, paving the way for deriving a QoE model for MMORPGs. Also, other MMORPGs should to be studied in order to generalize results. Such models could be possibly extended for other game types which rely on similar virtual world mechanics. Finally, while we have observed correlations between overall QoE and the identified quality features immersion, responsiveness, and fluidity, further studies based on multidimensional analysis and regression techniques may be used to identify additional QoE dimensions and their relevance in terms of overall QoE. Our long term goal may be seen as the development of a validated gaming QoE model which considers key system, user, and context factors, derived based on extensive experimental results.

Acknowledgements This work has been fully supported by Croatian Science Foundation under the projects 8065 and 5605.

References

1. Bernhaupt, R.: Evaluating User Experiences in Games: Concepts and Methods, chap. User Experience Evaluation: An Overview on Methods used in Entertainment, pp. 3–10. Springer (2010)
2. Beyer, J., Miruchna, V., Möller, S.: Assessing the impact of display size, game type, and usage context on mobile gaming QoE. In: Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on, pp. 69–70. IEEE (2014)
3. Chang, Y.C., Chen, K.T., Wu, C.C., Ho, C.J., Lei, C.L.: Online game QoE evaluation using paired comparisons. In: 2010 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR 2010), pp. 1–6. IEEE (2010)
4. Chen, K.T., Huang, P., Lei, C.L.: How sensitive are online gamers to network quality? Communications of the ACM **49**(11), 34–38 (2006)
5. Chen, K.T., Tu, C.C., Xiao, W.C.: Oneclick: A framework for measuring network quality of experience. In: INFOCOM 2009, IEEE, pp. 702–710. IEEE (2009)
6. Chen, K.T., Wu, C.C., Chang, Y.C., Lei, C.L.: A crowdsourceable QoE evaluation framework for multimedia content. In: Proceedings of the 17th ACM international conference on Multimedia, pp. 491–500. ACM (2009)
7. Chen, P., El Zarki, M.: 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, NetGames '11, pp. 2:1–2:6. IEEE Press, Piscataway, NJ, USA (2011). URL `http://dl.acm.org/citation.cfm?id=2157848.2157851`
8. Claypool, K., Claypool, M.: On frame rate and player performance in first person shooter games. Springer Multimedia Systems Journal (MMSJ) **13**(1), 3–17 (2007)
9. Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience. Harper and Row (1990)
10. De Moor, K., Ketyko, I., Joseph, W., Deryckere, T., De Marez, L., Martens, L., Verleye, G.: Proposed framework for evaluating quality of experience in a mobile, testbed-oriented living lab setting. Mobile Networks and Applications **15**(3), 378–391 (2010)
11. Denieffe, D., Carrig, B., Marshall, D., Picovici, D.: A game assessment metric for the online gamer. Advances in Electrical and Computer Engineering pp. 3–6 (2007)
12. Dick, M., Wellnitz, O., Wolf, L.: Analysis of factors affecting players' performance and perception in multiplayer games. In: Proceedings of 4th ACM SIGCOMM workshop on Network and system support for games, NetGames '05, pp. 1–7. ACM, New York, NY, USA (2005). DOI 10.1145/1103599.1103624. URL `http://doi.acm.org/10.1145/1103599.1103624`
13. Elo, A.: The Rating Of Chess Players, Past and Present. New York: Arco Publishing (1975)
14. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. Annals of Statistics **29**, 1189–1232 (2000)
15. Geel, I.V.: MMOData blog, Keeping track of the MMORPG scene, version 4.1 (2013). URL `http://mmodata.blogspot.com/`
16. Herbrich, R., Minka, T., Graepel, T.: TrueSKill^{TM}: A Bayesian Skill Rating System. Tech. Rep. MSR-TR-2006-80, Microsoft Research (2006)
17. Hoßeld, T., Schatz, R., Egger, S.: SOS: The MOS is not enough! In: Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on, pp. 131–136. IEEE (2011)
18. Hoßfeld, T., Heegaard, P.E., Varela, M., Möller, S.: QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS. Quality and User Experience **1**(1), 2 (2016)
19. Hoßfeld, T., Keimel, C., Hirth, M., Gardlo, B., Habigt, J., Diepold, K., Tran-Gia, P.: Best practices for QoE crowdtesting: QoE assessment with crowdsourcing. IEEE Transactions on Multimedia **16**(2), 541–558 (2014)
20. Howard, E., Cooper, C., Wittie, M.P., Swinford, S., Yang, Q.: Cascading impact of lag on user experience in cooperative multiplayer games. In: Network and Systems Support for Games (NetGames), 2014 13th Annual Workshop on,(Nagoya, Japan), pp. 1–6 (2014)
21. Hsu, C.L., Lu, H.P.: Why do people play on-line games? an extended TAM with social influences and flow experience. Information & management **41**(7), 853–868 (2004)
22. IJsselsteijn, W., Poels, K., de Kort, Y.A.: The game experience questionnaire: Development of a self-report measure to assess player experiences of digital games. TU Eindhoven, Eindhoven, The Netherlands (2008)
23. International Telecommunication Union: Opinion Model for Video-Telephony Applications. ITU-T Recommendation G.1070 (2012)
24. International Telecommunication Union: The E-model: A Computational Model for Use in Transmission Planning. ITU-T Recommendation G.107 (2015)
25. International Telecommunication Union: Reference guide to quality of experience assessment methodologies. ITU-T Recommendation G.1011 (2016)
26. ITU-T: Recommendation P.10/G.100 - Vocabulary for Performance and Quality of Service (2017)
27. ITU-T Study Group 12: Subjective test methodology for gaming based applications *(provisional title)*, Work item: P.GAME) (2017). Available: `https://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=13773` [Accessed: December 17, 2017].
28. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning - with Applications in R. Springer (2013)
29. Jarschel, M., Schlosser, D., Scheuring, S., Hoßfeld, T.: 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)
30. Kaiser, A., Maggiorini, D., Boussetta, K., Achir, N.: On the objective evaluation of real-time networked games. In: Proceedings of IEEE Global Telecommunications Conference, 2009, pp. 1 – 5 (2009)
31. Kim, H.S., Yoon, C.H.: Determinants of subscriber churn and customer loyalty in the Korean mobile telephony market. Telecommunications policy **28**(9), 751–765 (2004)
32. Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A.: Caret: Classification and Regression Training (2012). URL `http://CRAN.R-project.org/package=caret`. R package version 5.15-044
33. Kuipers, F., Kooij, R., De Vleeschauwer, D., Brunnström, K.: Techniques for measuring quality of experience. In: Proceedings of the 8th international conference on Wired/Wireless Internet Communications, WWIC'10, pp. 216–227. Springer-Verlag, Berlin, Heidelberg (2010). DOI 10.1007/978-3-642-13315-2_18. URL `http://dx.doi.org/10.1007/978-3-642-13315-2\_18`
34. Le Callet, P., Möller, S., Perkis, A.: QUALINET White Paper on Definitions of Quality of Experience. European

Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003). Version 1.2 (2013)
35. Metzger, F., Rafetseder, A., Schwartz, C., Hoßfeld, T.: Games and frames: A strange tale of QoE studies. In: Proceedings of the International Conference on Quality of Multimedia Experience, Lisbon, Portugal, pp. 1–2 (2016)
36. Möller, S., Antons, J.N., Beyer, J., Egger, S., Castellar, E.N., Skorin-Kapov, L., Sužnjević, M.: Towards a new ITU-T recommendation for subjective methods evaluating gaming QoE. In: Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on, pp. 1–6. IEEE (2015)
37. Möller, S., Schmidt, S., Beyer, J.: Gaming Taxonomy: An Overview of Concepts and Evaluation Methods for Computer Gaming QoE. In: International Workshop on Quality of Multimedia Experience, QoMEX, pp. 1–6 (2013)
38. NewZoo: The global games market will reach $108.9 billion in 2017 with mobile taking 42% . Tech. rep. (2017). URL `https://goo.gl/VeJuB3`
39. Poels, K., de Kort, Y., Ijsselsteijn, W.: 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)
40. Ribeiro, F., Florêncio, D., Zhang, C., Seltzer, M.: CROWDMOS: An approach for crowdsourcing mean opinion score studies. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2416–2419. IEEE (2011)
41. Ries, M., Svoboda, P., Rupp, M.: Empirical study of subjective quality for massive multiplayer games. In: Systems, Signals and Image Processing, 2008. IWSSIP 2008. 15th International Conference on, pp. 181 –184 (2008). DOI 10.1109/IWSSIP.2008.4604397
42. Slivar, I., Skorin-Kapov, L., Suznjevic, M.: Cloud gaming QoE models for deriving video encoding adaptation strategies. In: Proceedings of the 7th International Conference on Multimedia Systems, p. 18. ACM (2016)
43. Streijl, R.C., Winkler, S., Hands, D.S.: Mean opinion score (MOS) revisited: methods and applications, limitations and alternatives. Multimedia Systems **22**(2), 213–227 (2016)
44. Suznjevic, M., Dobrijevic, O., Matijasevic, M.: MMORPG player actions: Network performance, session patterns and latency requirements analysis. Multimedia Tools and Applications **45**(1-3), 191–241 (2009)
45. Suznjević, M., Saldana, J., Matijašević, M., Vuga, M.: Impact of Simplemux Traffic Optimisation on MMORPG QoE. In: ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016), pp. 1–6 (2016)
46. Suznjevic, M., Skorin-Kapov, L., Matijasevic, M.: The impact of user, system, and context factors on gaming QoE: a case study involving MMORPGs. In: Proceedings of Annual Workshop on Network and Systems Support for Games, pp. 1–6. IEEE Press (2013)
47. Takatalo, J., Hakkinen, J., Kaistinen, J., Nyman, G.: Evaluating User Experiences in Games: Concepts and Methods, chap. Presence, Involvement, and Flow in Digital Games, pp. 23–46. Springer (2010)
48. Ubicom Inc. Whitepaper: OPScore, or Online Playability Score: A Metric for Playability of Online Games with Network Impairments (2005). URL `www.kevingee.biz/wp-content/uploads/2011/04/IP3K-DWP-OPSCORE-10.pdf`
49. Verdejo, A.J., De Moor, K., Ketyko, I., Torben Nielsen, K., Vanattenhoven, J., De Pessemier, T., Joseph, W., Martens, L., de Marez, L.: QoE Estimation of a Location-Based Mobile Game using on-body sensors and QoS related data. In: Proceedings of the 2010 IFIP Wireless Days Conference âĂŞ Wireless Multimedia and Entertainment, pp. 1–5. Venice, Italy (2010)
50. Wattimena, A.F., Kooij, R.E., van Vugt, J.M., Ahmed, O.K.: Predicting the perceived quality of a first person shooter: the Quake IV G-model. In: Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games, NetGames '06. ACM, New York, NY, USA (2006). DOI 10.1145/1230040.1230052. URL `http://doi.acm.org/10.1145/1230040.1230052`
51. Weber, R., Tamborini, R., Westcott-Baker, A., Kantor, B.: Theorizing flow and media enjoyment as cognitive synchronization of attentional and reward networks. Communication Theory **19**(4), 397–422 (2009)
52. Xu, J., Xing, L., Perkis, A., Jiang, Y.: On the properties of mean opinion scores for quality of experience management. In: Multimedia (ISM), 2011 IEEE International Symposium on, pp. 500–505. IEEE (2011)
53. Yee, N.: The psychology of massively multi-user online role-playing games: Motivations, emotional investment, relationships and problematic usage. In: Avatars at work and play, pp. 187–207. Springer (2006)
54. Yee, N.: Experimental Motives for Playing Online Games. Journal of CyberPsychology and Behavior **9**(6), 772–775 (2007)
55. Zander, S., Armitage, G.: Empirically Measuring the QoS Sensitivity of Interactive Online Game Players. In: Proceedings of the Australian Telecommunications Networks and Applications Conference 2004 (ATNAC2004), pp. 511–518 (2004)
56. Zec, M., Mikuc, M.: Operating system support for integrated network emulation in IMUNES. In: Workshop on Operating System and Architectural Support for the on demand IT Infrastructure (1; 2004) (2004)