Quality-Influencing Factors in Mobile Gaming

vorgelegt von
Dipl.-Ing.
Justus Philipp Beyer
geb. in Leipzig

von der Fakultät IV – Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Ingenieurwissenschaften
- Dr.-Ing. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr.-Ing. Albayrak
Gutachter: Prof. Dr.-Ing. Sebastian Möller
Gutachterin: Prof. Dr. Lea Skorin-Kapov
Gutachter: Dr. Raimund Schatz

Tag der wissenschaftlichen Aussprache: 14. November 2016

Berlin 2017

Abstract

In the wake of the smartphone revolution, mobile games have not only become a spare time activity for the majority of phone owners, they have also created a prospering new industry. To thrive in an increasingly stiff competition, both game developers and service providers are seeking to improve their customers' gaming experience and understand how it is affected by external influences in order to distinguish themselves from their competitors.

However, playing experience is the result of a complex interplay of numerous factors: While the game itself sets the stage and determines the rules, look, and sound of the play, its implementation has to adapt to the player's device properties such as its screen size and available input methods, mobile network degradations, and respond to sudden interruptions such as incoming phone calls or contextual events like the player's arrival at the right bus stop gracefully. Although subjective effects of many influences have been studied for PC or console-based gaming in the past, this knowledge cannot be applied to mobile games straightforwardly as they differ from their stationary counterparts in various ways: Since smartphones and tablets are multi-purpose devices, they lack gaming-specific controls such as joysticks or game-pads and instead feature touch input which leads to the obstruction of manipulated parts of the screen and conveys no immediate haptic feedback.

Consequently, this thesis investigates the subjective effects of variations of the four quality-influencing factors game, device, network, and context in mobile touch-based gaming individually using experimental studies with test participants. Conclusions are then drawn on how each of these factors influences a player's gaming experience. As common interactive methods for assessing gaming quality are time-consuming and potentially unrealistic due to interruptions incurred by the subjective self-assessments, two additional studies are presented, which explore novel test methodologies. The first investigates the applicability of a standard non-interactive video assessment method for evaluating aspects of gaming quality, whereas the second examines using a physiological measure to obtain quality correlates as a substitute for having to interrupt and ask the player.

Finally, this thesis concludes with a discussion of how the found effects of game implementation, device size and network bandwidth affect future subjective gaming studies and considers further directions for research.

Zusammenfassung

Infolge der zunehmenden Verbreitung von Smartphones entwickelten sich mobile Spiele nicht nur zu einer Freizeitbeschäftigung für die Mehrzahl der Smartphone-Besitzer, sie schufen auch eine prosperierende neue Industrie. Um im wachsenden Wettbewerb bestehen und sich von ihrer Konkurrenz abheben zu können, streben Spiele-Entwickler und Service-Provider zunehmend danach, das subjektive Spieleerleben ihrer Kunden zu verbessern und zu verstehen, welchen externen Einflüssen dieses unterliegt.

Dieser subjektive Qualitätseindruck ist jedoch das Ergebnis eines komplexen Zusammenspiels einer Vielzahl von Faktoren: Während das Spiel selbst die Spielregeln, den visuellen und auditiven Eindruck bestimmt, muss sich seine technische Implementierung darüber hinaus an Eigenschaften des Endgeräts wie dessen Bildschirmgröße und verfügbare Eingabemethoden anpassen, auftretende Netzwerkstörungen kompensieren oder verschleiern und zudem angemessen auf auftretende Unterbrechungen wie eingehende Anrufe oder kontextuelle Ereignisse wie z.B. das Erreichen der richtigen Bushaltestelle reagieren.

Obwohl subjektive Effekte von zahlreichen Einflüssen für PC- oder Konsolenbasiertes Spielen bereits in Studien untersucht wurden, lassen sich deren Erkenntnisse nicht uneingeschränkt auf mobile Spiele übertragen. Sie unterscheiden sich von ihren stationären Pendants in vielfältiger Weise: Weil Smartphones und Tablets Mehrzweckgeräte sind, fehlen ihnen spielespezifische Eingabemöglichkeiten wie Joysticks oder Gamepads. Stattdessen werden die Geräte mittels Touchscreen bedient, wodurch es zu einer Verdeckung der berührten Bildschirmstelle kommt und zudem kein haptisches Feedback erfahren wird.

In dieser Dissertation werden folglich Änderungen der vier Einflussfaktoren Spiel, Gerät, Netzwerk und Nutzungskontext einzeln für mobile Touch-basierende Spiele im Rahmen von Nutzerstudien untersucht und hieraus Schlussfolgerungen abgeleitet, wie diese einzelnen Faktoren auf das subjektive Spieleerleben einwirken.

Da übliche interaktive Verfahren zur Bestimmung der Spielequalität zeitintensiv und durch wiederholte Unterbrechungen zur Abfrage subjektiver Selbsteinschätzungen potentiell unrealistisch sind, werden zwei weitere Studien präsentiert, die sich mit neuartigen Untersuchungsverfahren auseinandersetzen. Die erste hiervon untersucht die Anwendbarkeit von nicht-interaktiven Videobeurteilungsmethoden zur Untersuchung der Spielequalität, während

die zweite Studie die Eignung eines physiologischen Verfahrens untersucht, Qualitätskorrelate zu ermitteln, anstatt hierfür den Spieler unterbrechen und fragen zu müssen.

Schließlich werden in dieser Dissertation die gefundenen Effekte von Spielimplementierung, Gerätegröße und Netzwerkbandbreite und der widerlegte Kontexteinfluss diskutiert und mögliche Ansätze für weiterführende Forschung betrachtet.

Acknowledgements

During the past four years at the Quality and Usability Lab at Technische Universität Berlin, I have had the privilege to work with a team of excellent people. They have been an inexhaustible source of inspiration, guidance, and support. For this, I am immensely grateful.

Foremost, I would like to express my appreciation and gratitude to my supervisor, Prof. Dr.-Ing. Sebastian Möller. You have been an outstanding mentor and coach, guiding and encouraging me over the course of my research and always finding the time for some quick advice despite your most packed calendar.

I would also like to thank my committee members, Prof. Dr. Lea-Skorin Kapov and Dr. Raimund Schatz for your advice and for your agreement and commitment to co-examine my thesis.

Furthermore, to my colleagues at the Lab, Dr.-Ing. Jan-Niklas Voigt-Antons, Dr.-Ing. Tilo Westermann, Dr.-Ing. Benjamin Bähr, Dr. Benjamin Weiss, Dr.-Ing. Tim Polzehl, Dr.-Ing. Florian Hinterleitner, Dr. Dennis Guse, Dr.-Ing. Friedemann Köster, Steffen Zander, and Richard Varbelow, you not only provided invaluable advice and encouragement, but also made the work a thoroughly fun and joyful experience. I miss not only our discussions on research and plenty of other topics, but also (and particularly!) our regular meetings at the foosball table.

Thank you also to Irene Hube-Achter and Yasmin Hillebrenner for your organizational support. You solved so many complex bureaucratic and administrative challenges with remarkable endurance, dependability, and often refreshing creativity.

Tobias Hirsch and the Telekom Innovation Laboratories IT team deserve special thanks for their continuous support and flexibility in finding balances between Telekom´s corporate IT rules and the network requirements of my academic research projects.

Finally, this would not have been possible without my family. I am infinitely grateful for your patience and continuous support. Ina, Oskar, and Karl, you provided the foundation, balance, and strength which allowed me to complete this work.

Table of contents

List of Abbreviations

ACR Absolute Category Rating

ANOVA Analysis of Variance

CoD Call of Duty: Black Ops III

CPU Central Processing Unit

CODEC coder-decoder

CSMA/CA Carrier Sense Multiple Access with Collision Avoidance

DCR Degradation Category Rating

EEG electroencephalography

ERP Event-Related Potentials

FEC Forward Error Correction

FPS First-Person Shooter

GEQ Game Experience Questionnaire

GPU Graphics Processing Unit

GPRS General Packet Radio Service

GTA V Grand Theft Auto 5

ITU-T International Telecommunication Union - Telecommunication Standardization Sector

KSS Karolinska Sleepiness Scale

LAN Local Area Network

MANOVA multivariate analysis of variance

MIPS Millions of instructions per second

MMORPG Massively Multiplayer Online Role-Playing Game

MOS Mean Opinion Score

MTU Maximum Transmission Unit

PC Personal Computer

PDA Personal Digital Assistant

PGQ Post-Game Experience Questionnaire

QoE Quality of Experience

QoS Quality of Service

RAM Random Access Memory

SAM Self-Assessment Manikin

SI Système international d'unités

SSD Solid State Drive

TV television

UDP User Datagram Protocol

UMTS Universal Mobile Telecommunications System

VR Virtual Reality

WLAN Wireless Local Area Network

Chapter 1

Introduction

Long before the antique, cultures like the Babylonians and the Egyptians kept Astragalus bones from animals and used them to play games of dice [79]. As early as 2600 BC, Mesopotamians already played the *Royal Game of Ur*, an ancient race game played with multiple dice and a richly decorated board with 20 squares [14].

With the history of playing games going back far into the ancient human past, it seems that there were always some people who had an intuitive understanding of what constituted a good game and made it worthwhile to play. Since these early origins, however, a great number of games with an ever increasing complexity have been developed. Modern *digital* games are literally the product of hundreds of person years of work[1], which add to the even more years of work going into the underlying digital gaming platforms, algorithms, communication networks, etc. as the *technology* used for gaming becomes more and more sophisticated.

This growth in complexity is coupled to a great increase of the number of factors influencing the game: Whereas early dice games made from sticks, stones, or bones depended on a manageable set of influences like rule knowledge and experience of the players, material quality (e. g., wood, stone, bone), and enough light to see the positions of the game parts (i. e., the game *state*), a modern digital game's hardware requirements and recommendations text alone exceeds the length of the entire game description of many non-digital games.

Mobile digital games, running on a smartphone or a tablet, are furthermore played not only in a stationary setting, but allow playing virtually anywhere and at any time. However, in contrast to nearly all non-digital games, these games are not played on or with items which were made specifically for the game. Smartphones and tablets are devices created for a great variety of activities among which gaming is just one of many. Consequently, they are not

[1] https://en.wikipedia.org/w/index.php?title=List_of_most_expensive_video_games_to_develop&oldid=719731485 (last accessed: 2016-05-15)

ideally adapted to gaming and may even suddenly interrupt a game when other urgent events like an incoming phone call occur.

1.1 Challenges and Motivation

To understand the influence and effect of parameter variations in such a complicated system, intuition is no longer sufficient to achieve optimal results. At the same time it becomes exceedingly difficult to isolate the cause of errors, as the parameter space has become so big, that simple trial and error cannot possibly consider every parameter combination. However, elaborate mathematical models of game experience probably can. Yet, to develop such models and to understand how various factors influence games, methods are required to traverse the immense parameter space and quantifiably measure the result of individual factor variations.

In many regards, this quantified 'result' can be an objective metric like frames per second [Hz], start-up time [s], or computational complexity [e. g., Millions of instructions per second (MIPS) $= 1\,/\mathrm{s} = \mathrm{Hz}$]. When it comes to the subjective perception of a game, different methods and measures are required, as, unfortunately, the Système international d'unités (SI) currently lacks appropriate units for amusement, fun, and flow. However, applicable (non-SI) measures for subjective gaming experience exist in the literature and are presented together with various measurement tools for both objective and subjective metrics in Chapter 2.

While immense work has gone into optimizing performance aspects of games, considerably less research focused on understanding their subjective experience, leaving the interplay of technical parameters and aspects of gaming experience only partially understood. This is particularly true with mobile games, which are played on multipurpose devices such as smartphones and tablets, connected via inherently unreliable wireless networks. Game-playing with mobile games is therefore exposed to external influences to a much greater degree than that with stationary equivalents. Here, a new and highly relevant research field is opening up, as more than two thirds of smartphone users nowadays use their device also for playing[2]. This makes mobile playing of games not only a spare time activity for many, but also a concern to service providers and network operators around the globe.

The aim of this thesis is to identify factors influencing mobile gaming experience and to assess their subjective effects. To select factors from the huge number of possible candidates, the perspective of a telecommunications or network provider is taken, which has an economic interest to optimize its service to improve the subjective experience and ultimately the service

[2] http://www.emarketer.com/Article/Growing-Number-of-Smartphone-Users-Driving-Mobile-Gaming-Consumption/1013686 (last accessed: 2016-05-18)

acceptance of its customers, many of whom are mobile game players. In order to perform these optimizations, this provider needs to have an understanding and ideally a model to predict how changes to infrastructure parameters will affect the experience of those players.

As this provider has only limited knowledge about the players' expectations, gaming preferences and experience, these aspects are mostly beyond reach for modeling and optimization purposes. However, the provider does possesses knowledge about the games its clients are playing, which devices they are using, how they are connected to the network, and, approximately, where they are playing (e. g., public place or at home). These are the pieces of information it may use in its effort to improve its service quality. Yet, a model explaining and predicting how these factors play together and how they are influencing a player's experience does not yet exist. Furthermore, such a model can only be developed with the knowledge which of these factors exert a meaningful influence in the first place.

Therefore, each of these factors is evaluated in this thesis and conclusions for a future comprehensive model of mobile gaming experience are drawn.

1.2 Thesis Outline

To enhance the understanding of the influence factors discussed in the last section, each of them is evaluated and discussed in a chapter in this thesis.

After reviewing the fundamentals of quality assessment with mobile games in Chapter 2, the presented methods and metrics are used to study the effects of the selected influence factor variations. In Chapter 3, the most obvious factor, the game itself, is varied to find out how comparable different games are and what role their specific implementation plays. One part of that implementation is to make a game run on a possibly great variety of devices. However, these devices vary by numerous parameters, of which the most important and obvious is their size. The effect of device size variations is therefore investigated in Chapter 4. While Mesopotamians, Babylonians, and Egyptians could only play games they had physical access to, smartphone and tablets possess networking capabilities, allowing them to access games which are actually computed somewhere else. This cloud gaming paradigm and the effects network degradations exert on it are examined in Chapter 5. In Chapter 6, the influence of mobility and the ability to play in various contexts on gaming are investigated. With Chapter 7, the attention turns back to measures and measurement methods: Two promising new test paradigms, physiological and purely passive gaming tests, are explored and compared to more conventional experimental means. Finally, Chapter 8 summarizes this thesis' key contributions and closes with an outlook on future work.

Chapter 2

Assessing the quality of mobile gaming

Soccer, Rock-paper-scissors, dice, and First-Person Shooters - they refer to completely different activities which nevertheless all share a common denomination: *game*. Other activities like educational "games", simulators, and interactive movies present border cases. To determine this border and agree on what ingredients make an activity a game, a definition is presented in this chapter after discussing and defining another rather ambiguous term: quality. With these terms defined, a taxonomy of gaming quality aspects is presented and quantifiable measures and measurement tools are discussed, which may then be used to assess the quality of digital gaming.

2.1 Quality

Quality is a highly multi-layered term which, during the past two decades, has been repeatedly redefined (cf. [38, 60, 64]) and changed in scope multiple times. Generally, quality can be regarded from the perspective of a provider of a service or product, or from the view of the user of that offering. These two common, yet different perspectives have led to a differentiation into Quality of Service (QoS) and Quality of Experience (QoE). The Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T), as an institution dominated by service providers, first defined the term QoS in 1994 [64] and has since updated the definition to read as follows [63]:

> **Quality of Service**: Totality of characteristics of a telecommunications service that bear on its ability to satisfy stated and implied needs of the user of the service.

In turn, the Qualinet initiative[1], a European network of Quality of Experience experts, has published a working definition for QoE [84]:

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

The terms *application* and *service* are furthermore defined [84]:

> **Application**: "A software and/or hardware that enables usage and interaction by a user for a given purpose. Such purpose may include entertainment or information retrieval, or other."
>
> **Service**: "An episode in which an entity takes the responsibility that something desirable happens on the behalf of another entity."

Whereas the ITU-T definition of QoS emphasizes the characteristics of a telecommunications service, the Qualinet definition of QoE focuses on the "delight or annoyance of the user". As QoS and QoE are two perspectives on the same problem, they are inextricably related. For any service, a series of *influence factors* (i. e., QoS characteristics) can be defined which shape a player's subjective Quality of Experience. The term *Influence Factor* has been defined in [84] as follows:

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

As the perception of QoE is clearly not a one-dimensional construct, but has many aspects specific to the product or service under scrutiny, the term *QoE feature* has been introduced [84]:

> **QoE feature**: A perceivable, recognized and nameable characteristic of the individual's experience of a service which contributes to its quality.

In the literature, a concept of *game usability* is raised occasionally. However, in this thesis, the term *usability* is rather used in connection with productivity apps, as the definition of QoE is considered to cover all QoE features which may be relevant to the player of a game.

Later in this chapter, a framework is discussed which aims at relating influence factors from the QoS domain to subjectively perceived QoE features.

[1] http://www.qualinet.eu

2.2 Game

Juul proposed a definition for a game, that is based on six features [74]:

1. *Rules*: Games are rule-based.
2. *Variable, quantifiable outcome:* Games have variable, quantifiable outcomes.
3. *Valorization of outcome:* The different potential outcomes of the game are assigned different values, some positive and some negative.
4. *Player effort:* The player exerts effort in order to influence the outcome (Games are challenging.)
5. *Player attached to outcome:* The player is emotionally attached to the outcome of the game in the sense that a player will be winner and "happy" in case of a positive outcome, but a loser and "unhappy" in case of a negative outcome.
6. *Negotiable consequences:* The same game [set of rules] can be played with or without real-life consequences.

This definition places a strong emphasis on the rules of a game as these are considered to be "the most consistent source of player enjoyment in games" [74]. Game rules are expected to be easy to learn, and, as they add up, to be more than the sum of their parts: "For most games, the strategies needed to play are more complex than the rules themselves." [74]

In contrast to the goal of completing tasks with minimal effort in task-oriented human-machine interaction, the primary aim of games is to provide an entertaining activity, where challenges are put in front of the user on purpose and difficulty is optimized to meet the player's capabilities. That difference prevents the easy application of standard methods for determining usability (including effectiveness, efficiency, but also hedonic quality aspects), which are used in productivity-oriented human-computer interaction, as these standard methods aim at determining the effort and therefore the challenge of achieving the productivity goal. Furthermore, the outcomes of a game themselves are not necessarily the most rewarding aspect, but the process of overcoming the challenges by investing effort and achieving the desired outcomes is [83]. Productivity-oriented applications, on the other hand, are designed to minimize challenges while achieving the desired outcome which is the most rewarding aspect.

Based on the platform they are implemented on, digital games have for long been broadly classified into computer games, which are played on general purpose PC hardware, console games (Xbox, PlayStation, Wii, etc.), mobile games which run on devices such as

smartphones, tablets, or special gaming hardware such as the PlayStation Portable, and online games which are often browser-based and require a constant Internet connection. As many recent computer and mobile games also contain features of online games, these sets are not disjunct: Both single- and multiplayer games on computers, console, and mobile platforms make use of Internet connections to coordinate interactions or exchange information such as leader boards, high scores, or updates. A special case is the so-called "cloud gaming", where the code execution, game logic and rendering of a digital game are physically executed on a remote server farm (cloud), and just the display and input interpretation take place on the player's device. Cloud gaming and the network's influence on the quality of that game delivery paradigm are discussed in more detail in Chapter 5.

2.2.1 Characteristics of mobile gaming

Mobile gaming differs from stationary gaming primarily in the hardware that is used for playing. As outlined above, two fundamentally different device categories can be distinguished: Special-purpose gaming hardware like the Nintendo Gameboy[2], Nintendo DS[3], or Playstation Vita[4], and multi-purpose hardware like smartphones or tablets. Whereas the former have dedicated physical buttons and sometimes joysticks to control games, the latter are usually limited to a few general purpose physical sensors and buttons like volume controls and touch or multi-touch input using a touchscreen. This influences the design of mobile games, as touch-screen metaphors of joysticks do not adequately substitute the originals due to the lack of haptic feedback [98]. Instead, (multi-)touch input requires permanent visual feedback. The on-screen response therefore requires additional cognitive effort and competes with other game elements for attention.

Despite these shortcomings, smartphone and tablet-based playing has grown to vastly exceed[5] that of mobile gaming consoles (i. e., special-purpose gaming hardware), rendering their once high relevance increasingly negligible. However, the enormous success of smartphones also brought a great variety of different devices from various manufacturers, compared to a low number of popular mobile consoles. As a result, mobile games typically have to adapt to numerous devices' varying capabilities due to the fragmentation of the smartphone market. Adding to this challenge for developers [46], mobile games operate in a much more resource-constrained environment than PC or console titles. Despite rapidly growing capabilities of mobile CPUs and GPUs, available energy and the ability to dissipate

[2]https://en.wikipedia.org/w/index.php?title=Game_Boy&oldid=718972842
[3]https://en.wikipedia.org/w/index.php?title=Nintendo_DS&oldid=718060373
[4]https://en.wikipedia.org/w/index.php?title=PlayStation_Vita&oldid=718971929
[5]http://fortune.com/2015/01/15/mobile-console-game-revenues-2015/ (last accessed: 2016-05-08)

heat without active cooling severely constrain the computational complexity of mobile games. To mitigate this limitation, attention has turned to the devices' networking capabilities and the concept of offloading and performing complex computations not on a mobile device itself, but on a less resource-constrained and networked server. This offloading of computational load is considered promising, as the energy cost of wirelessly transmitting computed results from the cloud to the device can be lower[6] than that of comparable local computations. Where one end of the range of possible work divisions is completely local (i. e., offline) execution of a game, cloud gaming is the opposite end. In the latter, the entire complexity of game execution is moved to dedicated servers in the cloud. Between these extremes a great diversity of gradations of offloading exist [81]. A popular example is multi-player gaming: Here, a server creates and maintains a game state which is synchronized with the participating clients. Through the shared system state, players can interact with other players in the common game world. However, since the dependency on a well-functioning and stable network connection grows with increased integration of remote resources, it becomes a more and more important influence factor to the perceived quality of a game, as, in practice, network parameters often change dynamically. This is discussed in more detail in Chapter 3, where three games are compared with regard to their gaming experience and the influence network impairments impinge on this.

Taken together, smartphones and tablets are technically not ideal gaming devices, as their design is a compromise to fit multiple use cases and they are limited in their resources. However, their mobility and particularly the other parallel purposes they may be used for place additional requirements on mobile games, which are uncommon for stationary games: A game may be interrupted and suddenly stopped at any time, as the player may receive a call, or might wish to react to an incoming notification [123]. This requires developers to design mobile games accordingly. To guide producers in this development process Korhonen *et al.* have formulated a set of requirements as evaluation heuristics [80]:

1. The game and play sessions can be started quickly
2. The game accommodates with the surroundings
3. Interruptions are handled reasonable

The ability to quickly start, stop, and handle interruptions is also considered to be critical by Henning, who stresses: "Interruptions in mobile gaming can come from anywhere: maybe

[6] http://www.tomshardware.com/reviews/nvidia-shield-tegra-4-android-geforce-review,3576-12.html (last accessed: 2016-05-08)

the bus has reached your stop, and you need to stuff the phone in your pocket and disembark [...] and they [the Player] may just get a phone call"[7].

Due to the mobility of the devices, mobile games can be played in many different contexts. One particularly popular setting to play is during commuting. Liu *et al.* justify the great success of mobile entertainment and mobile gaming in countries like China with the people's long average commute. They reason that while the use of bigger devices such as laptops is impossible due to the crowded environment, the space is always sufficient for a smartphone. Additionally, they found the usage context to be the strongest predictor for playing mobile games. The context was furthermore identified to exercise an even greater influence on people's decision to play than their attitude [85].

While the prevention of boredom may be one of driving forces behind mobile gaming, social aspects may also be responsible: Dixon *et al.* found gaming to play a role in avoiding social interaction and potential embarrassment as the activity prevents unintended eye contacts from happening [39].

2.2.2 Classifications of mobile games

Despite numerous efforts, a generic and uncontested classification of games, and particularly of mobile games, has not yet been established. In "Genre and the Video Game", Wolf defines 42 different genres [124] based on the core activities performed in a game. Examples of these categories are:

- *Racing*: titles involving winning of a race, covering more ground than an opponent
- *Flying*: titles involving flying skills including steering, altitude control, takeoff and landing
- *Shoot 'Em Up (or Shooter)*: shooting at, and often destroying, a series of opponents or objects
- *Sports*: Games which are adaptations of existing sports or variations of them.

However, these genres are not an unambiguous classification, because many games belong to several of these categories. A game involving Formula 1 car races would clearly fall into the category of *Racing*, but also into *Sports*. Wolf notes [124]:

> "The idea of genre has not been without difficulties, such as the defining of what exactly constitutes a genre, overlaps between genres, and the fact that genres are always in flux as long as new works are being produced."

[7] http://blog.triplepointpr.com/mobile-game-design-dont-forget-the-basics (last accessed: 2016-05-08)

As in academia, no common categorization exists in the industry: The most popular market places for mobile game sales, Apple's App Store and Google's Play Store, each have their own system of game categories. Whereas the App Store knows 18 classes of games[8], including generic groups like *Family* or *Trivia*, the Play Store distinguishes between 17 classes[9], which, despite a large overlap, differ in details from Apple's catalog. In both stores, apps can be listed in multiple categories, rendering their classes indistinct.

As one-dimensional classifications have proven to be difficult, multiple systems based on a game-ontological approach have been proposed. Those works characterize games by identifying functional aspects and conditions which are important to a game. Although these typologies are not specific for mobile games, they cover that domain as well. Aarseth *et al.* proposed a multi-dimensional typology of "games in virtual environments" in 2003 which is based on 15 dimensions grouped into the 5 meta-categories *Space*, *Time*, *Player Structure*, *Control*, and *Rules* [1]. Based on the former model, but being more fine-grained, is the typology model proposed by Elverdam *et al.* They suggest 8 meta-categories which form pairs of in-game and real-world attributes like *Virtual Space* and *Physical Space*, and *External Time* and *Internal Time* [41]. One functional aspect for classification which both Aarseth *et al.* and Elverdam *et al.* use, is the visual perspective of the player into the virtual world which may be *Omnipresent* (the player sees the whole game world, e. g., Pac-Man, chess), or *Vagrant* (just an except from the game world is shown, e. g., side-scrolling games). Another criterion can be the *Player Structure* (Aarseth *et al.*) or *Player Composition* (Elverdam *et al.*) which distinguishes games based on the number of concurrent players and their relationship to each other (e. g., cooperative, competitive). The number and role of players and their relationship has been considered also by Fullerton, who differentiated between, e. g., single player against the game, several players against the game, several players against each other, cooperative game, team game, etc. [43] Dahlskog *et al.* [37] created a catalog of 75 games and used an extended version of the typology from Aarseth *et al.* to categorize them based on their features. They found that older games did not exhibit many of the categories used to characterize and differentiate modern games and hypothesized, that with future games also additional categories will have to be added [37]. This might mean that, in consequence, no generic typology may exist, and that useful, unambiguous, and agreed-upon classifications will be limited to aspects of games instead of providing an overarching scheme.

Some of the aspects used for classification purposes in the above models may strongly influence the effect a technical platform has on user-perceived QoE, e. g., the sensitivity to

[8] https://itunes.apple.com/en/genre/ios/id36?mt=8 (last accessed: 2016-05-09)

[9] https://play.google.com/store/apps/category/GAME (last accessed: 2016-05-09)

parameters like delay may be more influential to some types than to others. In Section 5.3.3 another classification is proposed based on a game's visual output and its delay sensitivity.

2.3 Taxonomy of gaming quality aspects

Following the concept of the "Qualinet White Paper on Definitions of Quality of Experience", cited and quoted in Section 2.1, a taxonomy was developed in [92] with three layers containing influence factors, interaction performance aspects, and quality features, which are relevant for computer gaming. This taxonomy has since also been used and adapted for Cloud Gaming [51].

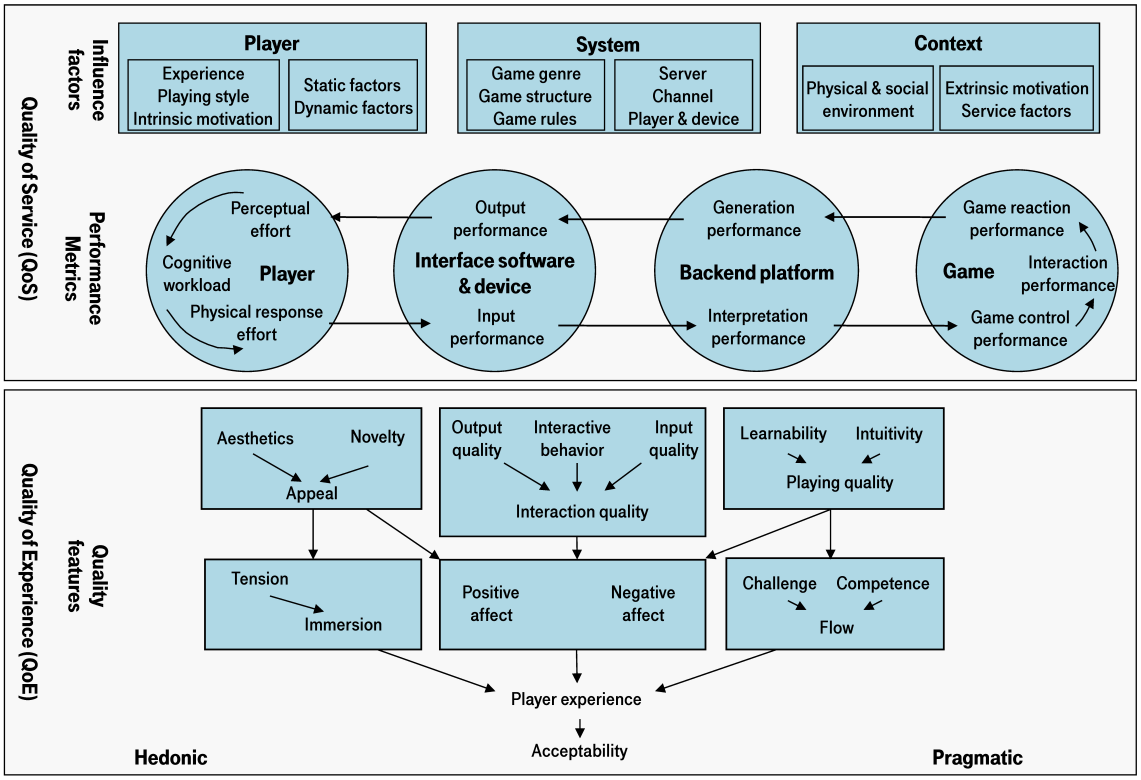


Fig. 2.1 Taxonomy of gaming QoE aspects, adapted from [16] and [92]
Upper panel: Influence factors and performance metrics; lower panel: QoE features.

In this thesis, the taxonomy has been slightly adapted to match the used terminology. For instance, in this text, the word *player* is preferred over *user*, as the latter term is more closely associated with productivity applications.

2.4 Influence factors

Factors influencing the quality of the gaming experience can be subdivided into the three groups: player factors, system factors, and context factors. This subdivision follows the structure proposed by the Qualinet initiative [84].

Player factors describe the impact that aspects of the player himself (i. e., the human being) have on the game experience. Notable examples of these influences are the player's experience with games (e. g., "newbie" vs. "pro gamer"), playing style (e. g., Bartle [10]: "achiever", "explorer", "socializer", and "killer"), intrinsic motivation, dynamic and static player factors. Many of these are difficult to control in an experimental study. However, a player's experience with games can be approximately gauged by the number of hours per week/month spent playing. This metric also allows inviting only participants with a minimum familiarity with gaming to studies. Factors, which are static at least for the duration of the experiment, are for example the player's age, gender, and native language. The player's emotional status, boredom, distraction, curiosity, etc. are considered as dynamic factors due to their change during the course of a study.

System factors not just refer to the game, but cover the setup as a whole. As such, relevant parameters are, e. g., the game and its content, rules, and implementation, the technical setup of the system with involved soft- and hardware and communication channels, and design characteristics which can be perceived by the player. This group of factors is of predominant interest in the following chapters, where the effects of variations in selected influence factors will be investigated with regard to their subjective effect.

Finally, the context factors encompass all situational influences, such as the physical environment (e. g., space, acoustics, lighting), the social context (e. g., relationships with other players or the presence of an experimenter), but also service factors like the price and availability of a service or game. A deeper look into the impact of varied locations with different physical and social contexts is taken in Chapter 6, where the subjective experience of playing is compared between a noisy public transportation setting and a quiet laboratory room.

2.5 Performance metrics

As in [89], a distinction was made between system- and player-related performance aspects. Furthermore, the system part was subdivided into the interface software and device, the back end platform, and the game. These modules may be spatially separated and interconnected using communication channels as in cloud gaming, where the player is interacting with a thin

interface software but the actual execution of the game takes place in a remote data center on the back end platform.

As a means to measure the performance of a game in a given environment, performance metrics such number of kills, deaths, level reached, the fastest time achieved, or points attained have been used. Using performance metrics to appraise the quality of a product has a long tradition in productivity applications. Here, an increased performance in terms of, e. g., more orders processed, more customers served, or less time or effort needed to perform a task is clearly desirable. The concept has, however, also been applied to games:

Beigbeder *et al.* monitored participants playing the First-Person Shooter (FPS) Unreal Tournament while delay and packet loss were simulated on the network connection. To elicit the effects of these degradations, they asked the subjects to perform a series of tasks in the game. In one, the time was measured that participants needed to move through an obstacle course. In another test, they calculated the fraction of precision shots that hit their intended target compared to the misses. In a 4-person multi-player setting, the number of accumulated kills and deaths was recorded [12]. In contrast to productivity applications, where performance metrics often have an absolute meaning (e. g., time below n seconds is considered good enough), in games they have no intrinsic sense of good or bad. Despite the comprehensive set of data collected, Beigbeder *et al.* can only make assumptions about the players' perceived degradation as the collected performance data is not coherently related to perceivable effects.

Similarly, Bredel *et al.* used another FPS in multi-player mode with bots playing against each other in an artificially degraded network environment and measured the scores of kills and wins. These numbers were then used as metrics to compare different network settings [25]. In a game from the Massively Multiplayer Online Role-Playing Games (MMORPGs) genre, Chen *et al.* investigated the player departure behavior from the game ShenZhou Online by obtaining a dataset of game traces. As metric, they used the duration of a gaming session and correlated that to various network impairments [30].

Comparable performance metrics have been used in numerous further studies investigating the effects of network impairments on games, e. g., [31, 33, 49]. However, the usability-inspired view represented by mere performance metrics does not reflect the subjective in-game experience to the full extent, as "the user's own goals when playing a digital game are not adequately captured by metrics such as 'time spent on task', or 'number of tasks successfully completed"' [49]. Further difficulties arise, when performance metrics are used to optimize a gaming system: Although lowering the number of deaths of the player's character and increasing metrics like number of points attained might seem welcome to the player, it likely interferes with the games ability to pose challenges, which cause the player to

exert “effort in order to influence the outcome” (cf. [74], Section 2.2). Consequently, as the game ceases to be challenging, it might become *less* attractive to play despite the increased awards. Furthermore, performance metrics are highly game-specific: First, metrics from a FPS like number of deaths per time unit are pointless in a racing game. But second, they are often even meaningless in another FPS title, because these individual games often differ in plenty of details (cf. Section 2.2.2), rendering performance metrics incomparable. Finally, these metrics are furthermore rendered elusive, as games may adapt their level of difficulty to the player’s capabilities and achievements in the current environment: Chanel *et al.* proposed a framework to adapt the difficulty in games based on real time measurements of emotions [27]. Antons *et al.* use a variety of parameters such as reaction time, and preferred modality to estimate capabilities of players in residential dementia care and adapt their game in real time [6]. In [86], Lopes *et al.* give an overview over existing indicators for the current player condition and methods to adapt game-play accordingly.

In consequence, performance metrics are useful only for the assessment of a limited set of gaming attributes as “objective parameters alone do not make a statement on the subjective game quality” [108].

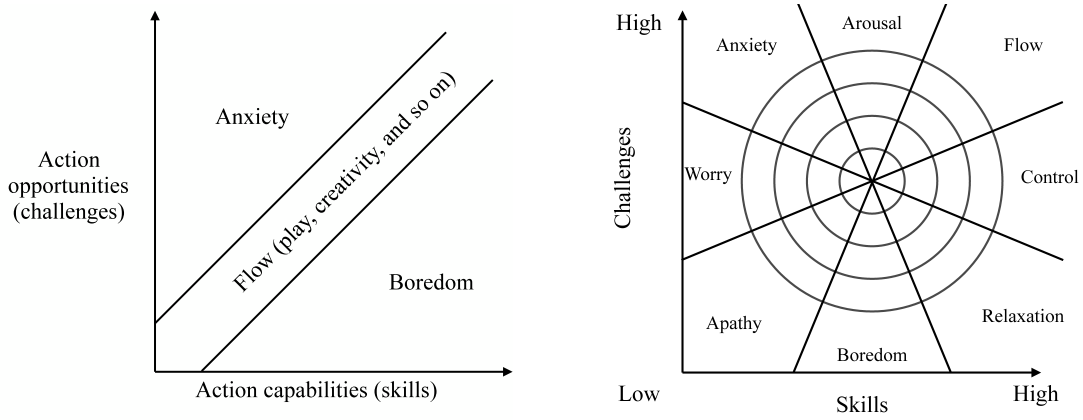
2.6 QoE features and subjective self-assessment

As objective performance metrics alone cannot adequately mirror the perception and *fun* of playing, the attention has turned to QoE features which can be measured using subjective self-assessment. Instead of using external ’objective’ observations, players are asked to reflect and describe their experience while interacting with a game. As defined in Section 2.1, a QoE feature is a “perceivable, recognized and nameable characteristic of the individual’s experience of a service which contributes to its quality.” [84] QoE features are reflected in the lower layer of the taxonomy in Figure 2.1, whereas the influence factors and performance metrics layers were depicted in the upper part together due to their objective nature.

In the literature, no consensus exists which QoE features best describe gaming experience. Instead, multiple one-dimensional measures and multi-dimensional frameworks have been proposed. One of these frameworks is the taxonomy from Möller *et al.* shown above, in which Flow, Immersion, and the Game Experience Questionnaire (GEQ)’s dimensions as one of most comprehensive models of gaming experience and other features from [89] such as the quality of in- and output are combined into a hypothetical framework of six groups of QoS features. In this section, Flow, Immersion, the GEQ and its dimensions, the Self-Assessment Manikin, the Karolinska Sleepiness Scale (KSS), and a simpler but less informative measure, the Mean Opinion Score, are introduced.

2.6.1 Flow

When Csikszentmihalyi studied the creative process of artists and intrinsically motivated activities of chess players and athletes in the 1960s, he found, that when their work was going well, they would single-mindedly persist and ignore hunger, discomfort, and fatigue for extended periods of time. This led him to develop the concept of *Flow* which he considered to be an equilibrial state between boredom and anxiety, and between requirements and capabilities (skills).



(a) Original model published in [36].

(b) Revised model from [97].

Fig. 2.2 Flow models according to Csikszentmihalyi (a) and Nakamura (b).

In [36], Csikszentmihalyi defined Flow as follows:

> "Poised between boredom and worry, the autotelic experience is one of complete involvement of the actor with his activity. The activity presents constant challenges. There is no time to get bored or to worry about what may or may not happen. A person in such a situation can make full use of whatever skills are required and receives clear feedback to his actions; hence, he belongs to a rational cause-and-effect system in which what he does has realistic and predictable consequences. From here on, we shall refer to this peculiar dynamic state – the holistic sensation that people feel when they act with total involvement – as *flow*."

In the original model, Flow was illustrated as a channel between boredom and anxiety (cf. Figure 2.2a), where action opportunities or challenges are met by capabilities, while both are at above average levels for the individual [36]. It was subsequently shown, that the resolution of the phenomenological map can be improved by subdividing the space into eight experiential channels (cf. Figure 2.2b) where the intensity of the experience intensifies within

a sector when challenges and skills move away from the person's average levels represented by the center of the circles [97].

The concept was later embraced for gaming and used to describe differing ideal zones in such a phenomenological map of capabilities and challenges for novice and hardcore players, where the flow area for more experienced and skilled players is shifted slightly upwards in the challenges dimension in comparison to down-shifted flow areas for less trained beginners [28]. Chen furthermore proposed that games may algorithmically adapt challenges to the player's skill and individual flow zone, to facilitate a flow experience for the player [28].

To measure the degree of flow experience and its aspects with a preferably short interruption of the task at hand, the "Flow-Kurzskala" (Flow Short Scale) was developed by Rheinberg *et al.* It is a 10-item questionnaire employing 7-point Absolute Category Rating (ACR) scales to assess the flow experience QoE feature immediately after or while conducting the according activity. The scale was used in numerous gaming studies, e. g., to assess the difference between human- or computer-controlled opponents [122], or to study the relationship between flow and immersion in a role-playing, a racing, and a jump and run game [121].

2.6.2 Immersion

According to Brown *et al.*, immersion describes the degree of involvement with a game. Following interviews with gamers as part of a qualitative study, they distinguish three stages of immersion called *Engagement*, *Engrossment*, and *Total Immersion* which *can* come after each other when the barriers to each level are removed [26].

The lowest level of immersion, *Engagement*, requires gamers to invest time, effort and attention besides needing to have access to the game in the first place. As players become further involved with the game and its "features combine in such a way that the gamers' emotions are directly affected by the game", they may enter the *Engrossment* stage and become "less aware of their surrounding and less self aware than previously". Finally, with *Total Immersion* "the game is the only thing that impacts the gamer's thoughts and feelings", a stage which requires the game to have an 'atmosphere' made by graphics, story, and sound elements, and the player to be able to empathize with a character or team in the game [26].

A player's immersion can either be measured using a purpose-built questionnaire by Jennett *et al.* [72], or using a set of items from Game Experience Questionnaire described in the following Section 2.6.3.

2.6.3 Game Experience Questionnaire

The Game Experience Questionnaire is a modular self-assessment questionnaire to "comprehensively and reliably characterize the multifaceted experience of playing digital games" [100], which is integrated into the lower layer of the taxonomy introduced in Section 2.3. The questionnaire consists of three modules: core questionnaire, post-game questionnaire, and social presence module. All three modules are intended to be administered directly after a gaming session.

The core questionnaire contains 36 items plus 6 additional spare items for 'translation purposes'. Each of these 42 items is related to one of seven dimensions of *Player Experience*. These dimensions are:

- *Sensory and Imaginative Immersion* - cf. Section 2.6.2
- *Tension* relates to emotional strain connected with attributes like feeling tense, pressured, or restless.
- *Competence* refers to having the skill, knowledge, and ability to reach the game's targets.
- *Flow* - cf. Section 2.6.1
- *Negative Affect* concerns unfavorable facets like boredom or distraction.
- *Positive Affect* refers to pleasant aspects of gaming experience such as fun or enjoyment.
- *Challenge* involves feeling the requirement to put effort into the game because the tasks are considered difficult.

Persons filling the questionnaire have to decide how much they agree with each statement (i. e., item) and rate this on a 5-point ACR scale labeled *not at all*, *slightly*, *moderately*, *fairly*, and *extremely*. To compute the respective values for the seven dimensions, the participants' answers to each related item are averaged using an arithmetic mean. The averaged ratings constitute the GEQ dimensions. Due to the questionnaire's core part's sizable nature, a shortened version called *In-game Questionnaire* is proposed [100] to be used during short interruptions of the game-play. It measures the same seven dimensions as the full core questionnaire, but is limited to two items per dimension, resulting in a total of 14 items.

The post-game questionnaire concerns players' feelings after they have stopped playing. It consists of 17 items related to four dimensions termed *Negative Experiences*, *Positive Experiences*, *Tiredness*, and *Returning to Reality*. While the first three are named quite

self-explanatory, the last refers to the difficulty of getting back to reality and associated disorientation.

A number of studies have used the GEQ successfully to, e. g., investigate the influence of social context [44], game level design modifications [94], or the use or non-use of sound and music [95].

2.6.4 Self-Assessment Manikin

Developed and published by Bradley *et al.*, the Self-Assessment Manikin (SAM) is a non-verbal pictorial assessment questionnaire to measure QoE aspects called *pleasure*, *arousal*, and *dominance* of a person's affective reaction to a presented stimulus [24].

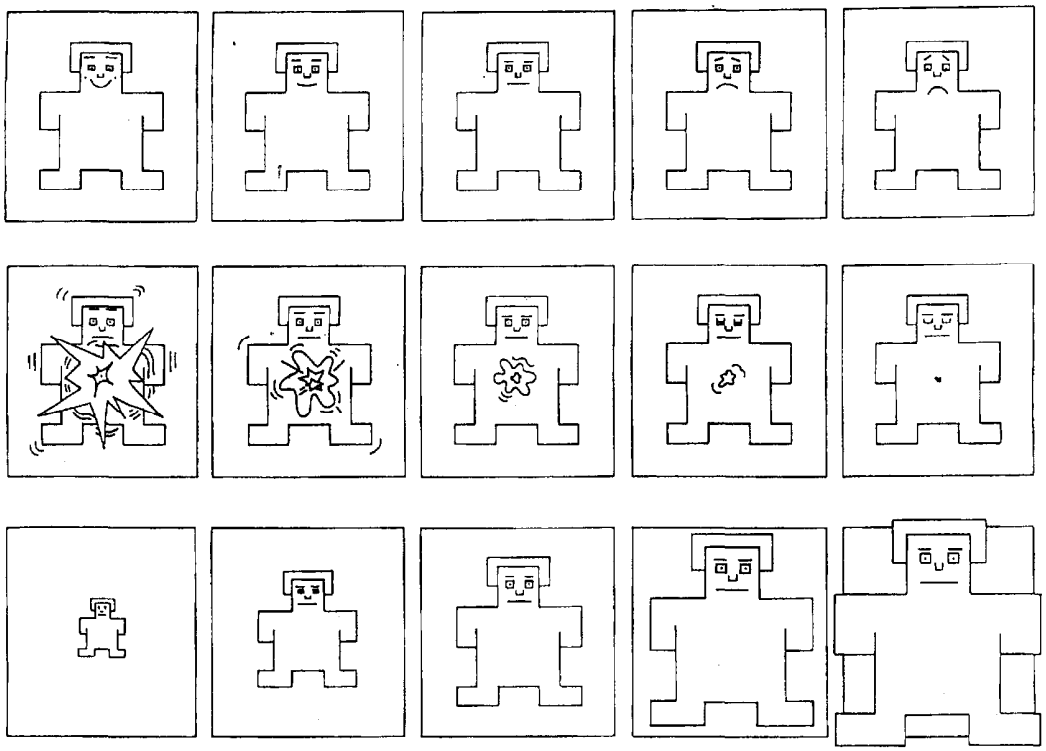


Fig. 2.3 Pictorial scales of the Self-Assessment Manikin used to rate the affective dimensions of valence (top), arousal (middle), and dominance (bottom) [24].

The questionnaire consists of three scales depicting a horizontal array of sketched 'manikins' showing visible emotional signs related to the respective dimensions (cf. Figure 2.3). The first of these scales, measuring the dimension *Pleasure*, is related to attributes like happiness, satisfaction, and relaxation. The second dimension, *Arousal*, refers to aspects such as stimulation, excitement, or feeling wide awake. It describes the perceived vigilance as a physiological and psychological condition of a person. The range reaches from excitation

to doziness or boredom. *Dominance*, the last dimension, concerns feeling in control versus being controlled, or feeling influential vs being influenced. This describes how much a person feels in control of a situation. A small manikin corresponds to a subject's feeling of having no power to handle the situation. Although the SAM was initially published as a 5-point scale, 7- and 9-point variations have also been created and published on the web[10].

The SAM was used successfully in the context of gaming to, e. g., measure the emotional appeal of a Tetris game with varying levels of difficulty [27], or to investigate the game play experience of elderly people [96].

2.6.5 Karolinska Sleepiness Scale

The Karolinska Sleepiness Scale (KSS) is a verbally anchored 9-point scale used to subjectively rate sleepiness, which, following the taxonomy in Section 2.3 is a dynamic player attribute. Of these 9 points, five are labeled as follows, while the steps between remain without text: *extremely alert* (1), *alert* (3), *neither alert nor sleepy* (5), *sleepy—but no difficulty remaining awake* (7), and *Extremely sleepy—fighting sleep* (9). [3]

With the scale it becomes feasible to easily monitor study participants' wakefulness state, as tiredness may interfere with cognitively demanding tasks, leads to slower reaction times, and causes participants to make more mistakes [75]. In games, these effects could distort the results as they might allow less success in games and therefore may increase frustration.

Although the KSS is essentially measuring a dynamic player attribute as stated above, it may also be considered as an indirect performance metric if it is repeatedly applied as done in the study in Section 7.2. There, the repeated application of this questionnaire is employed to appraise potentially tiring effects of high cognitive load caused by very bad visual quality.

2.6.6 Mean Opinion Score

The Mean Opinion Score (MOS) is an established measure for the assessment of the average subjectively perceived quality (i. e., the "opinion") of a system. In contrast to the GEQ or the SAM, the MOS is a one-dimensional overall quality rating. The score was originally developed for the assessment of transmission quality of telephone equipment and standardized for that purpose in ITU-T Recommendation P.800 [66] as the five-point ACR 'Listening-quality scale' shown in Table 2.1. Although ITU-T Recommendation P.800 makes no recommendations on the exact procedure and layout to be used to obtain the participants' ratings using this scale (e. g., paper-and-pencil-based, computer-form-based, etc.), examples presented in the document for other scales hint that a computer-based approach was intended,

[10] http://irtel.uni-mannheim.de/pxlab/demos/index_SAM.html (last accessed: 2016-05-12)

Table 2.1 Listening-quality scale as defined in ITU-T Recommendation P.800 [66] used to obtain subjective ratings from which the MOS can be calculated.

Quality of the speech	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

i. e., that participants press a labeled button after listening to a stimulus. However, both in paper- and computer-based questionnaires, a horizontal tabular display is common (cf. [90]). When subsequent stimuli grow worse (or better) in quality, this 5-point scale can suffer from saturation at its extremes. Furthermore, it allows participants only to provide a coarse answer due to the limited options available with no means to provide fine-grained answers between two categories such as *fair* and *good*. To mitigate these effects, an extended continuous rating scale has been proposed by Bodden *et al.* (cf. Figure 2.4), which carries the same five labels as the listening-quality scale shown in Table 2.1, but adds two items in the extremes labeled "extremely bad" and "ideal".

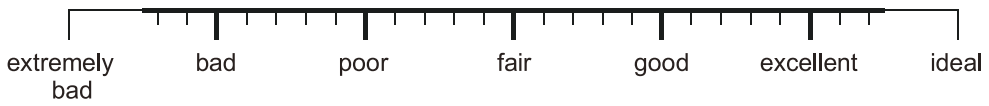


Fig. 2.4 Continuous rating scale after Bodden *et al.* [22] and Möller [90] labeled according to ITU-T Recommendation P.800 [66] with the addition of the labels "extremely bad" and "ideal" at the extreme ends of the scale.

Finally, the arithmetic mean of all ratings on these scales is called Mean Opinion Score (MOS). The ITU-T later adopted the same scale for use in the assessment of audiovisual (P.910 [68]) and video quality (P.911 [69]). Furthermore, the scale was embraced in multiple studies for the assessment of subjective gaming experience (e. g., [56, 118, 120, 125]), as performance metrics alone are insufficient to describe the quality of a gaming setup [108].

Jarschel *et al.* measured the perceived degradations of network delay and loss in a simulated cloud gaming environment, where the entire game execution is taking place on a remote server and only a video stream of the game is transmitted to the player over the network. While using the MOS to assess the quality of the system, they noted that the study design left it open to the participants to decide which aspects of the playing experience they valued most in their ratings [70]. Due to this ambiguity, the MOS in itself is a less informative

metric than multi-dimensional constructs of player experience such as the SAM or GEQ. It can still be used meaningfully as a generic summary metric in combination with other more specific questionnaires as it may cover further unexpected aspects of a participants experience due to its universality.

The MOS is frequently used as a target variable for modeling a system's perceived overall quality under variations of influence factors. Popular examples are the 'E-model' [65] for predicting the conversational quality of 3.1 kHz handset telephony, and the T-V-model for predicting IPTV quality [103]. In the domain of gaming, multiple models have been created to predict the quality of a game streamed in a cloud gaming setup [112, 114, 119]. These are discussed in more detail in Section 5.2.

2.7 Physiological methods

As self-assessment methods like questionnaires inherently place an additional burden on test subjects and interrupt the actual game experience, researchers are working on identifying physiological correlates with experience dimensions to obtain non-interruptive and continuous measures. As an example, the electroencephalography (EEG) has proven to be a valuable tool for research in the auditory and visual domains, as it can provide additional information about underlying processes [5, 9]. In the terms of the taxonomy in Section 2.3, physiological methods measure performance metrics. However, these particular metrics may be strongly linked to subjectively experienced QoE features.

EEG measures voltage changes due to brain activity by attaching electrodes to the scalp of a participant. Since Berger developed the EEG in 1929, it has been widely used for research of physiological correlates of perceptual and attentional processes [13, 40]. EEG data can mainly be analyzed in two different ways: on the one hand, by looking at the Event-Related Potentials (ERP), which are a time-locked reaction to an external stimulus measured as a change in voltage, and on the other hand, by taking a closer look at the spectrogram of spontaneous (not event-related) activity [107]. With respect to the latter, there are five different frequency ranges ascribed to specific states of the brain [107]: delta band (1–4 Hz), theta band (4–8 Hz), alpha band (8–13 Hz), beta band (13–30 Hz), and the gamma band (36–44 Hz). Activity in the delta band is mainly present during sleep, theta band activity during light sleep. Activity in the alpha band is related to relaxed wakefulness and to situations of decreased alertness. High arousal and focused attention lead to a high power in the beta and gamma bands [107].

2.8 Subjective assessment of gaming experience

Unlike other domains where the quality of an item or a system can be measured through, e. g., simulations or instrumental measurements, the assessment of subjective gaming experience requires having persons play games. Following ITU-T Recommendation P.911 [69] and the tradition of other domain-specific standardized test paradigms, to obtain both internally and externally valid results requires defining a set of experimental procedures, measures, and reference parameters.

While some parts of ITU-T Rec. P.911 such as the number of required test participants, guidelines on how they are to be instructed, reference viewing and listening conditions, and even some recommendations regarding the statistical analysis and result reporting may be applied to gaming, other parts such as the methods recommended for stimulus presentation and rating are inappropriate as gaming is an interactive process as opposed to the merely passive multimedia consumption considered in ITU-T Rec. P.911. Instead, participants of gaming studies are actively interacting with the system under test. As games used in a study may be unfamiliar to player, they are typically allowed to learn the game at the beginning and get used to the controls and game mechanics. Afterwards, the test conditions with variations of the influence factor(s) studied in the experiment are played in a consecutive manner. While the ACR method in ITU-T Rec. P.911 recommends stimuli lengths of around 10 seconds, this is unlikely to be enough for gaming. However, no consensus exists for the necessary duration of a stimulus to allow participants to experience, e. g., flow or immersion, and it is likely that this minimum duration depends on both the particular game and the factors varied in the test.

Another difference to experiments involving passive media consumption is that games may not end after the predetermined condition duration, requiring the on-going game session to be interrupted. As such interruptions may by themselves cause emotions, they have the potential to skew the players' experience. The most common method to measure this experience is through subjective self assessment using questionnaires such as those presented in Section 2.6. After this measurement, the condition is concluded and another may follow.

While the previously described test procedure in its basic form may be common to many gaming studies, it is not standardized in many facets, leading to different stimulus times, training phases, measurement methods, etc. The influence of these procedural aspects is, however, largely unknown and further work is necessary to establish a commonly applicable test paradigm. To coordinate efforts and facilitate collaboration, the ITU-T Study Group 12

has created a work item called P.GAME[11] with the goal of developing a set of test procedures which allow different labs' results to be truly comparable (cf. [91]).

2.9 Conclusion

In this chapter, the term *Game*, and the two perspectives on gaming experience *Quality of Service*, and *Quality of Experience* were defined. Founded on these, metrics and measurement tools for both objective and subjective experience were presented. In the following chapters, these means will be used to study the relationships between major influence factors and QoE features in mobile gaming.

[11] http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=9992 (last accessed: 2016-06-21)

Chapter 3

Influence of the game

3.1 Introduction

To investigate player, system, and context influences on a player's perceived quality, the theoretical ideal game would be a perfectly neutral one: a piece of software that could predictably excite the exact same emotional response as often as necessary, and that, at the same time, was representative for all imaginable games. Such a game would be well-balanced, in that all conceivable emotions could be raised and all possible kinds of input (e. g., touch screen, game-pads, joystick, device tilting, ...) and output (e. g., 2D, 3D, Virtual Reality (VR), ...) could be used.

Such a game cannot exist. The choice of games is therefore a pre-eminent question in the research of gaming experience. Digital games are complex multi-layered products with highly refined user (or rather: player) interfaces, typically employing plentiful artwork, and a variety of algorithms and rules to bring the interface to life. On each of the implementation's layers, a game producer possesses a great degree of freedom in how to achieve and implement a certain effect or behavior. As a consequence, games not only look, sound and behave differently (by design), but they may also react differently to changes of the execution environment they run in.

One of the aspects of a smartphone or tablet used for mobile gaming, is its network connectivity and the particular transmission channel parameters of, e. g., bit rate, packet loss, and delay. For multi-player games, which have to provide multiple players using different devices with a synchronized view of a shared virtual gaming environment, the network delay and mechanisms to hide its presence are of predominant importance. However, there is no uniform way to implement this concealment.

To find out how different games are subjectively experienced by players with varying network implementations, a study was conducted, in which test participants played three

different mobile multi-player games. To investigate the interplay of network delay and the games' implementations, the transmission delays in a simulated network between two players were varied. Not surprisingly, the games were perceived differently. Furthermore, the results illustrate, that substantial differences exist in the range of acceptable latencies and that games' network implementations vary strongly in their ability to gracefully alleviate the effects of network delay.

The results from this chapter's study were previously published in [17].

3.2 Related work

The influence of network parameters on specific games has been subject to research for a considerable time.

In 2001, Armitage [7, 8] monitored ping times of gamers playing the fast-paced first person shooter Quake III Arena on two public gaming servers in the United States. He found that, in the distribution of the players' network delays, the majority of active players had round trip latencies of less than 150 ms and only a fraction above that level. He concluded that ping times of 150 ms and above were not tolerable and gamers would rather switch to a different server with a lower latency. A similar study [49] using the first person shooter Half Life found the majority of players having ping times below 300 ms.

When Pantel and Wolf [99] investigated the effects of network delay on two commercial racing games in 2002, they observed that the delayed transmission of status updates often led to inconsistent states between the players: E.g., with two cars having started at the same time and performing the same accelerations, the local car would seem to be in the lead instead of being side-by-side to the opponent due to the delayed status reception of the other car's position over the network. In another comparable racing game, they artificially added a local delay to establish a synchronized state between players despite the network delay and measured performance metrics like the average time per round, best times, and the frequency of the racing car's departure from the track for different overall delays. They found that all three metrics increased with rising delay and concluded, based on participants' statements, that an overall delay (between input and game reaction) of 500 ms and more is not acceptable for a racing game.

However, similar to the cited works, most previous research focused on non-mobile gaming. While Schaefer *et al.* [108] and Wang *et al.* [118] actually investigated mobile playing, the games they used were stationary games adapted or streamed to mobile devices. These are, however, not representative of typical mobile games found in Google's or Apple's app stores, as these titles are specifically developed with touchscreen-interaction and the

smaller form-factor in mind. Consequently, a gap in research exists, which the present study may help to close.

3.3 Methodology

To investigate the interplay of network delay and game implementations, a study with test participants was conducted, in which three different Android mobile multi-player games were played over a controlled network with varied transmission delays.

3.3.1 Selection of games

Three Android mobile games from Google's Play Store were chosen for the experiment, namely *MiniMotor Racing*, *Curve Mania*, and *Blobby Volleyball*. In contrast to other studies, it was decided not to implement new games, as the complexity and necessary time investment for game developments, that are comparable in quality to well-polished commercial-grade apps, is disproportionate considering the scope of these studies (cf. [21]). The primary selection criterion was the games' ability to facilitate multi-player gaming within a local Wireless Local Area Network (WLAN) without the need for external servers, so the games could be tested in an isolated laboratory setup. The second criterion was the games' degree or frequency of interaction with the opponent: Whereas racing games such as MiniMotor Racing involve situations, in which one player tries to cut the other player's path in order to move his own car ahead of him, they happen not very frequently, whereas most of the time, one player follows their opponents car at a varying distance without direct interactions between the two. Other multi-player games feature much more frequent direct interactions. Curve Mania and Blobby Volley are examples of this category as illustrated below.

MiniMotor Racing

MiniMotor Racing[1] is a classical racing game where the player has to drive a car through a racing course faster than the opponent. Displayed in the lower part of Figure 3.1, the player can control his vehicle with three buttons: "L", to steer to the left, "R" to steer to the right, and "Nitro" to accelerate. A session with the game consisted of steering the car through the lap five times. The player, who achieved the lowest overall time, won the race.

Fig. 3.1 Screenshot from the game MiniMotor Racing.

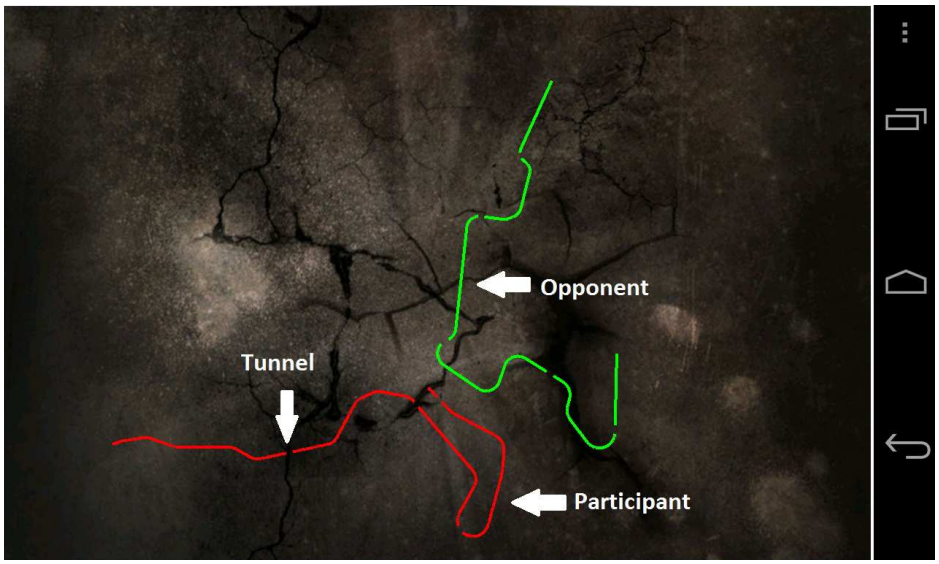


Fig. 3.2 Screenshot from the game Curve Mania.

Curve Mania

In the TRON-style, real-time networked multi-player game Curve Mania[2] each player steers one constantly moving colored dot on an otherwise dark screen. While moving, the player's and his opponent's dots draw lines on the screen. The player, who first drives his dot into either the other player's track line or the edges of the screen, loses the game. As one player

[1]https://play.google.com/store/apps/details?id=com.nextgenreality.minimoto (last accessed: 2016-04-13)
[2]https://play.google.com/store/apps/details?id=com.ratcash.games.curvemania (last accessed: 2016-04-13)

can win the session by drawing his line in such a shape that the opponent grows short of space to navigate in, and therefore inevitably has to drive his dot into the other line, this is the predominant strategy for winning this game. In practice, this frequently leads to situations where the two players' dots move along side by side with one player trying to cut in in front of the other, upon which the opponent also has to change course in a timely manner to evade moving his dot into the opponent's line. These are situations, in which precise timing is required, and therefore a high degree of interaction between the players exists.

Blobby Volleyball

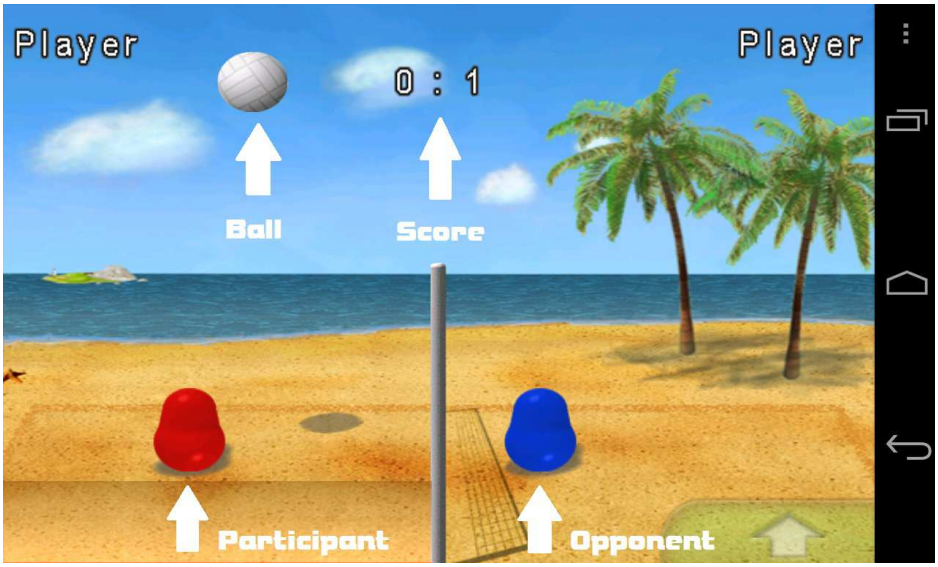


Fig. 3.3 Screenshot from the game Blobby Volleyball.

The third game, Blobby Volleyball[3], is a dynamic arcade sport game of volleyball as shown in Figure 3.3. As in the real game of volley ball, the player scores a point when the ball hits the ground in the opponent's part of the field. Touching the screen in the lower part of the player's own field will move his figure, while touching the upper part will make the figure jump. The ball is played by moving or jumping the figure with the intended angle at the ball. If the ball is touched by the figure more than three times, the opponent is awarded a point. The first player earning 10 points wins the match. Since the ball moves between the opponents' fields and the way of playing the ball influences its trajectory, upon which the opponent has to quickly react, interactions between the players happen very frequently.

[3]https://play.google.com/store/apps/details?id=com.appson.blobbyvolley (last accessed: 2016-04-13)

3.3.2 Network simulation

To allow the multi-player games to establish a transmission channel between the players' devices, both smartphones had to be within a common broadcast domain in the same network. Sinces a wired network connection is neither supported by the devices' hardware, nor is it a realistic use-case with a cable limiting the player's freedom to handle the device, a wireless link had to be used. However, whereas a cable network is shielded from external influences and interferences and offers simultaneous bi-directional data flow (full-duplex), a WLAN following the standard 802.11n [57] is susceptible to packet loss due to interferences caused by other users of the same unlicensed and therefore freely usable part of the spectrum, and can transmit data only in one direction at a time (half-duplex). Furthermore, WLAN is a shared medium: Of all devices operating in a given spectral band, only one can (successfully) transmit data at a time. If multiple stations send concurrently, their transmissions collide and the contents of the communications are lost. Although 802.11 defines a mechanism to minimize these collisions (Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) [58]), a low degree of packet loss is inevitable as long as multiple parties share the same part of the spectrum. To prevent messages originating from the smartphones from colliding, two separate access points (Apple Airport Express 2012[4]) were installed.

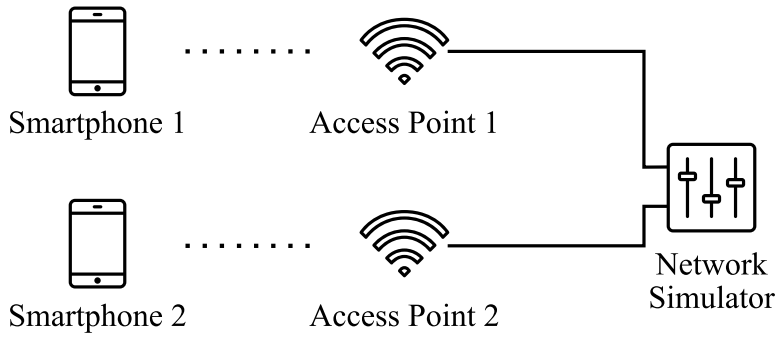


Fig. 3.4 Illustration of network setup using two separate WLAN access points linked by a network simulator to prevent interferences.

To minimize interferences from other users of the same spectral band, two otherwise unused channels in the 5 GHz part of the WLAN spectrum were chosen. The access points were configured as layer 2 network bridges, simply copying wireless communications to a wired network and vice versa. Each of the access points was then connected to a separate network interface in a PC (Intel Core i3-2120 3.3 GHz, 4 GB RAM, 120 GB Solid State Drive (SSD), Intel Server Network Adapter I350-T2) acting as a network simulator as illustrated in Figure 3.4. On that PC, Debian Linux 7.0 was installed and the two network

[4] http://www.apple.com/airport-express/specs/ (last accessed: 2016-04-11)

interfaces were configured as a bridge, causing data from one access point being forwarded transparently to the other. Delays in the forwarding of packets between the networks were then introduced using the Linux 'netem' network emulator kernel module [48].

3.3.3 Simulated parameters

As part of a pre-test, suitable ranges of delay were identified individually for each game so as to span from unnoticeable to strongly perceivable but still playable. In these ranges, four delay levels were chosen. This resulted in 500 ms, 1000 ms, 4000 ms, and 6000 ms delays for MiniMotors, 100 ms, 250 ms, 500 ms, and 1000 ms for Curve Mania, and 100 ms, 250 ms, 500 ms, and 2000 ms delays for Blobby Volleyball. While the range of 100 ms to 6 seconds might be considered as unrealistic for fixed Internet connections (e. g., DSL, Cable, etc.), which usually provide relatively constant transmission delays, these frequently occur in mobile networks during handovers between transmission technologies (such as WLAN to UMTS, or UMTS to GPRS in areas with poor coverage) and overload situations (e.g., underground public transportation during rush hours): "Vertical handovers between GPRS, WLAN and LAN [...] last from 200 ms up to several seconds, which is suitable for reliable flows but can be a problem for real-time flows" [47].

3.3.4 Measurements

To gather the test participants' impressions of the presented conditions, two questionnaires were used. The first part comprised a perception questionnaire with three items as seen in Figure 3.5. Whereas the first two items were created for this study to assess common gameplay degradations caused by the simulated network impairments, the latter item is used according to [22] and [67]. As stated in Section 2.6.6, the mean of all these overall quality ratings is referred to as MOS.

Afterwards, the full 42-item core module of the GEQ was presented to assess the effect of the varied network parameters on the participants' Player Experience.

3.4 Test procedure

The study took place in June 2013 at Quality and Usability Lab, Technische Universität Berlin in a lab environment with quietness and neutral lighting, fulfilling the requirements for audio-visual quality rating tests specified in ITU-T Rec. P.910 [68] and P.911 [69]. As the test plan followed a within-subjects design, each participant played all conditions.

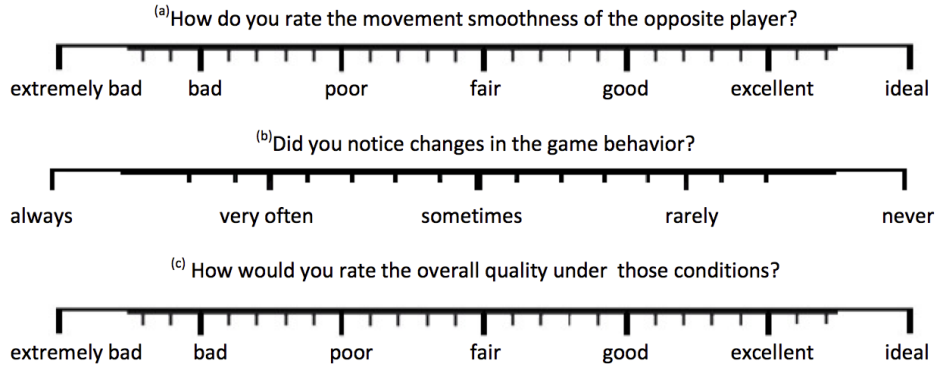


Fig. 3.5 Perception questionnaire with three items rated on a continuous rating scale to elicit perceived gameplay degradations caused by the artificially impaired network. The latter item is used according to [22] and [67].

In that room, the test participants sat on a comfortable chair at a desk, upon which a prepared smartphone, a 4.7-inch Google Nexus 4 device running the Android 4.2.2 "Jelly Bean", was placed. After they had sat down, the participants filled a demographic questionnaire and were introduced to the device and the games used in the experiment. No instructions were given on how to hold the smartphone during the test, but the persons could use the device as they deemed adequate and felt comfortable with.

The subsequent playing test consisted of three blocks. Each part began with a training session with the game under test without delay and was followed by four test sessions with varied levels of delay. The assignment, and therefore the order of the delay levels, was randomized. During each gaming session, the participants played against an experimenter, who was hidden from them, so that the playing conditions were kept approximately constant and comparable between experiments. The duration of each test session depended on the played game (e. g., time to finish five laps of racing in MiniMotor Racing), but was generally around three minutes. As a condition ended when, e. g., a race was won, the participants' gaming was not interrupted. After the end of a condition, the participants filled the questionnaire and continued with the next session.

The study was conducted with 19 casual gamers (less than 10 hours of playing time per week) of which 10 were male and 9 female. Ages ranged from 21 to 31 with a mean of 24 years. All were experienced in using mobile devices for gaming. Their average playing time per week was 2.2 hours. Of the 19 participants, 6 were already familiar with the PC-version of the game Blobby Volleyball but had not played it on a mobile before.

3.5 Results

In the following sections, error bars indicate the 95% confidence interval. GEQ items were coded with the values 1 = “Not at all” to 5 = “Extremely”. The GEQ’s Player Experience dimensions were calculated from these items according to [100]. The continuous scales used for perceptual measurements were coded in 0.2-intervals as 1.0 = “Extremely bad” to 7.0 = “Ideal” for the perceived smoothness of the opposite player and the overall quality. Finally, the item on noticed changes in the game’s behavior was mapped as 1.0 = “never” to 5.0 = “always”.

Each condition’s ratings were individually tested for normality using a Shapiro-Wilk test with a significance threshold of 0.05. Only a very small set of ratings differed significantly from a normal distribution according to the test. As observed deviations were only small, parametric statistical analyses were used in the following nevertheless, as the Analysis of Variance (ANOVA) was shown to be robust against minor violations of its normality assumption [109].

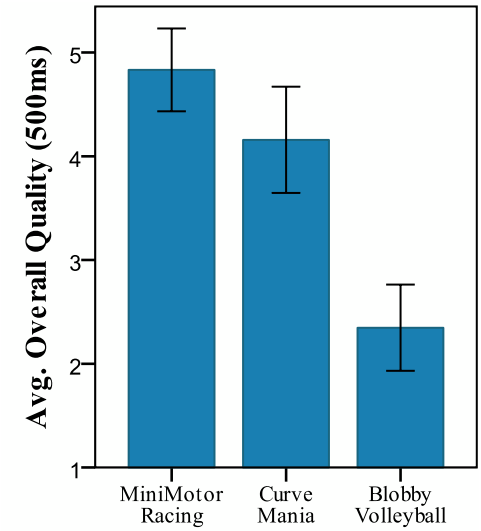
3.5.1 Overall comparison of the games

As the ranges of network delay levels, which were chosen for the games, vary between them, a direct comparison is limited to the common 500 ms setting. At that delay, the games’ overall quality item was rated very differently, as shown in Figure 3.6a. When using an ANOVA for repeated measures with a Greenhouse-Geisser correction, the mean scores for the overall quality item were statistically significantly different $(\mathrm{F}(1.663, 29.930) = 38.383, p < .001, \eta^2 = .681)$.

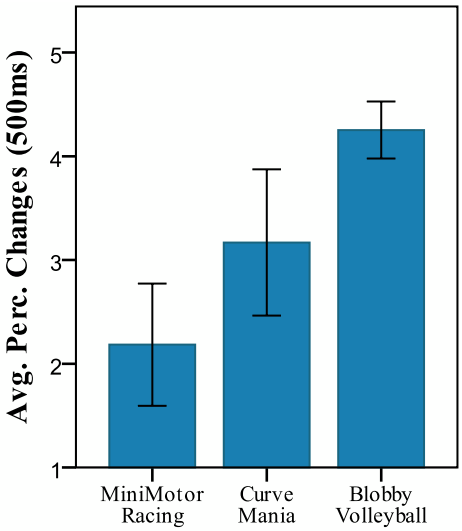
The perception of changes in the games’ behaviors also differed strongly, as seen in Figure 3.6b. Again, using the same repeated measures ANOVA as above, the observed differences in the mean ratings were statistically significant $(\mathrm{F}(1.698, 30.561) = 15.526, p < .001, \eta^2 = .463)$. Also, the perception of smoothness in the games’ actions differed as shown in Figure 3.6c. As with overall quality and perceived changes, this effect was significant $(\mathrm{F}(1.874, 33.727) = 32.602, p < .001, \eta^2 = .644)$. For all three overall quality, perceived behavioral change, and smoothness, the game Blobby Volley differs significantly $(p < .001)$ from MiniMotor Racing and Curve Mania, whereas between the latter two no significant difference was observed, although a trend is visible in Figure 3.6a and Figure 3.6b.

For the Player Experience dimensions from the GEQ as shown in Figure 3.7, changes are significant only for:

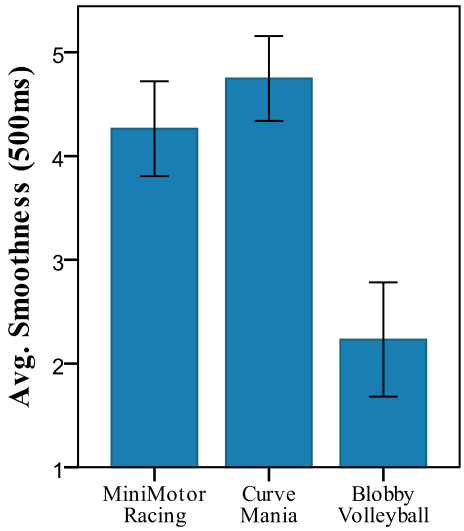
- Immersion $(\mathrm{F}(1.976, 35.573) = 10.745, p < .001, \eta^2 = .374)$,



(a) Overall quality
(1: “extremely bad” - 7: “ideal”).



(b) Perception of changes in the game behavior
(1: “never” - 5: “always”).



(c) Perception of the opponent’s smoothness
(1: “extremely bad” - 7: “ideal”).

Fig. 3.6 Ratings for the perception of the games at a simulated network delay of 500 ms.

- Tension ($F(1.856, 33.305) = 7.446, p < .01, \eta^2 = .293$), and
- Positive Affect ($F(1.747, 31.442) = 3.511, p < .05, \eta^2 = .163$).

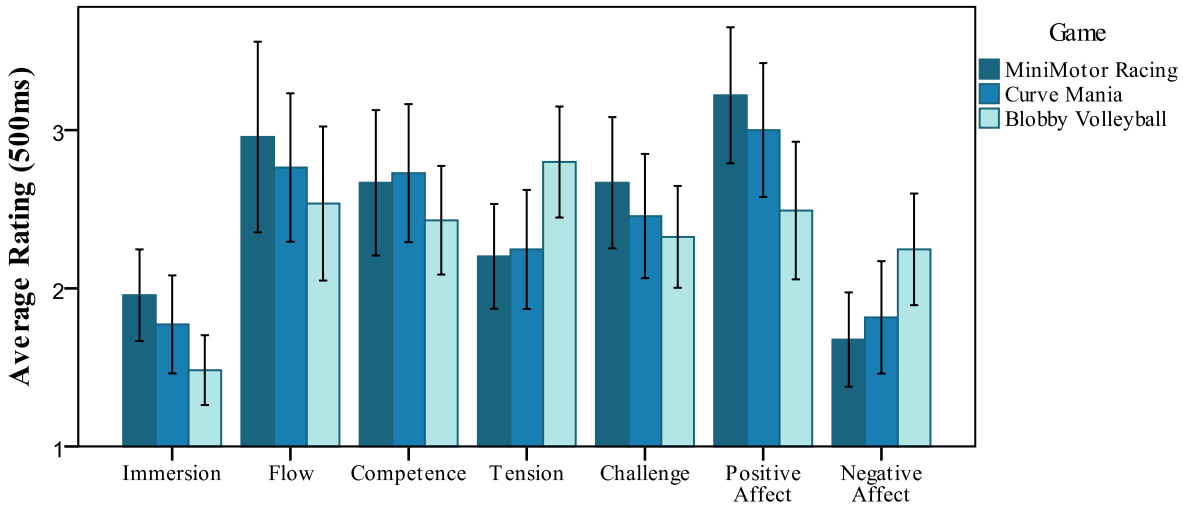


Fig. 3.7 Player Experience dimensions for the three games at a simulated network delay of 500 ms.

3.5.2 Influence of delay change

Due to the different ranges of delays, the absolute ratings cannot be compared between games. In the following, the user-perceivable effects of the varied delay are therefore presented on a per-game basis.

MiniMotor Racing

Although the range of tested delays was the broadest among the tested games with 5.5 seconds, significant effects could neither be found in the perception ratings (overall quality, change, smoothness, cf. Figure 3.8), nor in the Player Experience dimensions. In the game, the delay resulted in a delayed start and a time-shifted display of the opposing player. This was noted by multiple test participants when asked whether they had perceived changes in the game play:

> *I noticed that the player was starting a few seconds after me.* - Participant 4

> *It is hard to judge because I couldn't see the car. I was too fast! My opponent started every race some time after me.* - Participant 6

> *Why was I always starting as the first one?* - Participant 10

> *I had the feeling that the opposite player started the car deliberately later than I did.* - Participant 14

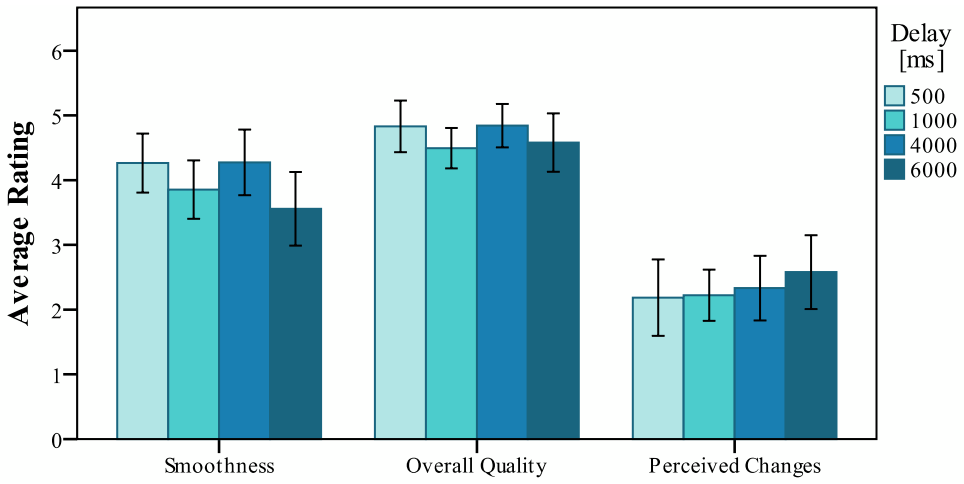


Fig. 3.8 Perceived smoothness and overall quality (1: “extremely bad” - 7: “ideal”), Perceived changes in gameplay (1: “never” - 5: “always”) for the game MiniMotor Racing with simulated network delays of 500 ms, 1000 ms, 4000 ms, and 6000 ms.

The time-shifted display also led to situations, where participants could see their own car crossing the finish line first, and still be informed by the game that they had lost the race, as they had completed the laps slower than the opponent.

Curve Mania

In the TRON-style game Curve Mania delay significantly affected perceived overall quality ($F(2.456, 44.207) = 4.349, p < .05, \eta^2 = 0.195$), and the perception of changes in gaming behavior ($F(2.358, 42.443) = 10.187, p < .001, \eta^2 = 0.361$) as shown in Figure 3.9. This influence was again tested using an ANOVA for repeated measures with a Greenhouse-Geisser correction. A weak trend can be seen for the perceived opponent’s smoothness but it does not reach significance levels.

However, none of the GEQ dimensions was significantly affected by delay, although a trend is visible in the Competence dimension in Figure 3.9. In this game, especially higher values of delay provoked multiple participants to note that they felt tricked by the other player, as these high delay scenarios would cause an asynchronous game state where, from the perspective of one player, the other could seemingly cross his line without losing, when, from the perspective of the other player, no crossing had yet occurred:

> *It’s impossible to win! The opposite player crossed the line multiple times and didn’t die. I also wanted to check if I am immortal.* - Participant 2

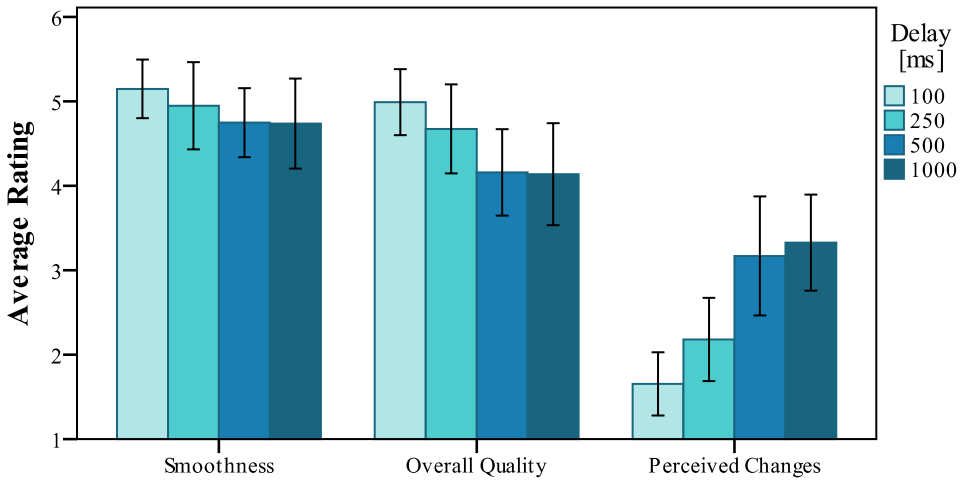


Fig. 3.9 Perceived smoothness and overall quality (1: “extremely bad” - 7: “ideal”), Perceived changes in gameplay (1: “never” - 5: “always”) for the game Curve Mania with simulated network delays of 100, 200, 500, and 1000 ms.

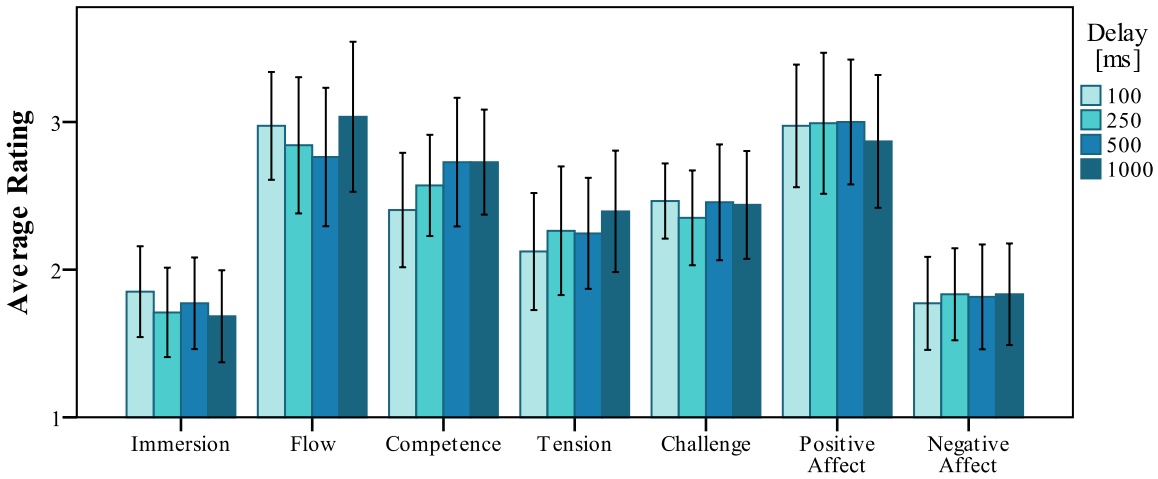


Fig. 3.10 Player Experience dimensions for the game Curve Mania played with simulated network delays of 100, 250, 500, and 1000 ms.

> *I suspect cheating. The opposite player could go through my line even though there was no escape hole in it. Once we were really close to each other and I thought I would win, but the game said totally the opposite!* - Participant 14

The game outcome was thus not correct from the player’s perspective. A total of 13 out of 19 participants explicitly mentioned having observed their opponent to have crossed their line without loosing, or that the game logic seemed to have changed.

Blobby Volley

In Blobby Volleyball, delay significantly influenced the participants' perception. An ANOVA for repeated measures with a Greenhouse-Geisser correction shows significant effects on smoothness of the opposite player ($F(2.5, 45.007) = 35.597, p < .001, \eta^2 = .664$), changes in game behavior ($F(1.795, 32.311) = 17.767, p < .001, \eta^2 = .497$), and overall quality ($F(2.454, 44.171) = 22.271, p < .001, \eta^2 = .553$) as depicted in Figure 3.11. Analyzing

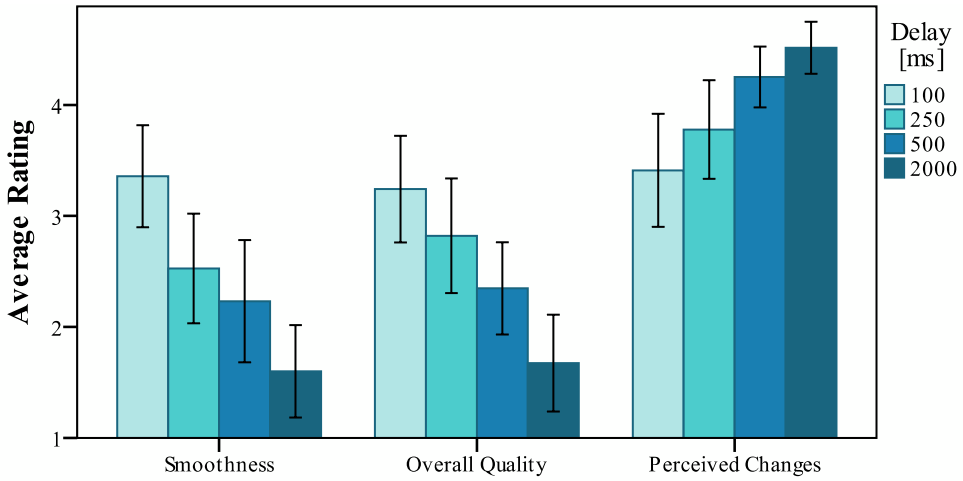


Fig. 3.11 Perceived smoothness and overall quality (1: "extremely bad" - 7: "ideal"), Perceived changes in gameplay (1: "never" - 5: "always") for the game Blobby Volley with simulated network delays of 100, 200, 500, and 2000 ms.

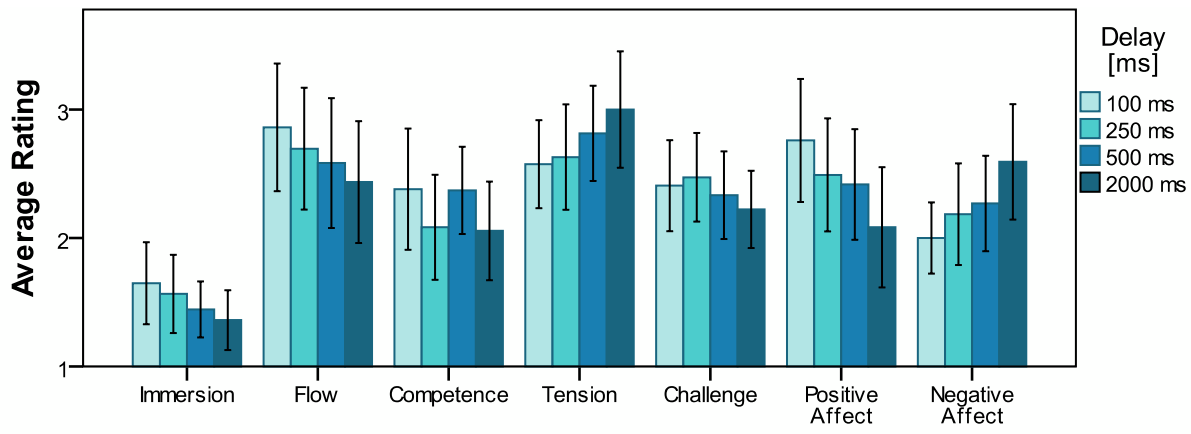


Fig. 3.12 Player Experience dimensions for the game Blobby Volleyball played with simulated network delays of 100, 250, 500, and 2000 ms.

the data using the same ANOVA as above, four out of seven Player Experience dimensions

as shown in Figure 3.12 turned out to be significantly affected by the increase in delay: Flow ($F(2.652, 47.730) = 3.574, p < .05, \eta^2 = .166$), Tension ($F(2.4, 43.198) = 4.293, p < .05, \eta^2 = .193$), Positive Affect ($F(1.918, 34.522) = 4.618, p < .05, \eta^2 = .204$), and Negative Affect ($F(2.426, 43.668) = 3.901, p < .05, \eta^2 = .178$).

Due to the "mechanics" of the game, delay led to situations, which were difficult to play due to frozen, or discontinuous movements of the ball, which also led to the perception of an unfair game:

> *The ball froze and landed at another area at random. The score didn't change accordingly.* - Participant 5

> *The ball teleported, touched the ground without giving me the point or came back to my side without the touch of the opposite player.* - Participant 6

> *The ball doesn't follow physical laws. Disappears and appears at random, counting of points doesn't work, the ball touched the sand and came back to play.* - Participant 18

3.6 Discussion

In the first part of this section, the games' performances with common network delay of 500 ms are compared, whereas in the second part the individual changes caused by the rise of the delay are discussed.

3.6.1 Comparison of game behaviors with common delay level

As expected, the three games in the test caused test participants to perceive the simulated network delay very differently. Whereas only infrequent changes in the gameplay were reported by the test participants for MiniMotor Racing in the 500 ms condition compared to the undelayed training session, the reported frequency was much higher for Blobby Volleyball as seen in Figure 3.6b. This is also reflected in the quality ratings for the games: Whereas the MOS of MiniMotor Racing was 4.8, which, on the rating scale, lies close to the label *good*, Blobby Volleyball's MOS of 2.3, is close to *bad* and therefore far worse.

At first glance, the games' different susceptibility towards delay might be solely explained by their differing degree of interactiveness between the participants in the multi-player session: Whereas in the racing game MiniMotor Racing a rise in network latency merely led to a delayed start of the opponent, which allowed the test participant to complete driving the laps rather undisturbed and without direct interactions with their competitor, the gaming

paradigm in Blobby Volley makes direct and frequent interactions between the players indispensable. Whenever the player's character in the game touches the ball, its direction and velocity change. This requires the other player to react accordingly to play the ball back and succeed in the game.

Although *Curve Mania* might also be referred to as a racing game, its game mechanics differ significantly from *MiniMotor Racing*. In this game, both players always share the same view of the game world. Contrary to the cars in *MiniMotor Racing*, which might get out of sight for the other player if the distance grows too big, each player's moving colored dot in *Curve Mania* is always visible to the other player. This continuous visibility and the resulting perceived competition might have led to the higher mean rating for *Curve Mania* in Figure 3.6c, when compared to *MiniMotor Racing*. Despite the existing possibility of evading direct opponent interactions by circling one's dot in different areas of the screen (cf. Figure 3.2), typical competitive sessions quickly lead to situations, in which one player tries to limit the other player's freedom in order to force him to involuntarily drive his dot into either one of the screen boundaries or the opponent's line and therefore loose the game. This causes a degree of interactivity between the players, which is comparable to *Blobby Volleyball*. As the reported breaks in the game logic (e. g., crossing lines, cf. Section 3.5.2) only occurred when both participants crossed the same position on the screen in a time frame shorter than the simulated network delay, it can be assumed, that most participants chose to play competitively and therefore closely interacted with their counterpart. The significant differences between *Blobby Volleyball* and *Curve Mania* are therefore most likely caused primarily by the way the games' implementations handled the transmission channel's latency. Whereas the display of the opponent in *Curve Mania* was merely delayed, yet smooth, the depictions of the ball and the opponent in *Blobby Volley* grew increasingly discontinuous and erratic.

Taken together, although both games were comparable in interactivity and subjected to the same degraded transmission channel, the user-perceivable implications of the delay varied substantially from another.

3.6.2 Comparison of game behaviors with changing delay levels

Comparing the progression of ratings for the games with increasing delay, strong differences can be seen. Consequently, the factor game appears to have a moderating effect towards delay and its perception by the player, as it co-determines the magnitude of network degradation's influence. This moderating effect is also raised by the particpants' ratings for the game *MiniMotor Racing*: Although the highest tested delay of six seconds is not entirely unrealistic in mobile networks, it is indeed very high for an interactive multi-player game. Nevertheless,

even in that most extreme condition, participants rated the game not significantly worse than in the lowest delay condition. In fact, not even a trend is visible in Figure 3.8.

For *Curve Mania*, a notably narrower range of network delays was simulated (100 ms to 1000 ms, cf. Section 3.3.3). Nevertheless, it led to a significant increase in perceived changes of the game's behavior, particularly due to arising issues in the game logic (cf. Section 3.5.2). Despite the majority of participants reporting what they observed as unfair behavior of their opponent, their quality ratings remained surprisingly high for all delay levels: The MOS fell only a marginal 0.85 points from 4.99 (*good*) to 4.14 (*fair*). A possible explanation for this phenomenon is, that the participants did not see the game itself at fault, but rather considered their opponent to be cheating, as noted in the comment by Participant 14 (cf. Section 3.5.2). If this finding should be substantiated in future studies, it would be an interesting analogy to the effect delay exerts in telephone conversations: There, an unfamiliar peer's delayed response is attributed to the person's personality, rather than the telephone system itself [110].

Blobby Volleyball was clearly the game reacting most sensitively to delay in the test. Not only was the drop in the MOS (3.2 down to 1.7) the most severe in all three games, the level of perceived changes in game behavior were also surprisingly high even in the lowest delay condition of 100 ms. While this level of delay might be high for a wired network, it frequently occurs in wireless networks under load. This unexpectedly high sensitivity to latency on behalf of the wireless transmission channel, and the participant's reports about multiple flaws in the game play lead to the impression that the game's implementation is not very well adapted to wireless networking environments. Yet, despite the game's irregularities, some players continued to find it fun to play, as can be seen by the surprisingly low drop in Positive Affect (cf. Figure 3.10) and players' written comments like these: "[...] It was fun" (Participant 17) and "[...] Funny game but hard to play" (Participant 13).

3.6.3 Limitations

Since the participants did not rate the games in an undelayed setting, it is not possible to clearly infer the latency's effect on the MOS and the GEQ Player Experience dimensions. Nevertheless, the participants were able to compare the games' behaviors to the undegraded performances as these were experienced in the training sessions.

Considering the high degree of perceived change for *Blobby Volleyball* even in the lowest delay condition (cf. Figure 3.11), it is possible that the entry-level delay was laid down to high in the pre-test. Except for the two written statements, it is therefore not possible to infer the participants general liking of the game from the ratings with delayed transmission, as already the lowest tested level introduced considerable changes into the game.

3.7 Conclusion

In this chapter, a study was presented, which investigated the moderating and shaping effect of multi-player games on the players' gaming experience in the presence of varying transmission channel delay. It was found, that the effect of delay strongly depends on the exact rules and implementation of the game. Whereas the least delay-sensitive game used in the test, *MiniMotor Racing*, was playable and well-rated even at the highest tested delay of 6 seconds, the most delay-susceptible game, *Blobby Volleyball*, showed strong signs of irregularities, such as rule violations in the gameplay, already at a low delay of 100 ms. While the aforementioned games demand differing degrees of interactivity between the players, the third tested game, *Curve Mania*, encompasses about the same intensity of player interactions as the highly delay-susceptible *Blobby Volleyball*. However, although this game also showed noticeable changes in the gameplay, it was much better received by the participants, as the game's appearance did not show unmistakable signs of malfunctioning, and rather led to multiple players' assumption of a cheating opponent. For the judgment of delay's impact on gaming QoE, it therefore seems that the exact nature of the impact of delay plays a role, i. e., whether the game rules apparent to the player are obviously affected or not.

As it has been shown that not only the game category or genre, but also the way the game is technically implemented influences player ratings significantly, the selection of comparable mobile games for use in the research of gaming quality influencing factors poses a serious challenge. Whereas a categorization of games based on aspects such as their game mechanics, input (e. g., touch-based, gamepad-based, movement-based such as using accelerometers or gyroscopes), or output (e. g., 2D, 3D, perspective) is basically possible with available or obtainable data, classifications based on internal implementation aspects and state synchronization algorithms is difficult since sufficient information about these are only available for a minuscule subset of games. But even if these details were readily available, they would likely be subject to frequent changes, as mobile games are usually updated many times. In a survey of the update frequency of the top 25 iOS apps, which include many games, Kimura et al. found an average update rate of 30 days[5]. Although only a fraction of these updates likely changes the core implementation of the games, these modifications nevertheless put into question previously obtained quality ratings.

It is therefore doubtful, if, in principle, an accurate and yet generic model of quality ratings for online mobile multi-player games can be built.

For the research of other influence factors, which is presented in the following chapters, the selection of games and the design of test beds is therefore performed in a way, which

[5] https://sensortower.com/blog/25-top-ios-apps-and-their-version-update-frequencies (last accessed: 2016-04-21)

evades settings in which the specific implementation of a games gains too much influence on the player's experience in the light of performed influence factor variations. Particularly in cloud gaming, where just a video stream of a game's output is sent to the user and input commands are transmitted vice versa, the effect of network impairments may be more generalizable as the implementation of the games is not directly affected (cf. Chapter 5).

As a more generic alternative, using a higher number of differing games in tests reduces the probability of observing very implementation-specific game behavior. This, however, comes at the price of increased test complexity and effort.

Chapter 4

Influence of the device

4.1 Introduction

With smartphone and tablet product announcements frequently promising increased gaming performance and improved playing experience, it is straightforward to assume an influence of the physical device and its properties on the subjective experience of games running on it. However, these advances in hardware capabilities can only transform into, e. g., more sophisticated imagery or more fluid animations if the game implementations are augmented and adapted concurrently. The perceivable output of a game is therefore the result of a complex interplay between the underlying system (i. e., the device, network) and the software (i. e., the game), and as such is largely dependent on the developer's implementation and optimization effort. Therefore, the result of changes to specific hardware parameters on subjective gaming experience is generally likely to be as strongly implementation-dependent, as has previously been shown for network delay in Chapter 3.

As only a small fraction of the population of actively used smartphone and tablet devices are equipped with the latest hardware generation, many game publishers try to increase the size of their target audience by supporting older devices. This requires limiting the games' hardware requirements to such a degree, that the audience's average phone can execute the game without complications. By resorting to conservative hardware requirements, differences between various device models and their processing capabilities are consequently alleviated to a high degree.

In this chapter, the focus is therefore placed on the device display and its size, as this part is one of the most important components of a mobile device, which is furthermore affected by the specific implementation of a game in a more generalizable manner: Bigger displays allow the same game to either simply display a larger version of its user interface, or render an adaption with, e. g., more detailed output, larger controls, or additional input methods.

Smaller displays, on the contrary, require more densely packed screens with less room for details and limited space for touch screen controls.

On the following pages, results from the Quality of Experience evaluation[1] of two commercially available games on four different smartphones and tablets with screen sizes between 3.27” and 10.1” will be presented and discussed. However, as the context (physical and/or social) is expected to be a confounding factor, it was simulated during the experiment to a certain degree as well by conducting the test in two different settings: a neutral lab and a simulated metro environment. The results show a considerable impact of display size on overall quality as well as four out of seven Player Experience dimensions. No significant impact of the simulated usage context on gaming QoE was observed, however.

This study has been published in [15].

4.2 Related work

In the past decade, the size of mobile devices changed quite dramatically with the public presentation of the iPhone and the onset of the smartphone revolution[2]. Before its beginning, displays on smartphones and Personal Digital Assistants (PDAs) were generally very small, as the devices often featured a hardware keyboard or an alpha-numeric keypad below the screen to take input. This was reflected in academia with skepticism regarding the suitability of the minuscule screens for media consumption such as the much-hyped mobile TV. In a paper entitled “Can Small Be Beautiful?”, published in 2005, Knoche *et al.* investigated various kinds of television (TV) content at different resolutions and scaled display sizes on an iPAQ PDA. They found that generally bigger is better and that participants favored the higher level of detail present in the bigger / higher-resolution displays [78].

In 2008, Maniar *et al.* evaluated the effect of display sizes ranging from 1.65" to 3.78" on video-based learning. It turned out that students using the smallest tested device had a significantly lower subjective opinion and learned considerably less than subjects using the bigger-sized devices [87].

When Kim *et al.* studied the psychological effects of different screen sizes on text reading or video watching in 2011, the range of sizes already went up considerably, going from 3.5" to 9.7". The analysis of the obtained data showed, that while the smallest device was praised for its higher perceived mobility, the biggest tested device also received the highest ratings for the level of enjoyment [76].

[1] The study was conducted in collaboration with Viktor Miruchna as part of a bachelor thesis.

[2] http://appleinsider.com/articles/14/05/06/before-apples-iphone-was-too-small-it-was-too-monstrously-big (last accessed: 2016-05-22)

Although no study regarding the influence of device or display size on *mobile* gaming was found, work was previously done for a stationary computer game by Hou *et al.* in 2012: Study participants played an action-adventure game on either a 12.7" or a much bigger 81" screen. The results showed that screen size significantly and favorably influenced the players feeling of involvement and participation in the game. It furthermore led to a "higher sense of being part of the game environment and more identification with the game avatar" [54].

Summarized, display size was found to be an influential factor in all cited studies. However, the relevance for mobile games, which are designed with small screens in mind is yet unexplored, and therefore investigated in the present study.

4.3 Methodology

For the study in August 2013 four popular screen sizes between 3.27" and 10.1" were chosen. To minimize effects of differing hedonic device quality, the selection of devices was limited to one brand: Samsung Galaxy Young (3.27"), Galaxy S4 (5"), Galaxy Tab 3 7.0 WiFi (7"), and Galaxy Tab 10.1 (10.1"). Although their build quality and case materials are comparable, different display technologies are used (AMOLED in the Galaxy S4, TFT in all others) and processing power differs. The Galaxy S4 ran Android version 4.2.2, whereas all other devices operated with Android 4.1.2. Yet, all were well capable to run the tested games without limitations.

As the usage context (physical and/or social) was expected to be a confounding influence on mobile gaming QoE, it had to be simulated as well. The first setting was the same laboratory room used in the study in Chapter 3, following ITU Recommendations P.910 [68] and P.911 [69], with participants sitting on an office chair next to a desk. The "metro" environment simulated a driving train with reduced lighting and train noises. The participant's space was limited using two gray partition screens very close to the sides of the player. In an effort to imitate the effects of a moving train, participants sat on a unsteady one-legged bar chair.

4.3.1 Selection of games

Two games were chosen based on their visual and input/control complexity: *Flipper Spiel Pinball*[3] and the more complex *Striker Soccer Euro 2012*[4].

[3] https://play.google.com/store/apps/details?id=com.PinballGame (last accessed: 2016-04-21)

[4] https://play.google.com/store/apps/details?id=com.uplayonline.strikersoccereuro_lite (last accessed: 2016-04-21)

Flipper Spiel Pinball

Flipper Spiel Pinball is a simple game representing a classic flipper. The player starts the game with a supply of four balls, depicted in the lower right corner of Figure 4.1a. These are shot onto the "table" (i. e., the gaming field) using a long press anywhere on the touchscreen. After the ball is launched, it will hit multiple of the round targets in the center of the screen, earning the player points in the process, and gradually roll down towards the bottom of the screen. There, the player has to use the two red levers to shoot the ball back up, trying to hit the targets as often as possible without losing the ball. The levers are operated by touches anywhere on the left side of the screen for the left lever, and on the right side for the right lever. One round of the game ends when all four balls have been played and lost. The game was chosen due to its simplicity: During playing, the screen remains mostly static except for the moving ball. To control the game, the entire touch screen can be used, making this game also easily playable on small devices without obstructing potentially important parts of the screen. Also, the low number of input options (launch ball, left lever, right lever) makes this game simple to learn and play.

(a) Galaxy Young (3.27").

(b) Galaxy Tab 3 7.0 WiFi (7").

Fig. 4.1 Screenshots from the game Flipper Spiel Pinball on a Galaxy Young and a Galaxy Tab 3 device.

When comparing the screenshots from the small Galaxy Young with its 3.27" (8.3 cm) measuring screen in Figure 4.1a to the bigger Galaxy Tab 3 tablet with a 7" (17.8 cm) display

in Figure 4.1b, it becomes apparent, that also the game's algorithm to fill the screen is rather simple: It merely scales the content to fit the display and even permits the resulting image to be skewed by the displays' differing aspect ratios.

Striker Soccer Euro 2012

Striker Soccer Euro 2012 is a soccer game where the player controls the actions of one team in a real-time soccer match, trying to score more goals than the other automatically controlled team. The currently controlled soccer player with a small red circle beneath him can be moved on the pitch using the indicated joystick imitation on the touchscreen in the lower left in Figure 4.1a. Concurrently, the player can choose to pass the ball to another player by shortly tapping on the other side of the screen, or to attempt a shot on the opponent's goal using a long press. The game is thus intended to be played using both hands simultaneously, with one finger resting on and operating the joystick and another finger handling the ball playing at the same time. Compared to *Flipper Spiel Pinball*, this game is significantly more complex as the controlled or selected character changes with each pass of the ball and the player is able to exercise different playing strategies. The game is also much more dynamic as multiple soccer players are moving concurrently (all automatically controlled except for the currently selected one), and the view of the pitch shows just the currently active segment of the soccer field (cf. Figure 4.2). In the game, a round ends when a preset time has passed. In the test, this was configured to three minutes.

(a) Galaxy Young (3.27").

(b) Galaxy Tab 3 7.0 WiFi (7").

Fig. 4.2 Screenshots from the game Striker Soccer Euro 2012 on a Galaxy Young and a Galaxy Tab 3 device.

A comparison of the game's interface on the different display sizes and aspect ratios of the Galaxy Young (cf. Figure 4.2a) and the Galaxy Tab 3 (cf. Figure 4.2b) reveals, that it adapts to the screen's dimensions. While the relative size of the soccer players remains

constant, the relative joystick dimensions vary to maintain an absolute size of approximately 2.5 cm in width and height, which is comfortably workable with a thumb.

4.4 Test procedure

The study was conducted using a within-subjects design with participants who were required to have prior experience in mobile gaming. After being instructed about the purpose of the experiment and filling in an introductory questionnaire, examining demographic information and prior experience with games and interaction with smartphones and tablets, the participants had to play a total of 12 game scenarios of approximately three minutes each in random order. After each test session, a three-part questionnaire had to be answered, containing the 42-item core part of the GEQ (cf. Section 2.6.3) one question for overall quality, and 4 further questions examining the suitability of the game for the present display. These questions had to be rated on an ACR scale labeled according to ITU-T Recommendation P.800 (cf. Section 2.6.6). Of the 12 tested conditions, 8 were situated in the neutral environment (both games on each device) and 4 conditions took place in the simulated metro (both games, only the biggest and smallest device). Due to the randomized test order, participants had to change between the two settings multiple times.

The study was conducted with 26 participants (17m, 9f; 22y-48y, avg. 25.5y) who were required in the invitation to be experienced in mobile gaming on smartphones or tablets.

4.5 Results

In the following sections, error bars indicate the 95% confidence interval. GEQ items were coded with the values 0 = “gar nicht” (i. e., “Not at all”) to 4 = “außerordentlich” (i. e., “Extremely”). The GEQ’s Player Experience dimensions were then calculated from these items according to [100]. The overall quality item was coded with 1 = “mangelhaft” (i. e., “Bad”) to 5 = “ausgezeichnet” (i. e., “Excellent”).

The collected GEQ data and the overall quality ratings from 312 sessions were tested for normality using a Shapiro-Wilk test with a significance threshold of 0.05. As no conditions clearly deviated from a normal distribution, the data was then analyzed using a multivariate analysis of variance (MANOVA) with the independent variables game, setting, and device and the dependent variables overall quality, sensory and imaginative immersion, competence, flow, tension, challenge, positive affect, and detail quality (suitability for display). The analysis showed that the overall quality MOS is significantly affected by display size ($F(3,300) = 38.87, p < .01, \eta^2 = .319$): Ratings using the smallest tested display size were significantly

lower (Scheffé post hoc test) than using the other displays. Among these bigger screens no significant differences were found (cf. Figure 4.3 and Figure 4.4).

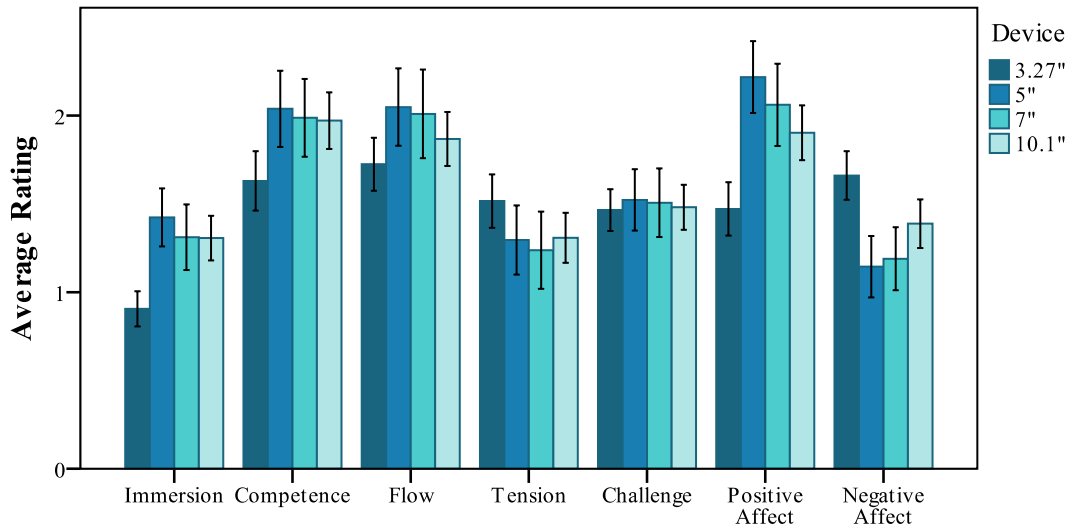


Fig. 4.3 Player Experience dimensions for the four tested display sizes averaged for both games and settings.

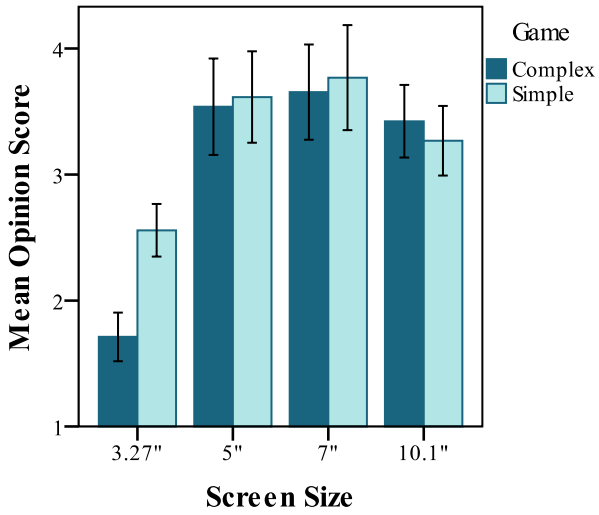


Fig. 4.4 MOS ratings for the two games on four tested display sizes averaged for both settings.

Significant influences of the display size factor were also observed for the quality dimensions shown in Table 4.1. While these effects exist for both games, they are more pronounced for the complex game (see Figure 4.4). Significant effects of the game factor are shown in Table 4.2. The environment factor showed no significant influence on any of the tested

Table 4.1 Significant MOS and Player Experience effects of the factor display size.

Dimension	Sig.	$F(3,300)$	η^2
MOS	$p < .01$	38.87	0.32
Immersion	$p < .01$	11.41	0.10
Competence	$p < .01$	4.58	0.04
Positive Affect	$p < .01$	10.33	0.09
Negative Affect	$p < .01$	6.48	0.06

Table 4.2 Significant MOS and Player Experience effects of the factor game.

Dimension	Sig.	$F(3,300)$	η^2
MOS	$p < .05$	4.78	0.02
Competence	$p < .01$	33.44	0.10
Tension	$p < .01$	43.40	0.13
Challenge	$p < .01$	80.00	0.21

dimensions. However, one participant remarked that he felt more comfortable in the metro situation, being hidden from the experimenter by the partition screens.

4.6 Discussion

The results confirm that the display size has a strong influence on the perceived quality of a gaming session. Although the screen sizes used in the experiments were not equally spaced on a continuum, there seems to be no linear link of quality with size. Instead, it seems that an acceptance threshold is reached as soon as the display has reached a certain size (in this study around 5”), and then quality and its sub-dimensions do not further increase significantly.

The data shows that smaller devices lead to lower playing experience ratings while gaming sessions with larger devices received higher marks. Considering the low ratings for Competence on the smallest device combined with the insignificance of the device’s influence on the Challenge dimension, it seems that the increased difficulty of playing on a small touch screen is not perceived as a challenge, but as annoyance, causing the observed higher Negative Affect scores (cf. Figure 4.3). As initially assumed, small devices are better suited for playing the simple than the complex game (cf. Figure 4.4). Although the games influenced ratings, the magnitude of their impact on the overall quality was lower than expected. It is possible that the participants focused primarily on the display sizes.

4.6.1 Limitations

In the study, the games' difficulty remained the same for all participants, potentially making them overly easy for some participants and too difficult for others. As the equilibrium of demanded skill and a player's abilities is a prerequisite for flow experience, games might need to be adapted in order to match the player's skills and represent an equal challenge in every case.

While the observed lack of influence of the simulated "metro" environment might mean that no context effect exists, the setting is not sufficiently realistic to completely dismiss its existence: The "metro" simulation may have been insufficient in that it did not take the social context into account to an adequate degree. Although the experimenter never interfered with the participants' playing, he was visible and his observation perceivable for the player in the neutral environment, whereas he was hidden in the "metro" setting by the partition screens.

4.7 Conclusion

In this chapter, a study was presented which examined the influence of a device on the gamer's playing experience. The parameter display size was chosen as an influential property of a handheld device and was therefore varied with four magnitudes in the test.

It was found that display size exerts a significant influence on the player's experience of a game and their ratings of the MOS and four out of seven Player Experience dimensions. The observed effect existed for both games used in the test, but was more pronounced for the more complex game, featuring a detail-rich in-game interface designed to be operated with both hands simultaneously. This rendered much of the screen invisible when playing on the smallest device in the test due to the fingers' obstruction of the display while manipulating the controls. The players' experience was therefore not only degraded by a smaller and less detailed display of the games' interfaces, but also by more difficult to handle controls. Which of these effects contributed more to the observed drop in player ratings in Figure 4.4 when moving from the 5" to the smaller 3.27" device may be an interesting subject for a future study.

As the display size of devices used in the test has emerged as an influencing factor, it has to be considered when planning future gaming studies. Strictly speaking only ratings obtained using a common display size are directly comparable. To study other influence factors, a constant and appropriate size should be preferred throughout the study. In field studies this is hardly possible. There, results might need to be grouped by similar display sizes.

Furthermore, the trend towards bigger displays in smartphones[5][6] might over time change player expectations. Whereas smallness was appreciated[7] before the debut of the smartphone revolution, the average size of sold phones has since risen[8] year after year.

The presented study did not indicate an influence of the playing context onto the participant's ratings. However, the performed simulation of different contexts was potentially insufficiently realistic. This was a motivation to conduct a combined laboratory and field study to confirm or refute the context insignificance and further test environmental effects on gaming experience. This study is presented in Chapter 6.

[5]https://medium.com/@somospostpc/a-comprehensive-look-at-smartphone-screen-size-statistics-and-trends-e61d77001ebe (last accessed: 2016-04-22)

[6]http://www.nielsen.com/us/en/insights/news/2015/super-size-me-large-screen-mobile-sees-growth-in-the-midst-of-a-small-screen-surge.html (last accessed: 2016-04-22)

[7]http://www.webdesignerdepot.com/2009/05/the-evolution-of-cell-phone-design-between-1983-2009/ (last accessed: 2016-04-22)

[8]http://www.pcworld.com/article/2455169/why-smartphone-screens-are-getting-bigger-specs-reveal-a-surprising-story.html (last accessed: 2016-04-22)

Chapter 5

Influence of the network

5.1 Introduction

Ubiquitous network connectivity is one of the main factors setting modern smartphone- and tablet-based gaming apart from older portable gaming consoles. Features such as online leader boards, turn-based and real time multiplayer gaming are therefore becoming more and more popular with mobile games. However, in Chapter 3, the interplay of three mobile games with varying network conditions was examined and found to be strongly implementation dependent, rendering a generalization of the subjectively observable effects of transmission channel parameter variations difficult for gaming setups where the major work of game computation is performed locally on the player's device. Yet, in use cases where the major computational load is concentrated on a remote server and all interactions with a game are equally required to traverse the network transmission channel, the perceivable effects may be more comparable. Consequently, the research discussed in this chapter focuses on this specific domain where local game-specific implementation details play a negligible role and network variations are likely to result in more similar and predictable effects: cloud gaming.

In this game delivery paradigm, the actual game execution is entirely decoupled from the display at the player's device, as the game's code runs on a remote *cloud* server and only a video of the game's output is streamed to client, which, in turn, sends back input commands. This division of work has fundamental consequences which apply to all cloud gaming systems: First, due to the transmission of commands and resulting output changes over a wide area network, additional delays are introduced to every interaction of the player with the gaming system. Second, the available bandwidth of the network limits the amount of information which can be transmitted between the server and the player. This necessitates the use of data compression, which, due to the amount of data reduction needed, typically results in the loss of information. Third, as the major burden of execution is performed on a remote

server, a loss of connectivity does not merely limit the usability of the service, but renders it entirely inoperable for the player. Gaming contexts without sufficient Internet access can therefore not access gaming services built using the cloud gaming paradigm in principle.

While cloud gaming with PC or console gaming titles has been subject to a multitude of studies, the application of the streaming concept to mobile touch-based games has so far not been thoroughly investigated. To examine the effects of additional input delay and reduced output quality due the data compression in this particular use case, a test bed called Stream-a-Game was developed and used in a laboratory study. This test bed and the study are presented in this chapter.

5.2 Related work

When the company G-cluster first publicly demonstrated a system[1] in 2000, which could stream the visual output and audio of Personal Computer (PC) games to a PDA in real time and could process commands received from that device at the E3 trade fair, it received interest from the commercial and academic world alike. While the business world was predominantly attracted by aspects such as the effective protection against piracy (the actual game code never leaves the server), novel business models (e. g., subscription-based gaming instead of single purchases), or reduced development effort (i. e., no adaption of the game to multiple platforms), the academic community embraced the concept's many technological challenges (e. g., load distribution and virtual machine placement [50], efficient video compression [2], hardware virtualization [45, 102], or network optimization [53]), but also the streamings' effects on the subjective experience of the gaming.

Compared to other services investigated by the QoE community, cloud gaming has a prominent position in that it is considered to be the most complex non-business-oriented service which at the same time has the highest degree of interactivity and is the most multimedia-intensive of all considered service categories [52]. As this complexity extends to the test bed needed to experimentally investigate cloud gaming, research on the topic was long hindered by the unavailability of freely available implementations of such a streaming system. Therefore, interested groups had to develop their own setup:

In 2009, Wang *et al.* [119] presented the first study examining the subjective perception of what they called cloud mobile gaming. They streamed three conventional PC games using a custom-built solution to an unspecified mobile client and varied resolution, frame rates, PSNR, delay, and packet loss. Although their publication leaves many aspects of their setup, study methodology, and their obtained results undetailed, they found that for the

[1] http://www.gcluster.com/eng/ (last accessed: 2016-05-13)

MMORPG World of Warcraft, subjective ratings began to lower at added network delays above approx. 120 ms. From the obtained subjective MOS ratings, they derived a prediction model designed in the style of the E-model [65] used in speech communication quality prediction. Their model is, however, limited to the specific games used in the test and is furthermore debatable, as major aspects (i. e., process of finding the specific factors in the equations, used hardware, available controls on the client, overall system delay, game scenarios, study group composition, test design, tested condition lengths, encoder settings, presence of audio, observable effects of packet loss, etc.) of its derivation remain unclear.

In 2011, Jarschel *et al.* [71] addressed QoE effects of simulated network delay and packet loss in a cloud gaming scenario built using a special-purpose streaming appliance called "Spawn Box". The simulated parameters for delay ranged from 0 to 300 ms, whereas packet loss levels spanned from 0 % to 1.5 %. They grouped the three games used in the test into the categories "slow", "medium", and "fast" depending on the pace of their action and found that the perceived quality (MOS) under simulated loss and delay depended on that category: The "fast" game's ratings appeared to be more tolerant to loss but reacted sensitive on delay when compared to the "slow" game, which, in turn, was less sensitive towards delay but reacted more delicately to lost packets. Unfortunately, the overall end-to-end delay of the used setup was not reported, making it difficult to compare the tested delay levels to other studies.

The intrinsic system delay of a cloud gaming setup (i. e., not considering network delays), however, was shown to vary significantly between different cloud gaming systems by Chen *et al.* [29]. In their measurement study, the processing times (in this case: time between sent command on network level to response data received) varied between 110 ms and 471 ms. Although these numbers do not represent the whole user-perceivable delay, which is higher due to additional local processing and input and output delays, they illustrate nevertheless, that the server-side cloud gaming implementation contributes significantly to the overall delay. Yet, due to the complexity of cloud gaming, all previous works relied on incomparable custom-built solutions or on existing commercial black box systems such as StreamMyGame or hosted services like OnLive where details of their implementation could not easily be varied in studies.

The presentation of the open-source cloud gaming system GamingAnywhere (GA) [55] in 2013 by Huang *et al.* changed that situation and first allowed the execution of fully repeatable experiments as researchers were in full control of the entire cloud gaming system. Since then, GamingAnywhere has been continuously developed as an open-source project and gained a rich set of features such as support for the emerging H.265/HEVC [115] video compression standard. GamingAnywhere was subsequently used in several QoE studies: Slivar *et al.* [114] compared native game-play of the World of Warcraft MMORPG with a version which

was streamed at 3 Mbit/s using GamingAnywhere in the periodic screen capture mode in an experimental in-home streaming setup where the game's output was streamed from one computer to another in the Local Area Network (LAN) and input commands were sent back vice versa. Study participants consistently had lower willingness to continue playing when they experienced the streamed version of the game. Using the participants' ratings, a model was created to predict the MOS of the tested in-home streaming setup under the influence of delay and packet loss on the external Internet uplink.

Claypool *et al.* [35] also used GamingAnywhere when they tested the effects of added network delay on a streamed PC skill game involving rolling marbles around obstacles in hillocky 3D world by tilting that world. They found subjective ratings to significantly drop at delay levels above 100 to 150 ms.

Despite the existence of the open-source cloud gaming toolkit, QoE studies continue to be conducted with commercial streaming setups such as Steam In-Home Streaming: Slivar *et al.* [113] investigated the interaction of frame rate and transmission bit rate with a fast-paced FPS and a slower role-playing game. They found that reducing the frame rate never resulted in raised ratings for gaming experience when the bit rate was kept constant. A reduction to 15fps, however, resulted in significantly lowered ratings. Compared to these substantial quality decreases, a reduction of the bit rate from 10 to 3 Mbit/s led to only minor quality degradations.

In another more recent publication from Slivar *et al.*, models were created [112] using laboratory ratings from 52 study participants to predict the MOS of an FPS and an online collectible card game based on the frame rate and bit rate at which the games were streamed using the commercial Steam In-Home PC cloud gaming system.

While the models created by Slivar *et al.* and Wang *et al.* may not be generic in that they can predict quality ratings for games other than they were created and trained for, they do, however, suggest that games in a cloud gaming setup respond to changes of the system settings and the network channel in a generalizable way. As the parameters of the proposed models differ between games, they are not directly applicable to mobile games which have different interaction models (i. e., usually direct manipulation using touch-based input). Here a new research field opens up, to which the study presented in this chapter contributes.

5.2.1 Suitability of games for cloud gaming

Considering the necessary compression of the transmitted video, multiple measures have been proposed to describe a game's output in terms of its visual complexity. Claypool [34] describe the motion complexity of a game's visual output using the percentage of forward/backward or intra-coded macroblocks (PFIM) of an MPEG-compressed video recording of the game.

To describe the scene complexity, he uses the average intra-coded block size (IBS) present in the file. These metrics were shown to correlate moderately well with users' ratings of a games' motion and scene complexity.

Chen *et al.* [32] describe a game's suitability for cloud gaming using three parameters: *screen dynamics* computed from the encoded video's motion vectors, *command heaviness* (quotient of screen dynamics and the rate of input commands), and a derived *real-time strictness*. Suznjevic *et al.* [116] compared PFIM and IBS metrics proposed by Claypool with measures of spatial (SI) and temporal complexity (TI) standardized by the ITU-T in Recommendation P.910 [68] for a broad variety of PC games. Both PFIM/TI and IBS/SI were shown to exhibit a high degree of accordance.

5.2.2 Mobile cloud gaming

Previous works on streaming games to mobile devices have concentrated on delivering desktop class games to less capable battery-powered handheld devices through the means of cloud gaming (e.g., [29], [119]). This device category change entails the need to adapt the input mechanisms expected by the games (e.g., keyboard, mouse, controller) to means available on the mobile device. While dedicated mobile gaming devices with support for cloud gaming such as SONY Vita or Nvidia Shield offer input options comparable to a console game controller, other solutions encompassing general purpose mobile devices such as smartphones or tablets typically employ custom gestures or overlay buttons, which are displayed on-top of the streamed game output. Although these substitute input mechanisms permit bridging the gaps between different device categories, they require the gamer to adapt and may not reach the versatility of the original control they replace. These latter methods of cloud gaming are therefore not considered to be truly comparable to ordinary mobile games, which are designed with the (usually touch-based) input options and limitations (e. g., small screen) of the mobile device in mind. In this chapter, therefore an alternative approach is taken, which uses the cloud gaming concept with preexisting unmodified mobile games.

5.3 Methodology

As a prerequisite for the research of subjective effects of streaming mobile games using a cloud gaming paradigm, a test bed is required. Since existing solutions including the open-source GamingAnywhere currently cannot stream this category of games, a new system had to be developed. To investigate the subjectively perceivable effects of network degradations

on the gaming experience of games streamed using that test bed, a study with test participants was conducted.

5.3.1 Stream-a-Game test bed

This test bed for streaming mobile games was called *Stream-a-Game*, published as an open source project[2] and demonstrated publicly at the NetGames conference in 2015 [18]. In contrast to previous works mentioned in Section 5.2, this mobile cloud gaming platform does not bridge device category boundaries but streams smartphone and tablet games to those very devices. This allows conducting research of the implications of network degradations and delay onto the Quality of Experience (QoE) of mobile cloud gaming with realistic use cases (i.e., with games which are designed and optimized to be played on mobile devices).
The Stream-a-Game test bed consists of four distinct building blocks:

- The compute component runs an Android system inside a virtualization environment,
- the rendering component receives OpenGL instructions and textures from the virtual Android and renders them to pixel-based images using a hardware Graphics Processing Unit (GPU),
- the streaming component compresses these rendered images and provides a video stream on the network and
- the client accesses and displays this video stream and transmits input commands back to the server.

These four components compose a pipe in which visual output flows from the virtualized Android to the client and input commands are forwarded vice versa. This modular design allows each component to be independently developed and configured (e.g., the version of the Android system inside the compute component may be altered without implications to the rest of the system).

5.3.2 Selection and variation of parameters

Wide area networks in general or the particular transmission channel between server and client can be characterized by numerous parameters such as bandwidth, end-to-end or round-trip delay, delay jitter, packet loss rate, packet loss distribution, packet corruption rate, and more. As set out above in Section 5.2, frequently used criteria in gaming quality

[2] https://github.com/streamagame/streamagame

research are bandwidth, round-trip delay, and packet loss. Whereas the former two are adopted, the latter parameter is skipped in this study despite its importance in inevitably lossy wireless connections: The latest generation of commercial cloud gaming systems employ Forward Error Correction (FEC)[3] dynamically to protect a stream's contents against the loss or corruption of information. In other domains, this technique has successfully been used to correct transmission errors in, e. g., digital TV broadcasting in DVB-T2 or DVB-S2 [82, 93, 117]. Although this error protection comes at the cost of increased data volume caused by added redundancy, translating into lower usable net bandwidth, this development is considered to substantially change the subjectively perceivable effects of packet loss. While packet loss's influence on subjective gaming experience in a cloud gaming setup is acknowledged, it is nevertheless skipped in the present study, as Stream-A-Game currently provides no protection against loss or corruption and such behavior is considered to be unrealistic for a serious commercial service provider in the face of recent developments. Results using Stream-A-Game with packet loss would therefore likely not be generalizable.

For the remaining two parameters, bandwidth and delay, suitable levels were identified in a pre-study. These levels were 384, 768, 1536, and 3072 kbit/s for the bandwidth factor, and 0, 100, 200, and 300 ms for network-level round-trip delay. The selection of these values was also guided by previous research such as by Claypool *et al.*, who found player ratings to significantly drop at delays bigger than 100-150 ms [35], and Slivar *et al.*, who reported high subjective ratings at a 3 Mbit/s level. The selection of bandwidth and delay levels was therefore considered to cover the critical range, where subjective quality perception would likely become affected by these degradations. While the different transmission bit rates were achieved by reconfiguring the video compressor during run-time using a purpose-built extension[4] of GamingAnywhere in Stream-A-Game, network delay was selectively added using the Linux 'netem' network emulator kernel module [48] on inbound User Datagram Protocol (UDP) control packets on the rendering system. The delay created with 'netem' added to the existing intrinsic system delay.

The subjectively perceivable effects of particularly the lower bit rate levels with the Stream-A-Game test bed in high motion scenes are high degrees of blockiness, discolorations, and streaks behind moving objects. In low motion scenes, however, only very small amounts of blockiness are visible. Added network-level delay, on the other hand, leads to the impression of a sluggish and delayed system response.

[3] http://netgames2015.fer.hr/presentations/FranckDiard.pdf (last accessed: 2016-06-16)
[4] https://github.com/streamagame/gaminganywhere/commits/feature/live-reconfigure

5.3.3 Selection of games

As shown, e. g., by Claypool, games vary with regard to their suitability for cloud gaming due to different visual complexities and dissimilar delay requirements [34]. The goal of the game selection process described in this section was to identify games which differed possibly strongly with regard to these dimensions and were still quick to learn and play to be usable in a study.

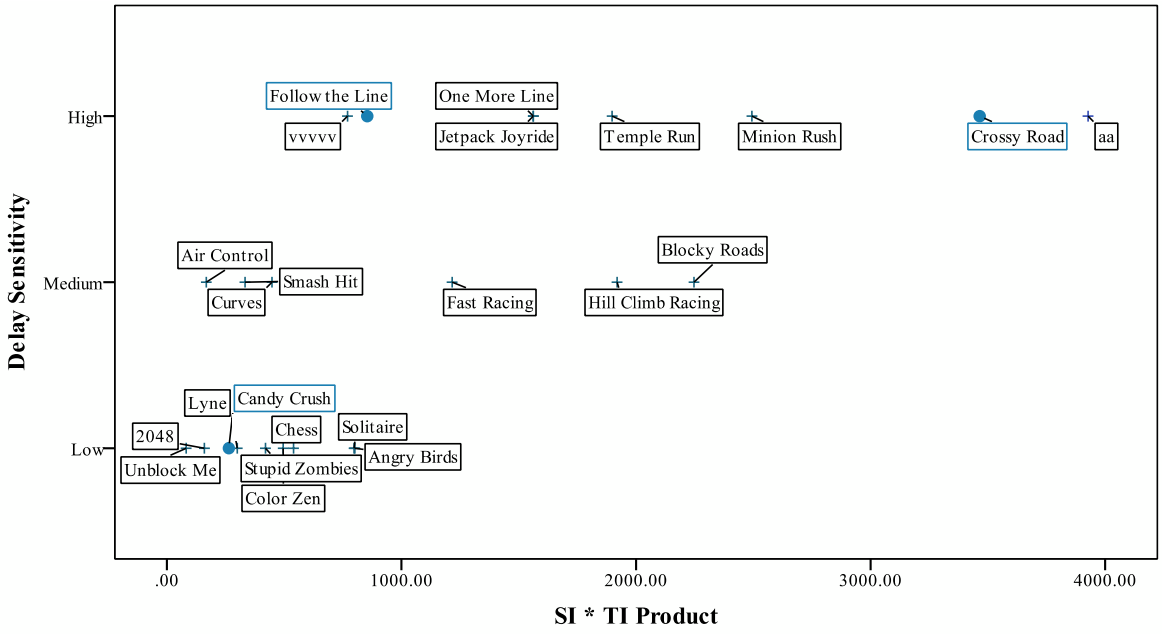


Fig. 5.1 Scatter plot of the SI · TI product and the delay sensitivity of 23 popular Android games. Titles selected for the study are shown with a circle.

Following the procedure from Suznjevic *et al.* [116], mean spatial (SI) and temporal complexity (TI) values were calculated for 23 popular mobile games from Google's Play Store using a set of multiple 5 second long video recordings per game with a resolution of 1280x720 covering typical game scenes. To derive an estimate of the amount of visual information generated per time unit, the product of SI and TI was computed. As the video compression in cloud gaming removes visual information to shrink the data volume needing to be transmitted, games which deliver a more complex output image should suffer from a stronger visual degradation than titles with an intrinsically simpler output. Additionally, each of the games was classified regarding its delay sensitivity as part of an expert review, in which three experienced mobile gamers evaluated the respective games and agreed on a sensitivity judgment based on the time between a game's visual clue and the required reaction from the player to succeed in the game. The results of this survey are compiled in Figure 5.1.

As can be seen, the SI·TI product varies strongly between games. However, it also seems that the average visual complexity of games rises with higher delay sensitivity. It furthermore looks as if titles with low delay sensitivity are generally restricted to lower SI·TI values. This may be caused by many of these games' mainly static screen during periods where player input is awaited. As the sample of 23 games is small and may not be representative for all games in the Play Store, these findings rather pose working hypotheses than substantiated proofs.

Because the total number of games used in the test was limited to three to keep the test duration manageable, these titles were chosen to be possibly far apart in the plot in Figure 5.1.

Candy Crush / Candy Frenzy 2

Candy Crush Saga[5] is a very popular casual game with millions of downloads, depicting a matrix of differently shaped and colored little sweets (cf. Figure 6.1). The player's task is to create a possibly long array of similar items with a single swap of sweets from adjacent cells. This line of candies then vanishes, whereby points are awarded, and new items flow in from the top.

Fig. 5.2 Screenshot from the game Candy Frenzy 2.

[5] https://play.google.com/store/apps/details?id=com.king.candycrushsaga (last accessed: 2016-04-23)

Due to a technical problem with embedded code in *Candy Crush* specifically compiled for the ARM processor architecture, the original game could not be used on the Intel x86-64-based Stream-a-Game test bed without an additional compatibility layer[6]. However, multiple clones of the game exist, which precisely copy its style and game interaction model. One of these is *Candy Frenzy 2*[7], shown in Figure 5.2 which is used in this study.

Like *Candy Crush*, *Candy Frenzy 2* does not pressure the player to act quickly as no quick reactions to visual changes in the game's output are required. Furthermore, no time limits are imposed and a slower, more careful interaction with the game is possible and has no negative consequences. Due to these considerations, *Candy Crush* was considered to be very insensitive to delay in Figure 5.1 and due to its similar game paradigm, the same is assumed for *Candy Frenzy 2*. Additionally, both games' output remains completely static while the games await player input, resulting in a low SI·TI product for *Candy Crush* of 264 (SI=88, TI=3) and making both titles hypothetically well suitable for streaming at low bit rates and considerable delay.

Follow The Line 2

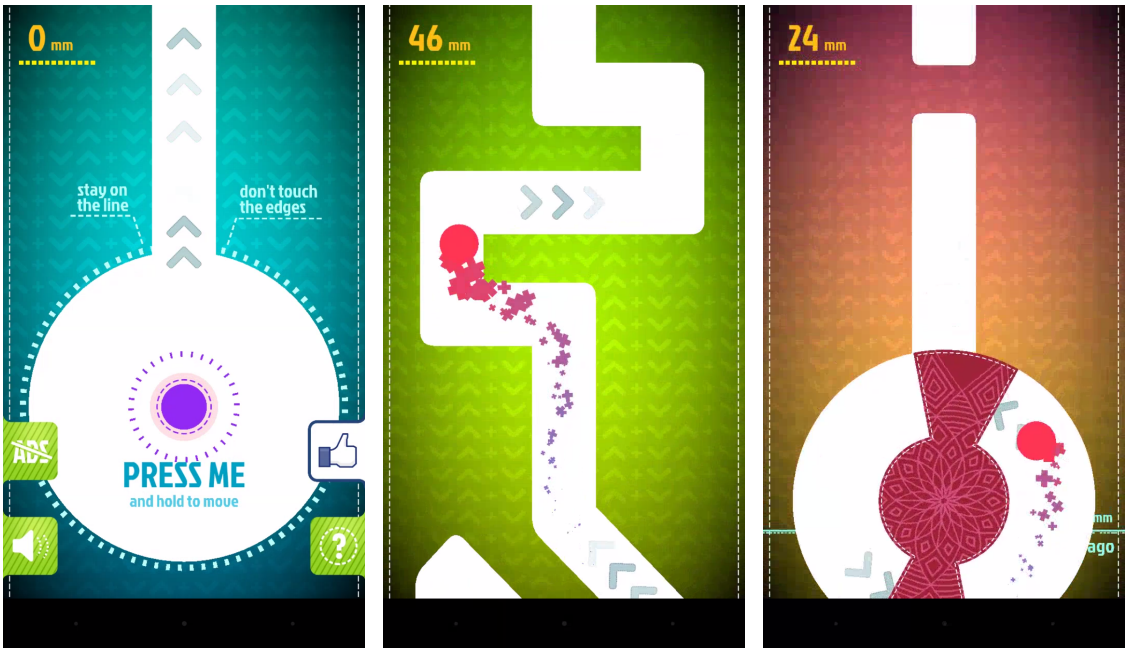
Follow The Line 2[8] is a skill game, in which the player has to draw a path with his finger tip along a white line or through white spaces without touching the boundaries or upcoming obstacles, as the course gradually moves by. As shown in Figure 5.3, the position on the touchscreen where the finger touch is registered is highlighted with a red circle which leaves a trail as it moves along the path.

As some of the obstacles move continuously or periodically, precise timing is necessary to prevent the red circle from leaving the white path and touching the boundaries. As a consequence, the game was considered highly delay sensitive (cf. Figure 5.1). Although the field of view is continuously changing as long as the player's finger touches the screen, it moves in a uniform motion creating a high degree of similarity between one frame in the video stream and the next. For the game, an SI·TI product of 854 (SI=61, TI=14) was calculated which is one of the lowest of the surveyed highly delay sensitive games.

[6]https://commonsware.com/blog/2013/11/21/libhoudini-what-it-means-for-developers.html (last accessed: 2016-06-17)

[7]https://play.google.com/store/apps/details?id=com.appgame7.candyfrenzy2 (last accessed: 2016-06-17)

[8]https://play.google.com/store/apps/details?id=com.crimsonpine.followtheline2 (last accessed: 2016-06-17)



(a) Start screen with minimal instructions. The game starts when the circle is touched.

(b) The course has to be followed without touching the white line's borders.

(c) Some obstacle move or rotate, requiring the player to react with precise timing.

Fig. 5.3 Screenshots from the game Follow The Line 2 with the red dot signaling the position where the player's finger tip is sensed.

Crossy Road

In *Crossy Road*[9] the player controls a chicken using tap and swipe gestures and has to make it hop across busy roads and train tracks, and cross rivers by jumping from one floating trunk onto the next (cf. Figure 5.4). The chicken dies when gets into contact with a vehicle or a train, drowns when it jumps into a river, and has to return to the starting point when the player acts too slow. In any such event, a single press on a retry button suffices to immediately begin another attempt. The game's goal is not to reach a destination, but to move the chicken as many steps as possible before it inevitably dies.

The visual style of *Crossy Road* is deliberately blocky and pixelated (cf. Figure 5.4). However, despite the many isochromatic areas inherent in that visual style, edges never proceed in parallel to the screen borders and its pixel matrix: The game's visual perspective is slightly rotated, i. e., the chicken does not move straight in an upward direction on the screen, but also slightly to the right. This aesthetic with its many high-contrast edges and a high degree of movement on the screen due to passing cars, trains, and tree trunks lead to a very high SI·TI product of 3465 (SI=105, TI=33). Additionally, the game requires quick

[9] https://play.google.com/store/apps/details?id=com.yodo1.crossyroad (last accessed: 2016-06-17)

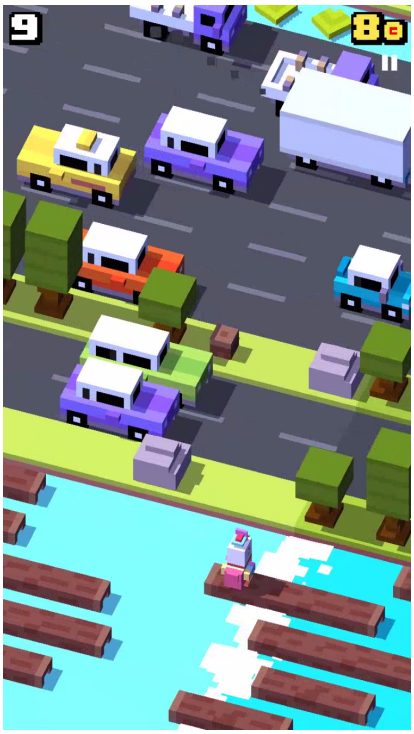


Fig. 5.4 Screenshot from the game Crossy Road.

reactions and precise timing to maneuver the chicken alive through all the perils, making the game fall into the category of highly delay-sensitive games in Figure 5.1.

5.3.4 Study set up

To study the network influence on mobile cloud gaming experience in this chapter, the Stream-a-Game compute component with Android 5.1.1 was set up as a virtual machine (VM) equipped with 4 virtual CPU cores and 2 GB RAM on a DELL Precision WorkStation T7500 (2x 4-core Intel XEON X5550 2.67 GHz, 48 GB RAM) with the open source virtualization environment XenServer 6.5.0-90233c. This VM was connected to a switched Gigabit Ethernet network with a standard 1500 bytes Maximum Transmission Unit (MTU) size. On the virtual Android device, the three selected games from Section 5.3.3 were installed.

Connected to that same network was another purpose-built computer (4-core Intel Core i5-4460 3.2 GHz, 32 GB RAM, AMD Radeon R9 290X, Ubuntu Desktop 15.10 with the fglrx GPU driver 15.201.1151) running the rendering and streaming components. The GamingAnywhere-based streaming component was configured to generate a video stream at a resolution of 704x1248 pixels (upright image) according to H.264's Main profile and to use the x264 *ultrafast* encoding preset with *zerolatency* tuning. It was allowed to use only a

single (i. e., the previous) frame as reference in video encoding (for performance reasons) and send keyframe information at least every 250 frames. The system's frame rate was variable in the range from 40[10][11] to 50Hz[12] depending on changes of the screen: Without the need for updates to the screen's content, the compute component does not send any commands to the rendering component, causing no new frames to be generated. To allow key frames to be generated nonetheless at regular intervals in the event of absent content updates, the streaming component generates duplicated frames to maintain a minimum frame rate of 40 Hz (80 % of the nominal frame rate, cf. [18]). Consequently, key frame information is sent at least every 6.25 seconds. During the development of the platform, it was noticed, that the handling of key frames is critical for both the performance and the visual fidelity of the system: Conventional key frames contain sufficient information to reconstruct a full image of the video stream without referencing previous frames. This inevitably causes spikes in the transmission bit rate of the stream as these full frames require more information to be transmitted than (P-)frames that are allowed to reference image data from previous frames. While this behavior is not problematic for delay-insensitive streams as long as the average bit rate resulting from buffering does not exceed the limits of the transmission channel, games require both low delay (i.e., no buffering) and a constant frame transmission latency (i.e., frame sizes have to be relatively homogeneous). In the platform this is achieved through a video coding feature called "Periodic Intra Refresh" (PIR), which omits full key frames in the video stream and instead gradually delivers reference-free blocks of image data to the client, spreading the overhead to recover one full image over many frames instead of one [82, 111]. The upper end of the frame refresh range (50 Hz) was deliberately set below the typical 60 Hz used by contemporary smartphones and tablets to avoid potentially accruing a backlog of waiting frames in the play-out buffer of the client device due to a display clock which can be slightly slower than the server's frame generation rate.

The bit rate control algorithm in x264 was configured to use constant bit rate (CBR) mode encoding with a low rate control buffer size of 768 kbit, which enforced a highly homogeneous output stream bit rate even during intervals of strong visual changes in the compressed video. The x264 CODEC was furthermore set to subdivide each frame into four slices which it compressed using a similar number of threads concurrently, thereby distributing the load of the video compression as evenly as possible on the computer's four physical Central Processing Unit (CPU) cores and consequently further reducing frame

[10] https://github.com/streamagame/streamagame/blob/master/conf/streamer.conf#L14

[11] https://github.com/streamagame/gaminganywhere/blob/devel/ga/server/event-posix/ga-hook-gl.cpp#L119

[12] https://github.com/streamagame/streamagame_platform_sdk/blob/streamagame-lollipop-x86/emulator/opengl/host/libs/libOpenglRender/RenderControl.cpp#L149

encoding time. Finally, the size of each of these frame slices was limited to 1450 bytes to fit well into a single UDP packet.

The client component was installed on an iPhone 6 running iOS 9.2.1 and using FFMPEG / x264 software-based video decompression while color space conversions from the stream's YUV to the screen's RGB were performed on the device's GPU (cf. [18]). The device connected to the compute, rendering, and streaming units' network using a dedicated Apple Airport Express 2012 802.11n access point[13] operating on an otherwise unused 40 MHz-wide channel in the 5 GHz spectrum. To prevent involuntary interactions with the device's native iOS operating system (e. g., opening the notification or control center using unintentional swipe gestures), the "guided access" mode[14] was enabled, ignoring any user input not directed at the foreground app - the streaming client.

5.3.5 Measurement of end-to-end delay and test bed verification

Since the overall delay between a player's touch input and the visual response appearing on the screen is not merely the result of network delay, but also influenced by numerous other latency contributor such as video encoding, decoding, game processing, screen refresh, etc., experimental results can only be compared by the overall system delay. In [20], a method was presented to measure that system parameter using a low-cost Arduino device. For the present setup with an iPhone 6, the time from touch input to visual response in the virtual Android environment streamed using Stream-A-Game without any added network delay was observed to be 144 ms for all used video compression bit rates. Further measurements were performed to ascertain, that the chosen levels of network delay added to the intrinsic system delay in a linear manner. The effective player-perceivable overall delay levels occurring in the study are therefore 144 ms (no additional network delay), 244 ms, 344 ms, and 444 ms. In the following, only these values are used in this chapter.

5.3.6 Subjective assessment method

As means to measure the degradation of the subjective gaming experience by the impaired visual quality and the delayed system response, the ACR self-assessment method (cf. Section 2.6.6) with a continuous rating scale as in Figure 2.4 was used to let participants rate their individual experience of overall and video quality. To assess potential emotional effects of the varied system behavior, the SAM questionnaire (cf. Section 2.6.4) was used as its

[13] http://www.apple.com/airport-express/specs/ (last accessed: 2016-04-11)
[14] https://support.apple.com/en-us/HT202612 (last accessed: 2016-06-16)

Table 5.1 Selected delay and bandwidth conditions to test in the study.

Bit rate levels	System delay levels			
	144 ms	244 ms	344 ms	444 ms
3072 kbit/s	*	*	*	*
1536 kbit/s	*			
768 kbit/s	*			
384 kbit/s	*			*

three items may be filled in a very brief period of time, thereby allowing a greater number of conditions to be tested in a limited time than would have been possible with, e. g., the GEQ.

5.4 Test procedure

As part of the preparation of each experiment session, the test device was charged, the Stream-a-Game setup run, and the games' data reset to discard any previous high scores and to mitigate potential game adaptations (e. g., higher challenges due to a highly skilled previous participant).

Study participants were invited using a web portal and required to play mobile games for at least four hours per month, and to have basic knowledge of the English language to not be confused by non-German messages and texts. Upon the arrival of a participant, he/she was accompanied to a sound-proof and air-conditioned laboratory room following ITU-T Recommendations P.910 [68] and P.911 [69]. There, a written introduction was read, an informed consent signed, and a demographic questionnaire filled. After that, the actual game testing began.

A full factorial test with a within subject design would have required 3 games · 4 latencies · 4 bit rates = 48 test conditions of multiple minutes each, resulting in an infeasible total of more than two hours of uninterrupted playing and rating. Therefore, a partial factorial design was created, which reduced the number of delay and bit rate combinations for each of the three games as shown in Table 5.1. This test plan retained all delay conditions at the visually least degrading video bit rate and vice versa all bit rates with the lowest possible system delay. Additionally, a combination of the worst levels of bit rate and delay was preserved from the full factorial design to allow creating an estimate of the subjective severity of the combination of these two types of impairments.

While a fully randomized condition order may have been desirable to be able to put all test conditions in relation to each other, it would have been time-consuming as the games require around 30 seconds to start. It was furthermore considered to be highly unrealistic,

frustrating, and exhausting to keep switching games in a rapid manner for an extended period of time. Instead, letting participants play conditions for each game en bloc was deemed more appropriate, as it allowed them to grow accustomed to each of the games, improve their skills, and successively exceed their own previous high scores or achievements. To nevertheless minimize order effects, the order of the game blocks and the sequence of test conditions within them was randomized.

Following ITU-T Recommendation P.911 [69], each gaming block was begun with a training session, in which study participants were introduced to the game under test and allowed to play the best (3072 kbit/s, 144 ms overall system delay / no added network delay) and the worst (384 kbit/s, 444 ms overall system delay / 300 ms network delay) conditions. After that introduction, the game's actual test session with 8 conditions was begun. After one minute of playing a condition, a bell was rung to signal the start of filling the questionnaire. Participants were, however, allowed to continue to play the current round if they wanted.

After the last gaming block was finished, participants were thanked for their participation, informed about the purposes of the study, and given €15 as compensation for their effort in participating.

The study was conducted from 2016-06-03 to 2016-06-08 in a laboratory room at Technische Universtät Berlin. Altogether 20 subjects (9 females and 11 males; mean age = 28.25 years; SD = 5.408; range = 19-41) participated in the study, of whom the majority were either students (60 %), or employees (25 %). 12 had previously played the game *Candy Crush*, 4 knew *Crossy Road* from personal experience, and only one had played *Follow The Line* before. From the 20 subjects, just one had previously participated in a gaming study. Together, the participants played and rated 480 sessions.

5.5 Results

The error bar in all following figures indicates a confidence interval of 95 %. The continuous rating scales used for the overall, and video quality MOS were mapped to the range from 0 = "*extremely bad*" to 6 = "*ideal*". Ratings on the SAM pictorial scales were coded to the range from 1 to 9.

To analyze the obtained data, the distribution of the ratings for each condition was tested for normality using a Shapiro-Wilk test with a significance threshold of 0.05, which was preferred over a Kolmogorov–Smirnov test due to the small sample size. This test revealed significant violations of the normality assumption for multiple items in numerous conditions. Consequently, in the following, non-parametric tests are used.

Table 5.2 Spearman's correlation coefficients r_s of questionnaire items' data points for each condition. For each condition the obtained data points are all significantly correlated at the $p < .01$ level.

	Overall MOS	Video MOS	Pleasure	Arousal	Dominance
Overall MOS	1	.869	.376	-.251	.392
Video MOS	.869	1	.279	-.135	.289
Pleasure	.376	.279	1	-.473	.689
Arousal	-.251	-.135	-.473	1	-.448
Dominance	.392	.289	.689	-.448	1

For each condition, all questionnaire items' data points were inter-correlated at the $p < .01$ level. The Spearman's correlation coefficients r_s are shown in Table 5.2. According to these coefficients, the ratings of video and overall MOS show a very high degree of similarity ($r_s = .869, p < .001$).

5.5.1 Influence of video bit rate variation

In this section, the subset of obtained data points with a common system delay of 144 ms but varying bit rates is analyzed. In Figure 5.5, the mean ratings for all obtained dimensions are shown with ratings for the three games averaged. For all five displayed dimensions, a clear influence of changed bit rate is visible as higher bit rates improved overall and video quality ratings and led participants to feel less aroused but more pleased and in control. According to non-parametric Friedman tests of differences among repeated measures, these visible differences are significant for overall quality ($\chi^2 = 46.74$, $p < .001$), video quality ($\chi^2 = 50.40$, $p < .001$), Pleasure ($\chi^2 = 26.66$, $p < .001$), Arousal ($\chi^2 = 8.69$, $p < .05$), and Dominance ($\chi^2 = 25.94$, $p < .001$).

As the games were selected by the visual complexity of their output, assuming they might differ in their suitability for cloud gaming and therefore being differently influenced by bit rate reductions, their mean overall quality ratings (MOS), video quality ratings (MOS_V), and the three SAM dimensions at the four tested bit rates were analyzed and plotted for each game in Figure 5.6. The significance of the caused differences was again tested with repeated-measures non-parametric Friedman tests and the results are reported in Table 5.3.

For all three games, bit rate variations significantly effect ratings for overall quality, video quality, and the SAM Pleasure dimension. For the remaining two SAM dimensions Arousal and Dominance, the influence is more mixed: In the game *Follow The Line 2* they are both significantly affected, whereas in *Crossy Road* only Dominance is (strongly) effected, and no effect is seen for neither of the two in *Candy Frenzy 2*.

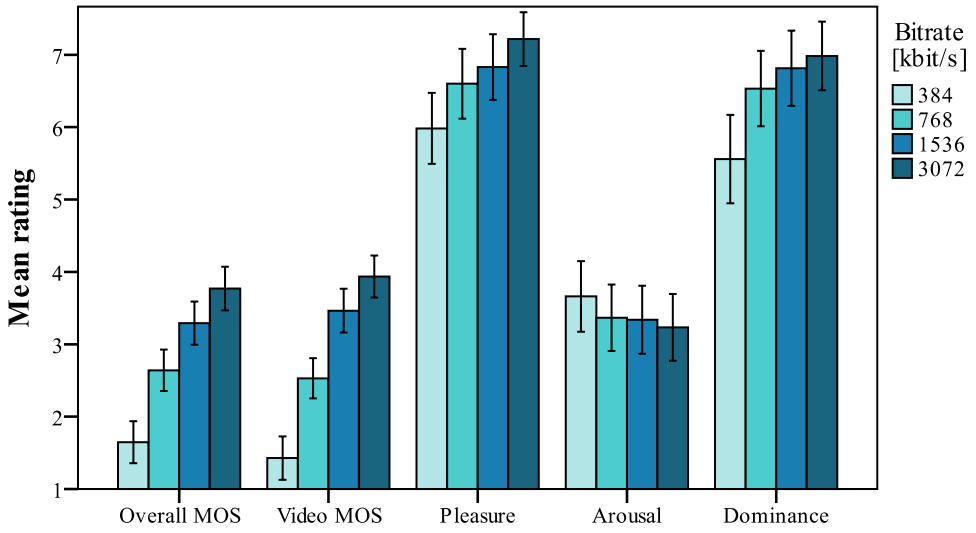
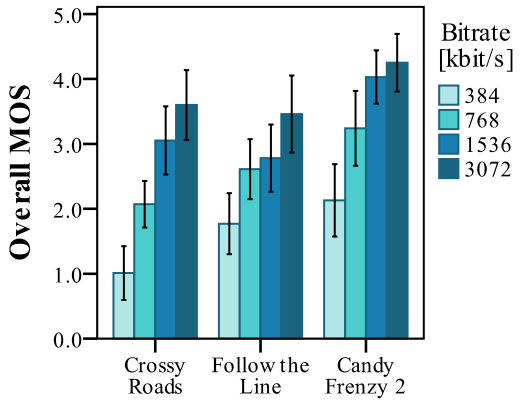


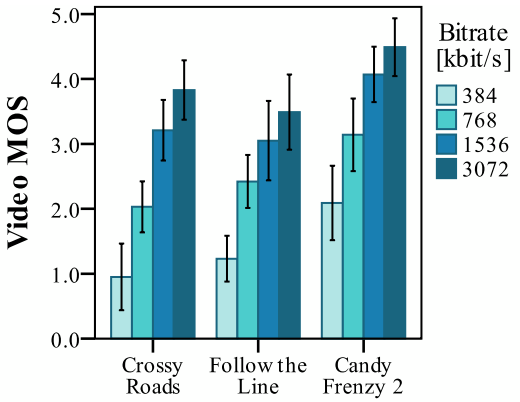
Fig. 5.5 Overall quality MOS, video quality MOS, and SAM ratings for the four tested bit rate levels averaged over all three used games at a 144 ms system delay.

Table 5.3 Results from non-parametric Friedman tests of differences among repeated measures of overall and video quality and the three SAM dimensions with varying video streaming bitrate at a common 144 ms system delay. Significantly influenced items are printed in bold.

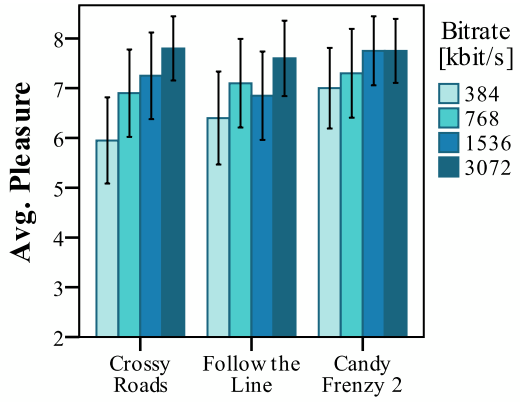
Bit rate influence on:	Crossy Road		Follow The Line 2		Candy Frenzy 2	
	χ^2	Sig.	χ^2	Sig.	χ^2	Sig.
Overall quality	42.58	**p < .001**	28.81	**p < .001**	32.63	**p < .001**
Video quality	47.86	**p < .001**	39.59	**p < .001**	35.89	**p < .001**
SAM Pleasure	20.25	**p < .001**	20.97	**p < .001**	09.77	**p < .05**
SAM Arousal	00.81	p > .05	12.70	**p < .01**	06.44	p > .05
SAM Dominance	21.36	**p < .001**	13.39	**p < .01**	05.45	p > .05



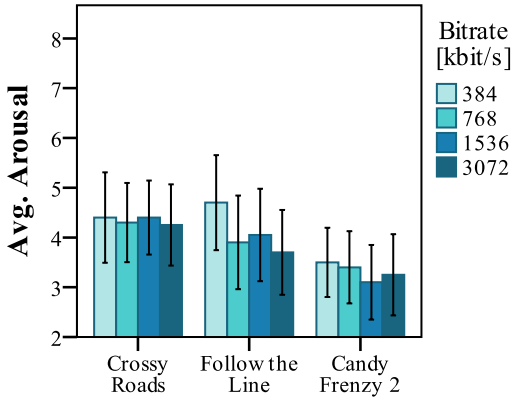
(a) Overall quality
(0: “extremely bad” - 6: “ideal”).



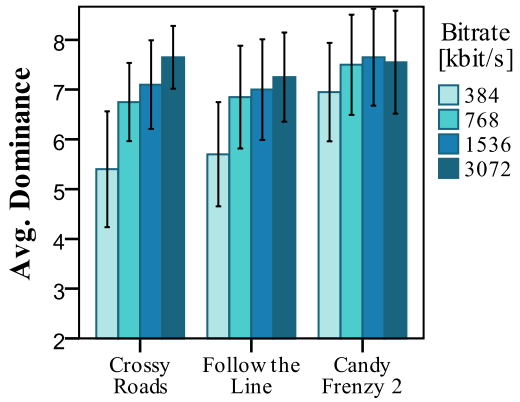
(b) Video quality
(0: “extremely bad” - 6: “ideal”).



(c) SAM: Pleasure
(1: “Annoyed” - 9: “Pleased”).



(d) SAM: Arousal
(1: “Unaroused” - 9: “Aroused”).



(e) SAM: Dominance
(1: “Controlled” - 9: “Controlling”).

Fig. 5.6 Overall quality MOS, video quality MOS, and SAM ratings for the four tested bit rate levels at a 144 ms system delay.

5.5.2 Influence of system delay variation

In this section, the subset of obtained data points with a common system bit rate of 3072 kbit/s but varying system delay levels is analyzed. The mean ratings for overall quality, video quality, and the SAM's Pleasure, Arousal and Dominance dimensions, averaged over the three games, are shown in Figure 5.7.

Despite the small differences in the means of overall quality, the ratings are statistically significantly different according to a non-parametric Friedman test of differences among repeated measures ($\chi^2 = 13.352$, $p < .01$), as the mean rating for the 344 ms delay condition differs significantly from the other delay levels (Wilcoxon Signed-Rank tests, $p < .01$). Similarly, the means of the games' video quality ratings vary significantly with changing delays ($\chi^2 = 15.578$, $p < .01$) as, again, the 344 ms delay condition differs significantly from the other three. Pleasure is also significantly affected ($\chi^2 = 15.730$, $p < .01$). Here, however, a general trend of sinking pleasure with growing delay is registered and the 344 ms level is not exceptionally different. No significant differences exist for Arousal, but Dominance is again significantly affected ($\chi^2 = 13.528$, $p < .01$) as the perception of being in control lowers with growing delay.

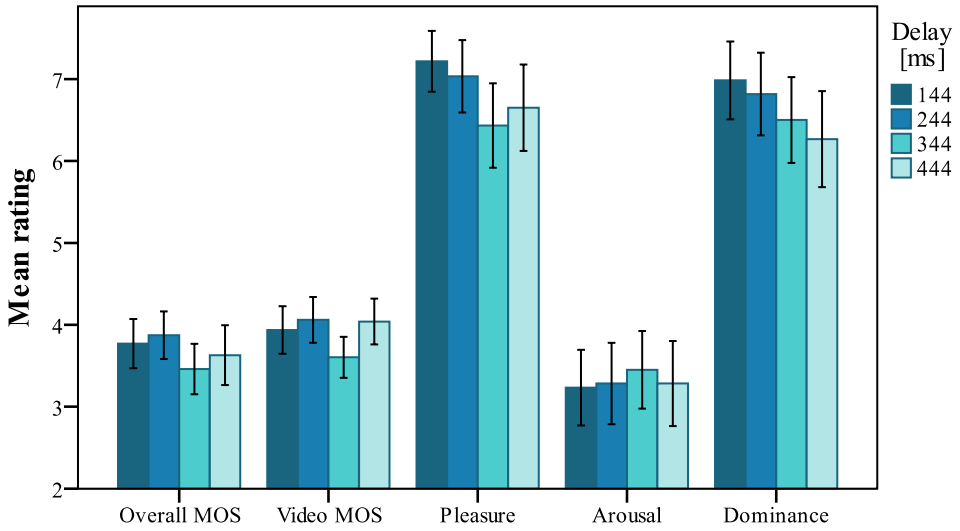


Fig. 5.7 Overall quality MOS, video quality MOS, and SAM ratings for the four tested system delay levels averaged over all three used games at a 3072 kbit/s bit rate.

As the games were thought to differ with respect to their delay sensitivity, the ratings for overall quality, video quality, and the three SAM dimensions are graphed in Figure 5.8, and the results from non-parametric Friedman tests of differences among repeated measures statistically analyzing the effect of delay on the ratings are reported in Table 5.4. For the game *Candy Frenzy 2*, which was considered to be delay insensitive in Figure 5.1, no influence of

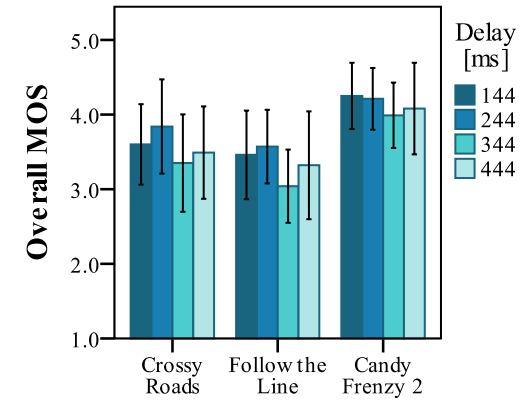
added network delay was found in the collected data. For the other two games, which were considered to be highly delay sensitive, only some dimensions were affected by delay: In *Crossy Road*, additional network delay lowered Pleasure marginally ($\chi^2 = 9.27, p < .05$), whereas in *Follow The Line 2* Pleasure sank slightly more ($\chi^2 = 19.57, p < .001$) and ratings for the Dominance dimension also decreased with growing delay ($\chi^2 = 22.29, p < .001$).

Table 5.4 Results from non-parametric Friedman tests of differences among repeated measures of overall and video quality and the three SAM dimensions with varying system delay at a common 3072 kbit/s bit rate. Significantly influenced items are printed in bold.

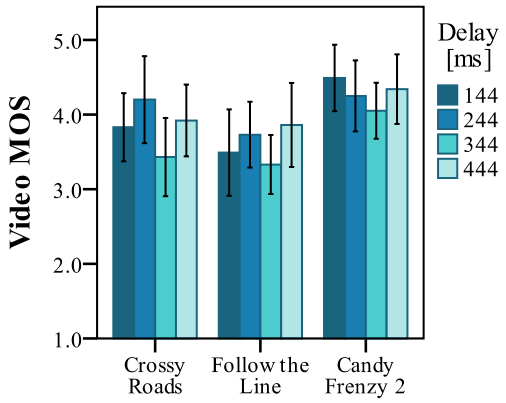
Delay influence on:	Crossy Road		Follow The Line 2		Candy Frenzy 2	
	χ^2	Sig.	χ^2	Sig.	χ^2	Sig.
Overall quality	04.48	p > .05	03.05	p > .05	02.71	p > .05
Video quality	06.45	p > .05	04.54	p > .05	07.39	p > .05
SAM Pleasure	09.27	**p < .05**	19.57	**p < .001**	01.63	p > .05
SAM Arousal	02.51	p > .05	06.47	p > .05	04.44	p > .05
SAM Dominance	07.33	p > .05	22.29	**p < .001**	01.40	p > .05

5.5.3 Influence of combined bit rate and delay impairments

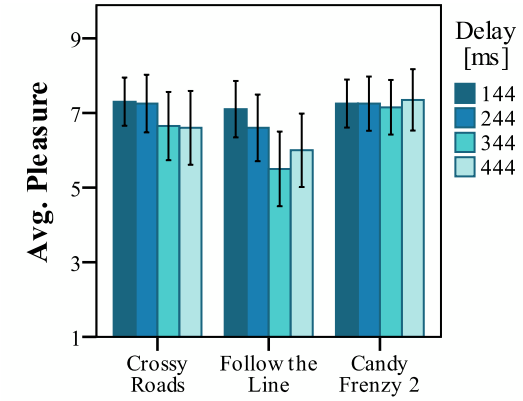
One of the condition combined the worst system delay (444 ms) with the lowest transmission bit rate (384 kbit/s). Overall quality ratings differ not significantly between this and the (144 ms, 384 kit/s) condition for *Candy Frenzy 2* and *Crossy Roads*. For *Follow The Line 2*, however, ratings are significantly different (Wilcoxon Signed-Rank Test, $Z = 16.5$, $p < 0.05$) as the MOS drops from 1.8 ($SD = 1.0$) to 1.3 ($SD = .86$).



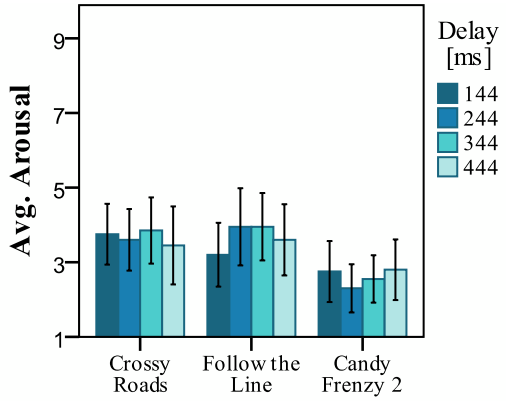
(a) Overall quality
(0: “extremely bad” - 6: “ideal”).



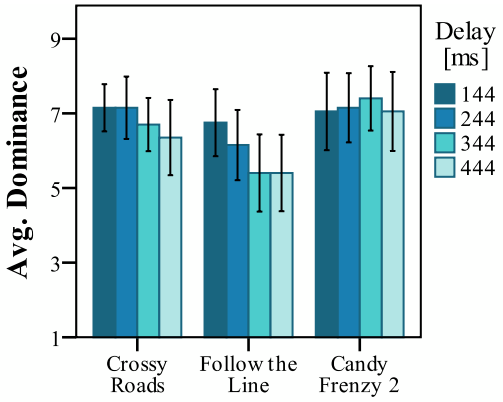
(b) Video quality
(0: “extremely bad” - 6: “ideal”).



(c) SAM: Pleasure
(1: “Annoyed” - 9: “Pleased”).



(d) SAM: Arousal
(1: “Unaroused” - 9: “Aroused”).



(e) SAM: Dominance
(1: “Controlled” - 9: “Controlling”).

Fig. 5.8 Overall quality MOS, video quality MOS, and SAM ratings for the four tested system delay levels at a 3072 kbit/s bit rate.

5.6 Discussion

The results show that variations of visual quality caused by bit rate changes influenced almost all ratings significantly, while for the remaining (Arousal and Dominance for *Candy Frenzy 2* and Arousal for *Crossy Road*) trends are visible in the plots in Figure 5.6. Generally, higher visual quality creates a better mobile cloud gaming experience, as conditions with higher bit rate receive better marks in overall quality, video quality, but also in the SAM's Pleasure and Dominance dimensions. This is both expected and in line with previous works concerning PC or console-based game streaming, e. g., [112]. Furthermore, similar to PC-based cloud gaming, a bit rate of around 3 Mbit/s in *mobile* cloud gaming appears to be the lower boundary for a gaming experience which participants rate as "good" (cf. Figure 5.6).

Games differed in how strong they were affected by lowered video transmission bit rate: While in *Candy Frenzy 2* the drop in MOS from the highest to the lowest bit rate was 2.1 points of the range from 0 ("extremely bad") to 6 ("ideal"), the drop in *Crossy Road* was 24 % higher with 2.6 points (cf. Figure 5.6), showing that this game was stronger affected by the compression's data reduction.

Looking at the absolute bit rate requirements, *Candy Frenzy 2* was rated "fair" at a bit rate of 768 kbit/s, whereas twice the number of bits per second was required for *Crossy Road* to reach the same quality level. For the lower two bit rates, the game *Follow The Line* lies between *Candy Frenzy 2* and *Crossy Road* in overall quality, video quality, Pleasure, and Dominance. Higher bit rates, however, do not seem to benefit the game as much as the other two titles. While participants filled the questionnaires, the smartphone used for playing remained on the table and the last display from the finished session was still visible. In *Follow The Line* the screen following a (failed) session contained a high degree of animations and pulsating buttons causing blockiness to remain visible during the rating process even at the 3 Mbit/s setting. This behavior might have negatively influenced ratings at the higher bit rates.

Considering the selection of the games by their respective SI·TI product, which was 264 for *Candy Crush* / *Candy Frenzy 2*, 854 for *Follow The Line*, and 3465 for *Crossy Road*, the order of the scores matches the observed sequence of the games' overall and video quality ratings for the lower bandwidths. However, beside that matching order, the size of the interval between highest and lowest quality ratings for each game does not match well with the games' strongly different SI·TI products. If that were the case and the relation a linear one, then *Crossy Road* would have had to be rated much worse than the other two titles at lower bit rates, which, on the other hand, should have been rated almost similar with regard to their visual quality. This is not observable in the results. However, multiple players made exclamations with regard to the unacceptably bad quality of the game when

they were first shown *Crossy Road* in the worst condition as part of the training session. It seems, that initial training with the best and the worst quality condition in each gaming block led participants to use these extremes to scale their opinion to the items' rating ranges. Although this procedure was chosen because such initial training is recommended by ITU-T Recommendations P.910 and P.911, it may be detrimental to the external validity of such ratings, as participants compare stimuli to the previously demonstrated extremes rather than to their personal quality expectation.

The completely absent meaningful influence of changed system delay on the participants' ratings of the mobile cloud gaming experience is surprising and in contradiction to prior studies involving PC games: With 300 ms of added network delay, Jarschel *et al.* recorded MOS reductions from 5 to around 3 for a slow game, and from 4.6 to 1.3 for a fast-paced game [71]. When this study's participants were asked what changes they had perceived in the experiment after completing the last condition, only 6 noted having observed "lag" or delayed responses. However, multiple stated during the test or afterwards that the difficulty of the games varied or the game reacted unexpectedly. Participant 17 exclaimed "Here one has no control at all!" (German: "Hier hat man ja überhaupt keine Kontrolle!") while playing *Follow The Line* in a condition with increased delay. The significant influence of delay on the SAM's Dominance dimension (labeled "Controlled" - "Controlling") in *Follow The Line* (cf. Table 5.4) testifies that added delay did indeed influence the game's experience. It seems that the delayed and less predictable behavior was at least partly attributed to the games themselves and not to the gaming system. However, with touchscreen-based game interaction, another explanation is also possible: During interaction (i. e., the touch), the manipulated part of the screen is obscured by the finger. The circle signaling the recognized finger's position in *Follow The Line* (cf. Figure 5.3) therefore is invisible to the player most of the time. Consequently, the effect of the delayed input is not directly visible, but only indirectly perceivable as the game does not respond as expected and collisions with walls in *Follow The Line* are detected by the game, leading to the end of the session, although the player moved his finger properly along the white line. While the consequence of this unperceived late system response is detrimental to the success in the game in *Follow The Line*, it has no consequences in *Candy Frenzy 2*: Sweets, which are moved to an adjecent field in the grid may react later due to the added system delay, but this is not noticed as during the time the finger is still covering the display. The combination of the not perceivable visual effect due to screen obstruction and absent negative consequences of later input in the game may explain the complete absence of observable delay influence for *Candy Frenzy 2* in the ratings in Table 5.4. On the contrary, PC- and console-based games have an different input model, where a player moves his limbs, sings, or manipulates buttons without restraining the game's

output modalities, hence, the effect of actions can immediately be observed. While indirect control of the chicken is also possible in the game *Crossy Road*, many participants were observed to be using a swiping gesture to move the animal across the challenges, thereby again obscuring the immediate effect of their action with their finger or hand.

The reason for the small but significant difference between the 344 ms condition and the remaining delay levels in the averaged ratings of overall and video quality in Figure 5.7 remains unknown, as a technical cause is improbable as the test bed's ability to reproduce the correct delay levels was validated with measurements using the technique described in Section 5.3.5 multiple times over the course of the study.

5.7 Conclusion

In this chapter, an experimental study was presented, which was conducted to test network influence on mobile games implemented using a cloud gaming setup, where the actual game execution is performed on a remote server and the client on the user's device just displays the remotely rendered image and forwards input commands. The obtained results show the expected detrimental influence of reduced network transmission bit rate on virtually all tested quality metrics - an effect that has been shown in the literature for streamed PC and console games but not for natively touch-based mobile games. Despite the demonstrated reduction of quality with lowering bit rate, the obtained results for the tested bit rate levels show that mobile cloud gaming in wireless cellular networks is technologically feasible today at quality levels that were rated as "good" by the participants in the study: The highest used level of 3 Mbit/s can easily be transmitted with modern 3G and 4G networks. However, even with a reduction of the bit rate to 1.5 Mbit/s the MOS averaged over all tested games was still better than "fair". For the least visually complex game used in the test, *Candy Frenzy 2*, an even lower 768 kbit/s is still sufficient for a "fair" rating. In practice, the success of cloud gaming business models will therefore likely not be limited by technical factors such as insufficient transmission bit rate but by service factors: The data volume that is transmitted during extended gaming sessions quickly exceeds a player's phone contract allowance as one hour of gaming at 3 Mbit/s causes around 1.4 GB of compressed video data to be transmitted - more than most contracts in Germany currently allow.

The ratings of the games with added system delay raise new research questions, as a meaningful delay influence in the individual games was only perceived by participants in their feeling of being in control. This finding is in strong contrast to previous works on the effect of delay on player experience in PC or console-based cloud gaming, which found delay to be a highly impairing influence factor. Further research is required to better understand

how touchscreen-based gaming is affected by delay as the touching finger or hand obscures the manipulated part of the screen and a visual response may not be visible to the user regardless of whether it is delayed or not. However, games with a more indirect control model (e. g., soccer games with virtual joysticks such as in Figure 4.2) might be affected by network delay much more than the games used in this study.

This study's results have given rise to doubts whether the training procedure recommended by ITU-T Recommendations P.910 and P.911 is helpful in obtaining externally valid opinions from the participants: When the worst and the best level are shown prior to the ratings, they may serve as references to which all subsequent stimuli are compared to as opposed to the persons individual intrinsic expectations. Since games may be unfamiliar to participants, the training phase cannot be skipped altogether. However, to prevent presenting a quality reference which is common for all participants, the training session could be performed / played using either a randomly chosen test condition, or one that does not represent an extreme of any varied factor instead of always training with the best or worst system setting.

Furthermore, it was found that allowing participants to see the last running game during their rating process may skew the results as the then visible display of the game may not be representative for contents shown during the rest of the session. Instead, displaying a neutral gray as proposed as part of the ACR method in P.910 [68] or hiding the device from the participant during rating may be beneficial.

It is conceivable, that, since participants were not informed about which degradations they were supposed to rate, they were only concentrating on the most obvious difference: visual quality. The very high correlation between the overall quality and video quality items seen in Table 5.2 hints at participants considering the two questions almost equal in this study. Consequently, a between subjects test design with a division of letting one group assess video quality changes and another delay may make study participants more sensitive to the respective changes and let them make more informed opinion ratings.

Although Chapter 3 focused on the influence games and their implementation have on gaming experience, variations of network delay were used there as well to study the games' differing behaviors. The results showed that the particular implementation of a game strongly influences a player's gaming experience with non-perfect network conditions. The differences were so fundamental, that a mathematical model to describe the network influence on locally executed software's experience was considered to be infeasible without additional knowledge about the games' internal mechanisms. In comparison, the results obtained in the present study with mobile cloud gaming are much more homogenous, as all tested games were perceived with lower quality with sinking streaming bit rate (albeit to a different amount) and none of the games' experiences was seriously degraded by the added network delay.

Consequently, developing quality prediction models for mobile cloud gaming should be well feasible. One of the remaining challenges is, however, to find an appropriate characterization of a game's sensitivity to limited bandwidth and, potentially, delay. The SI·TI product may in itself be such a metric or be part of one. Due to the issue with possibly skewed data due to the trainings presenting references as mentioned earlier, the data from the present study is insufficient to decide that, but a point is given from which future research may start.

Chapter 6

Influence of the context

6.1 Introduction

Since *mobile* games are designed to be played in different contexts such as at home, at work, or during commutes[1], a context's influence on gaming experience is of profound interest to researchers and game developers alike. In the study presented in Chapter 4, this influence was investigated by simulating a space-constrained and noisy metro setting. The rated playing experience from that environment was then compared to a standard lab setting. However, no significant differences were observed, leading to the conjecture that the simulation was insufficiently realistic.

As no previous experiences existed, which of the many aspects that differ between a laboratory setting according to ITU-T Recommendations P.910 [68] / P.911 [69] and a metro environment actually influence mobile gaming, a comparative study[2] was conducted, in which one setting was a real metro in the field. The results show surprisingly little influence of the environment on the participants' ratings, as none of the core GEQ dimensions are affected, and only one parameter of the Post-Game Experience Questionnaire (Returning to Reality) differs significantly.

6.2 Related work

In the literature, multiple contributions consider the context of use as influential for exercising gaming activities: Dixon *et al.* studied user requirements for mobile gaming with regard to the usage context in 2002 with a combination of interviews, focus groups, and analysis of

[1] https://gigaom.com/2012/03/22/where-is-mobile-gaming-happening-at-home-in-bed/ (last accessed: 2016-07-01)

[2] The study was conducted in collaboration with Stefanie Hecht as part of a master thesis.

recorded game playing. They found participants to play in numerous contexts for varying motives and that "in different contexts of use, users demand very different experiences from mobile gaming" [39]. An influence of the context was also agreed to by Liu *et al.*, who noted that the "use context strongly and significantly influences the formation of peoples' perceptions of all aspects of mobile games, including perceived ease of use, perceived usefulness, perceived enjoyment and cognitive concentration." However, they attribute these effects more to the players' happiness about evading the boredom of the usage context itself, than to the entertaining factors of the game as "being able to play a game in certain environments, such as during a commute, makes users happy, apart from the playability of the game itself" [85].

The influence of the context was also studied in other media-related domains such as in mobile TV consumption: Jumisko-Pyykkö *et al.* [73] investigated how study participants' ratings for service acceptance, satisfaction, entertainment, and the ability to recognize information in artificially degraded video stimuli varied between three different contexts: waiting for a train at a station, riding a bus, and spending time in a cafe. Although Jumisko-Pyykkö *et al.* found no significant differences for acceptance and satisfaction, participants felt less entertained in the bus scenario and were able to recognize more information in the cafe context.

Consequently, the usage context is assumed to be an influence factor in the QoE community. Indeed, an earlier definition for QoE developed by participants of the Dagstuhl seminar "From Quality of Service to Quality of Experience"[3] in 2009 still read "*Degree of delight of the user of a service. In the context of communication services, it is influenced by content, network, device, application, user expectations and goals, and* ***context*** *of use.*" However, only surprisingly little research exists comparing different contexts regarding their effects on perceived gaming experience.

One comparative study was conducted by Engl in Regensburg, Germany in 2010 with 35 participants, who played a puzzle game and a game of skill in both a living room and a tram of the local public transport system, and rated their experiences using the GEQ (cf.. Section 2.6.3). Although significant differences were found in the Immersion and Negative Affect dimensions as both were rated higher in the stationary context, these variations were very small and the other five dimensions of the GEQ remained virtually unchanged [42]. Two reason give rise to seek a refined repetition of that study: The selected stationary context (meeting room) was in no way resembling a standardized test environment as recommended by the International Telecommunication Union - Telecommunication Standardization Sector (ITU-T) and likely did not remain fully static in the course of the experiment. More

[3] http://www.dagstuhl.de/de/programm/kalender/semhp/?semnr=09192

importantly, however, the participants were allowed to fully concentrate on the game in the mobile context without having to care for leaving the train at the right station. The lack of that secondary task, which binds attention and keeps a part of the perception directed on the environment, may be influential in determining the effect of the environment on gaming. Since the question regarding the gaming context's influence is a pressing one, as it directly determines the external validity of (mobile) gaming experiments conducted in the lab, a need for further research exists, which is addressed by the study in this chapter.

6.3 Methodology

Commutes not only take place in metros as simulated in Chapter 4. Instead, a number of different transportation means like buses, trams, different types of trains, and even ferries comprise the public transportation system. Although each of these subsystems in itself provides a valid context for playing, the number of means of transportation had to be narrowed down to one to make different study participants' experiences more comparable.

Generally, throughout the day, the conditions in public transport systems change significantly. Not only does the number of passengers change in the course of a day, depending on the type of vehicle, other factors like daylight, temperature, or delays due to congested roads might also affect the transit, and indirectly cause variations in playing experience. However, the changes in these context factors vary between different means of transportation. Whereas daylight is directly perceivable in all surface-based vehicles through windows, underground travel is shielded from daylight and the artificial lighting is consistent throughout the day. Furthermore, this part of public transportation is completely unaffected by road congestion and travel times are therefore much more predictable. An underground section of the Berlin metro line *U2* was therefore chosen for the experiment. To give participants enough time for playing and to allow them to get immersed in the game, a route of five stops (between *Ernst-Reuter-Platz* and *Theodor-Heuss-Platz*), taking approximately 7 minutes, was deemed appropriate and sufficiently realistic.

In marked contrast to the noisy public transportation, a very quiet soundproof neutral room in accordance with ITU-T Recommendations P.910 [68] and P.911 [69], equipped with a comfortable armchair and a table, comprised the laboratory setting.

As conducting a field study with a public transportation system required including a considerable margin of error in the test schedule due to possibly (or likely) unpunctual or canceled trains, the test was very time-consuming and the number of participants had to be limited. A within-subjects design was therefore chosen, to still be able to recognize differences in the data, as that design would be more resilient to individual differences

between participants, since the same persons would rate both scenarios. To furthermore prevent order effects, the order of the metro and laboratory settings was balanced.

The same LG Google Nexus 4 device previously used in the study in Chapter 3 was also chosen for this experiment. With its 4.7" screen it is comparable to the 5" Samsung Galaxy S4 used in the study in Chapter 4, which then had allowed favorably-rated gaming sessions and did not impair the participants' gaming experience with an overly small screen size. To prevent test participants from accidentally leaving the game by touching one of the Android system buttons at the bottom of the screen (cf. Figure 6.1), these were deactivated during the experiment. At the time of the study, Android 5.0 (Lollipop) was installed on the Nexus 4 device and the automatic adaptation of screen brightness to the environment light was enabled.

6.3.1 Selection of games

In Section 2.2.1 it was stated, that the player of a mobile game may be interrupted at any point in time by events from either the system (e. g., incoming phone calls, notifications), or from his environment. Whereas a laboratory is explicitly built to eliminate unwanted interruptions, public transport, on the contrary, is usually a rich source of distractions and uncontrolled stimulation. Game titles, whose mechanics require a player to uninterruptedly pay attention, might therefore be experienced differently in a quiet lab than in the noisy public transport. Thus, the primary selection criterion was a game's ability to be interrupted without experiencing disadvantages in the gameplay (e. g., lost points).

Whereas participants could be asked to play the provided games while being seated in the lab, they would have to stand in a crowded metro when no free places were available. Standing in a crowded environment might impair the ability to interact with games controlled by movements (e. g., by tilting the device) due to limited space, or accelerations or decelerations of the vehicle. A secondary selection criterion was therefore established that the games should be controllable solely by touch input. It was furthermore taken care that the selected games could be played without audio feedback.

Candy Crush Saga

As previously presented in Section 5.3.3, *Candy Crush Saga*[4] depicts a matrix of differently shaped and colored little sweets (cf. Figure 6.1) and the player's task is to create a possibly long array of simlar items with a single swap of sweets from adjacent cells. The game is therefore exclusively touch-based, in that a single yet precise touch is sufficient to proceed to

[4]https://play.google.com/store/apps/details?id=com.king.candycrushsaga (last accessed: 2016-04-23)

the next step, and physical movement of the device does not influence the game state in any way.

Fig. 6.1 Screenshot from the game Candy Crush Saga.

Although time-limited advanced game modes exist, the game does not pressure the player to keep his attention at the game. Thus, the gaming session can be interrupted at any time without negative consequences in the game.

Smash Hit

Smash Hit[5] revolves around smashing glass pyramids and obstacles on the way while the camera unstoppably moves forward through a three-dimensional world as depicted in Figure 6.2.

The player smashes glass items by "throwing" shimmering metal balls in their direction. These balls are launched by tapping on the touchscreen at or above the desired target. Since the balls are always launched with equal velocity and follow a downward-bent path due to simulated gravity, the player has to adjust how far above an item he aims. The game furthermore requires the precise timing of touches to hit the targets, as the game world passes by at an increasing speed.

[5] https://play.google.com/store/apps/details?id=com.mediocre.smashhit (last accessed: 2016-04-23)

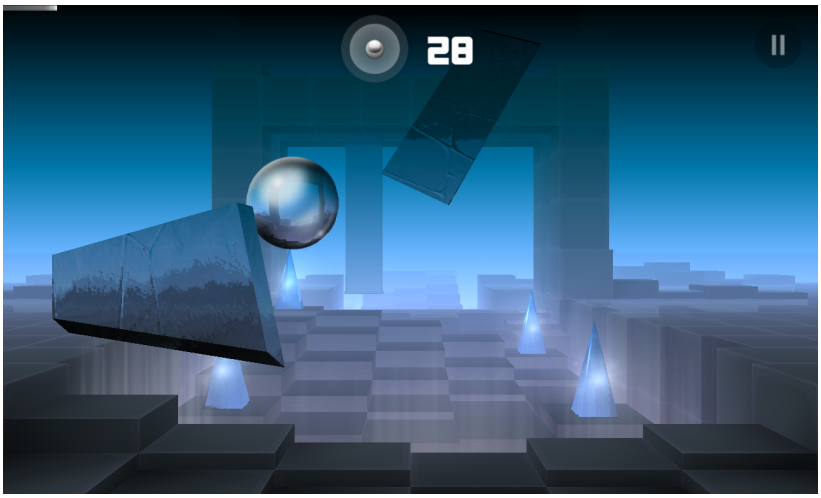


Fig. 6.2 Screenshot from the game Smash Hit.

To challenge the player, the supply of usable balls is limited (cf. number at the top center of the screen in Figure 6.2) and decreases as they are thrown. Additional balls are earned by smashing glass pyramids. In the course of the game, obstacles repeatedly appear. These have to be removed by throwing balls at them. As collisions with these barriers are penalized with the reduction of the player's available ball count, even brief distractions of the player can be detrimental to his success in the game. Nevertheless, the game can be paused, but only through active intervention by the player. In this aspect, *Smash Hit* differs from *Candy Crush Saga*, which requires no such action to prevent unfavorable events in the game from happening. The session ends when the supply of balls is extinguished, upon which the achieved "distance" of travel in the game is recorded and added as new high-score if it exceeds previous achievements.

6.3.2 Measurement instruments

As in the previously presented studies, the Game Experience Questionnaire (GEQ) from IJsselsteijn *et al.* was used to elicit the Player Experience dimensions. However, in this study, besides the 36-item core module, the so-called Post-Game Experience Questionnaire (PGQ) with 17 items was made use of, which is aimed at determining how people feel after they finish playing (cf. Section 2.6.3, [100]).

In an effort to determine and characterize the different physical properties of the contexts in the study, ambient brightness and loudness were measured. For this purpose, two specialized devices were procured:

The *Sekonic L-758 Cine Digital Master*[6] is a portable light-measuring tool primarily built for cinematographers, videographers, and photographers. The device is capable of measuring the intensity of light emitted or reflected by a surface, or of metering the ambient brightness by gauging the amount of light shining upon a white carlotte.

The *NTi Audio AL1 Acoustilyzer* is a professional handheld sound pressure level meter[7]. The device requires an additional measurement microphone, as which an *NTi Audio MiniSPL* was used.

6.4 Test procedure

Participants for the study were recruited using a web platform[8] and invited to appear at an appointed date and time at the laboratory at Technische Universität Berlin. Upon their arrival, they were welcomed and informed to turn off their own phones for the duration of the test. Before the actual beginning of the experiment, they were asked to fill a demographic questionnaire. Due to the effort and time required to walk to and from the underground station, the order of the test conditions could not be fully randomized (i. e., completely randomly mixed metro and laboratory conditions). Instead, one group first played in the metro context, whereas the other started in the laboratory. The assignment of participants to these groups was balanced. Participants due to start with the metro context part of the experiment would be accompanied from the laboratory to the *U2* station after being instructed and introduced to the games used in the test. The subjects were briefed to independently enter the train, play the game in question and leave the train at the destination station on their own. During the passage, the experimenter would maintain a distance but keep the participant in sight. This was done, as people had relied on the experimenter to tell them when to disembark in the pre-test. The conductor would intervene only if the participant failed to notice the station. During the approximately 7 minutes in the metro, the tester took notes about the approximate number of persons in immediate proximity of the test participant, observations about special events, and measured the ambient brightness and noise using the respective devices. After they had left the train, the participants were immediatly met by the experimenter, asked to seat at the station and fill the supplied questionnaire. Afterwards they would wait for the return train and, during that passage, play the other assigned game. The laboratory part was comparable in nature to the previously presented studies: Participants sat on a comfortable chair next to a table and were not instructed to hold the device in any particular way. After

[6] http://www.sekonic.com/germany/products/l-758cine/overview.aspx (last accessed: 2016-04-24)
[7] http://www.nti-audio.com/en/products/acoustilyzer-al1.aspx (last accessed: 2016-04-24)
[8] https://proband.prometei.de (last accessed: 2016-04-24)

the beginning of a session, the experimenter left the room, and only came back if questions arose or the time for playing had passed. After this, the participants filled the questionnaire and proceeded with the next game. After participants had completed both the metro and the laboratory part of the experiment, they were interviewed about notable observations, thoughts, and opinions, and received a financial compensation for their participation.

In total, each test run took approximately two hours. Throughout the test, the order of the played games was randomized. Each game was tested twice - once in the metro context and in the laboratory each.

The study took place in the days from 2014-12-12 to 2014-12-19. A total of 30 people were invited, but only 26 showed up as 4 failed to appear. Among the participants, 16 were female and 10 male. Their ages ranged from 18 to 35, with a mean age of 26 years ($M = 26.54$). A university degree was possessed by 14 persons (53.8 %). 20 were students (76.9 %) at the time of the study, whereas one person was receiving non-university education, and five (19.2 %) were employees. 17 participants (65.3 %) were familiar only with the game *Candy Crush Saga*, one person knew only *Smash Hit*, and two had played both games before. The majority of participants was unfamiliar with the section of the metro line *U2* used in the test ($n = 17$, 65.4 %).

With 26 subjects and the two independent variables with two levels each (game 1: *Candy Crush Saga*, game 2: *Smash Hit*, context 1: laboratory, context 2: metro), 104 game sessions were played and a total of 5.512 data points was generated using the 36 items from the GEQ and the 17 items from the PGQ.

6.5 Results

The participants' ratings on the ACR scales were coded with 0 = "gar nicht" (i. e., "not at all") to 4 = "außerordentlich" (i. e., "extremely"). From the two questionnaires' coded data, the GEQ dimensions (Competence, Immersion, Flow, Tension, Challenge, Negative Affect, and Positive Affect) and the PGQ dimensions (Positive Experience, Negative Experience, Tiredness, and Return to Reality) were then calculated following [100]. These computed dimensions are henceforth used as the 11 dependent variables of this test. Error bars in this chapter refer to the 95 % confidence interval.

The GEQ and PGQ ratings grouped by the context are shown in Figure 6.3 for the game *Candy Crush Saga* and in Figure 6.4 for *Smash Hit*. To test, whether significant differences between the two contexts and games exist, each dimension's data from each condition was first checked for normal distribution using a Shapiro-Wilk test with a significance threshold of 0.05. This showed, that the ratings were only normally distributed in each condition for

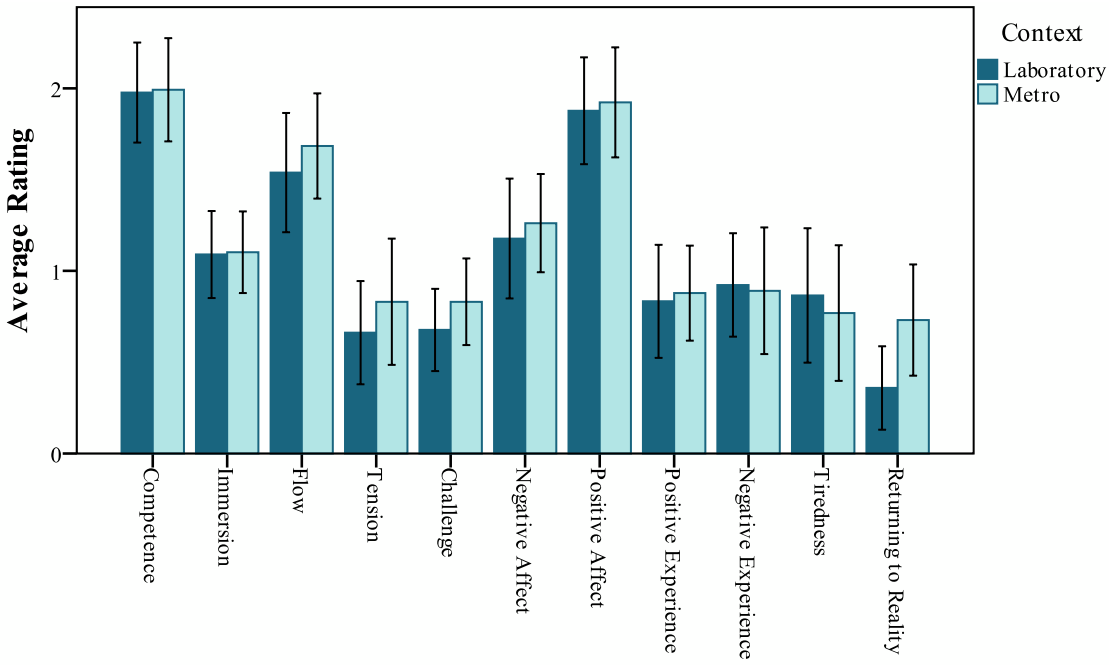


Fig. 6.3 Player Experience and Post-Game Experience Questionnaire dimensions for the two contexts Metro and Laboratory for the game *Candy Crush Saga*.

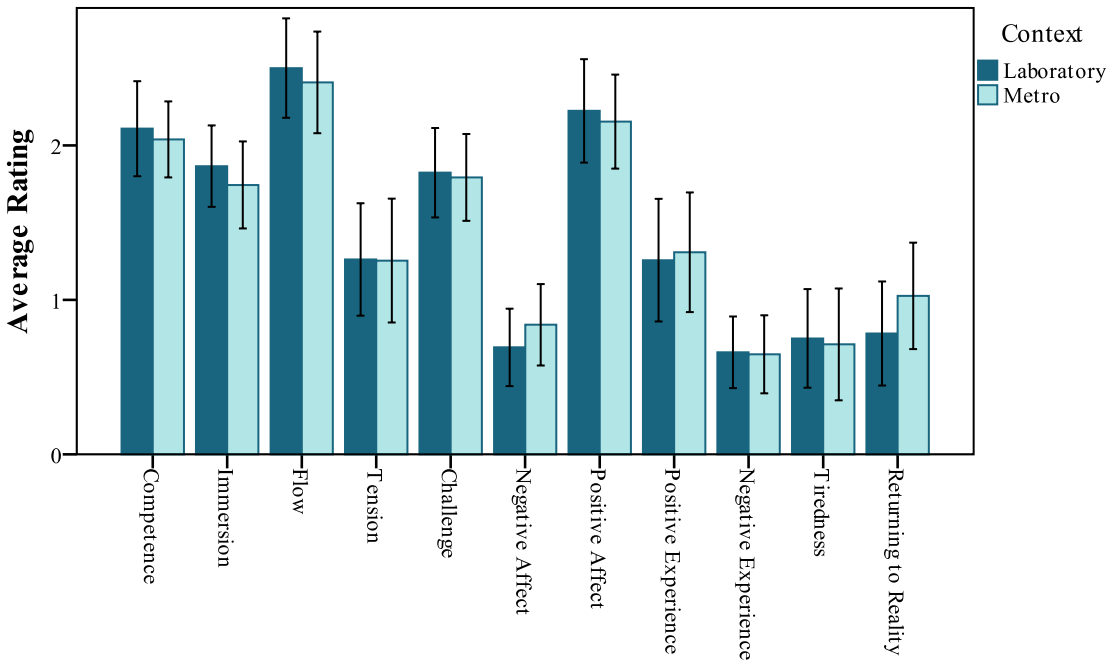


Fig. 6.4 Player Experience and Post-Game Experience Questionnaire dimensions for the two contexts Metro and Laboratory for the game *Smash Hit*.

Table 6.1 Statistical analysis of the influence of the independent variables game and context on the Player Experience dimensions Competence, Flow, and, Positive Effect using a repeated measurements ANOVA.

Dimension	Game			Context		
	Sig.	$F(1,25)$	η^2	Sig.	$F(1,25)$	η^2
Competence	$p = 0.481$	0.511	0.020	$p = 0.742$	0.111	0.004
Flow	$p < 0.001$	42.845	0.632	$p = 0.841$	0.041	0.002
Positive Affect	$p = 0.072$	3.542	0.124	$p = 0.922$	0.010	0.000

Table 6.2 Statistical analysis of the context influence on the dimensions Immersion, Tension, Challenge, Negative Affect, Positive Experience, Negative Experience, Tiredness, and Returning to Reality. A non-parametric Wilcoxon signed-rank test was computed for the two games separately. A significance ($p < 0.05$) means an influence of the context exists.

Dimension	Sig. (*Candy Crush Saga*)	Sig. (*Smash Hit*)
Immersion	$p = 0.777$	$p = 0.253$
Tension	$p = 0.165$	$p = 0.888$
Challenge	$p = 0.107$	$p = 0.610$
Negative Affect	$p = 0.662$	$p = 0.291$
Positive Experience	$p = 0.760$	$p = 0.577$
Negative Experience	$p = 0.635$	$p = 0.654$
Tiredness	$p = 0.361$	$p = 0.450$
Returning to Reality	$p < 0.01$	$p = 0.098$

the dimensions Competence, Flow, and Positive Affect. For these dimensions, a repeated measurements ANOVA with the independent variables context and game was computed. The results are listed in Table 6.1. To analyze the remaining dimensions for a context influence, a non-parametric Wilcoxon signed-rank test was employed to test the significance of differences between ratings for the non-normally dimensions separately for the games. The results from these computations are listed in Table 6.2.

6.5.1 Ambience measurements

As part of the metro passages, two measurements of ambient iluminance and sound pressure levels were conducted during each ride: one while the train was on its way from one station to the next (i. e., without external light sources), and another while the train was at a station. The means of the observations are reported in Table 6.3.

Table 6.3 Recorded ambient iluminances and sound pressure levels (L_{eq} measured over 20 s with dB(A)) in the Berlin metro line *U2* and in a sound-proof laboratory room (L_{eq} measured over 5 minutes with dB(A)) at Technische Universität Berlin.

Setting	Ambient Illuminance		Sound Pressure Level	
	M	*SD*	*M*	*SD*
Tunnel	28.85 lx	9.81	71.8 dB(A)	4.52
Station	31.12 lx	9.79	70.3 dB(A)	3.66
Laboratory	60 lx		37.0 dB(A)	

6.6 Discussion

When comparing the average ratings for all dimensions except "Returning to Reality" in Figure 6.3 (Candy Crush) and Figure 6.4 (Smash Hit), the level of similarity between the laboratory and the metro settings is staggering. All Player Experience dimensions are virtually identical in the quiet lab and in the much more noisy and dimly lit (cf. Table 6.3) metro. This observation from the graphs is reflected in the results from the statistical analyses in Table 6.1 and Table 6.2: Except for "Returning to Reality", no significant influence of context could be found on any of the dimensions and the Null hypothesis has instead to be adhered to: There is no effect. While this lack of influence is in line with the study results in Chapter 4, it contradicts findings of Engl [42], who found significant, yet small differences for the dimensions Immersion and Negative Affect. These effects cannot be found in the current data: While Engl observed consistently higher ratings for Immersion and Negative Affect in the stationary setting, there is not even a trend for Immersion in the present data, and the trend for Negative Affect, small and insignificant as it may be, even goes in the opposite direction: Here, the laboratory condition is rated marginally more favorably (i. e., lower rating for Negative Affect). However, the present study and [42] agree in so far, as both find no difference in the dimensions Competence, Flow, Tension, Challenge, Positive Affect, Positive Experience, and Negative Experience.

The only dimension showing a significant, but small effect is "Return to Reality" (cf. Table 6.2) for the game *Candy Crush Saga*. The same trend is visible for the game "Smash Hit", but fails to reach significance level ($p = 0.098$). This finding is consequently in partial contradiction to [42], who found no significant variation in this dimension regardless of the game. The corresponding questionnaire items to this aspect are: "I found it hard to get back to reality", "I felt disoriented", and "I had a sense that I had returned from a journey". Considering, that participants agreed with these statements more after playing in the metro than in the lab, it might mean that the game let them forget their environment more than in

the laboratory, which would substantiate the claim of Dixon *et al.* that mobile gaming is also an effective form of social isolation to avoid undesired contact. There is, however, another possible explanation: The situation following the end of gaming differed notably between the lab and the metro context: Whereas the participants remained seated in the lab and started filling the questionnaires without further interruption, they first had to leave the possibly crowded train, walk to a seat in the station, and find the tranquility to fill the questionnaires as part of the metro conditions. The degree of accordance between the ratings from both contexts is even more surprising in this respect.

6.6.1 Limitation

Since this study was conducted using a within-subjects design, each participant experienced all conditions and therefore the same persons rated the games in the metro and in the laboratory environment. As such, it is theoretically possible, that the ratings from the first experience of a game were memorized and then influenced the scoring after the second encounter with the game in the other context. Since the two games were, however, consistently rated differently (i. e., *Candy Crush Saga* ratings differed from *Smash Hit* ratings), independently from the context, an indication exists, that the participants did indeed rate their experience with the games anew. The theory of memorized questionnaire responses is furthermore made less likely by the high number of items (53, cf. Section 6.3.2) to remember.

Movement-controlled games might have been more difficult to play in an accelerating and shaking environment, such as a metro, compared to a static setting, such as a laboratory. The exclusion of such games might therefore incorrectly have shaped the results. In the previously cited study from Engl, one of the employed games was a skill game requiring the player to navigate a ball through a maze using careful movements of his device[9]. However, the decision to avoid movement-controlled games did not seriously restrict the choice of available games, as many popular titles depend on touchscreen input only.

While the metro is an oft-used means of public transportation, it is not representative for all travel options. Surface-bound or aerial travel is subjected to daylight and sunshine, which might cause complications with screen readability on a smartphone due to the high ambient brightness.

[9] https://en.wikipedia.org/wiki/Super_Monkey_Ball (last accessed: 2016-04-25)

6.7 Conclusion

In this chapter, a study was presented, which examined the influence of a player's context on his playing experience. With a metro and a laboratory environment, two very different contexts were chosen and evaluated in the test. The study's results show that the participants' Player Experience was similar in both settings, and ratings coincided to a surprising degree. The only difference observed was an increased perception of change when ending the game and returning to reality.

The core finding of this study is very meaningful for the research of game interactions: Instead of having to arrange complex field studies with plenty of uncontrolled factors endangering the success of the experiment, easier and more controlled laboratory studies can be conducted, as they are not significantly different with respect to the gamer's Player Experience ratings. The results imply, that data obtained in a laboratory environment can be ecologically valid, despite the artificial nature of a lab setting.

Chapter 7

Considerations on test methodologies

As opposed to other forms of media quality assessment such as with video or audio stimuli, a standardized test paradigm is not yet established for the research of gaming quality. A variety of assessment methods have been used in the literature, but their results are often difficult to compare because little information exists on how these methods influence and shape the obtained results. In this chapter, two comparative studies are presented, which explore promising new means of assessment.

A common denominator between virtually all gaming studies is that participants actively, or rather: interactively, play games. However, for some aspects of quality, it might be sufficient to just view recordings of game-play to appraise differences between conditions. This viewing-only approach heralds the advantages of being significantly more efficient at evaluating large quantities of stimuli, maybe being even more effective and sensitive as individual rater's gaming capabilities are not influencing the progression of events within the game and they can focus on visual and audible quality aspects, but most notably, viewing-only tests can be performed in a comparable manner, as they are standardized in ITU-T Recommendations P.910 [68] and P.911 [69]. In the first presented study, the interactive and passive (i. e., viewing-only) test paradigm are compared. For this purpose, participants interactively play scenes from two games and rate them, but also perform an assessment of audiovisual stimuli showing further scenes from the same two games, which have been subjected to the same degrees of visual quality degradation.

In the second presented study, the appraisal of stimuli using self-assessment is compared to the evaluation using physiological methods. These physiological methods, such as electroencephalography (EEG), might have the potential to obtain quality-related information from participants without disturbing their gaming experience and requiring them to reflect on and actively scale their impression. For this purpose, participants played long-lasting sessions of an FPS with strongly different degrees of visual degradation, while an EEG

device continuously recorded voltage differences on their scalp. Afterwards, they rated their experience with self-assessment questionnaires.

As the entirety of influencing factors on gaming quality is not yet well understood, both studies attempted to adhere closely to the aforementioned ITU-T Recommendations P.910 and P.911. These documents not only specify properties of the environment, in which comparable tests are conducted, but also contain guidelines on various aspects of the stimuli presentation. Among these are parameters which concern the screen used to present stimuli and the participants' viewing distance. These specifications are difficult to comply to with mobile devices, as participants can and should hold them, as they see fit. Furthermore, the second presented study required players to be immersed in the game for prolonged periods of time without getting bored. Although such mobile games exist, they are rare. Console or PC games, on the contrary, may often entertain players for multiple hours at a time. Foremost, however, non-mobile games allow the player to remain seated virtually motionless, with just a slight shifting of their arms, as they operate the controls. This was considered as highly beneficial for the EEG usage, since its electrodes may slightly shift and lose electrical contact to the scalp and therefore be negatively influenced by unnecessary player movement. Consequently, for these studies, PC games were used and controlled using keyboard and mouse, or with a gamepad.

As opposed to the mobile games used in the previous chapters, these PC games allow the player a high degree of freedom in choosing his actions. To maintain sufficient similarity between different players' gaming experiences, participants were therefore instructed which goals to achieve, or what tasks to complete in the game.

7.1 Comparing interactive and passive test methodologies

Passive tests are an established method for the assessment of audio, video, and audiovisual material. In an ITU-T standardization meeting of Study Group 12 in 2014, Briard et al. proposed using passive test methodology as a supplement to interactive playing of games to assess their visual quality [61]. However, passivity precludes subjects from exercising effort and influencing the progression of events, which touches the core of the definition of a game: According to the Classic Game Model of Juul cited in Section 2.2, a game is where "the player exerts effort in order to influence the outcome;" and it is required to have "variable and quantifiable outcomes" [74]. In a passive test paradigm it is, however, impossible to influence the outcome because it is not variable. On the other hand, the variability of games is, in part, responsible for the complexity of experiments involving gaming, as this allows different participants to have diverging experiences with a game depending on their skill

and the effort they exert in the game. In a passive test, on the contrary, all participants are compelled to witness the same game experience, opening the possibility of more comparable and sensitive ratings. To research, how well this passive rating reproduces the results from a realistic interactive gaming experience, is the subject of the study[1] presented in this section.

7.1.1 Passive (non-interactive) audiovisual test methods in ITU-T Rec. P.911

A passive audiovisual test paradigm for the assessment of multimedia applications is standardized by the ITU-T in Recommendation P.911 [69]. This document contains recommendations about properties of the used stimulus material, test designs and procedures, viewing and listening conditions, selected subjects, and their instruction.

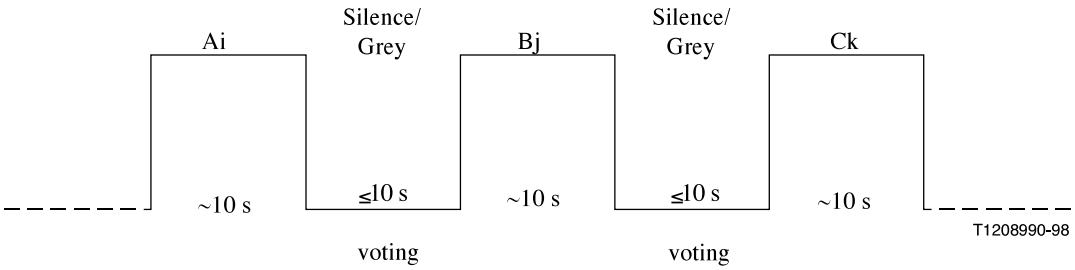
Following ITU-T Recommendation P.911, source stimuli should last around 10 seconds and be of the highest possible quality. These stimuli should contain at least four different types of scenes to avoid boring test participants, be relevant for the service, and span the full range of spatial and temporal information which might be of interest for the users of the service. The recommendation then suggests four different methods for rating these stimuli:

With the Absolute Category Rating (ACR) method, stimuli are presented sequentially. After each stimulus, the screen turns gray for up to 10 seconds (cf. Figure 7.1), during which participants are required to rate the previously seen scene on a scale, for which ITU-T Rec. P.911 offers a five- and a nine-level quality scale recommendation. Both scales have in common, that participants rate absolute quality without being given a direct reference to compare to. To make their rating decision, subjects therefore have to refer to either an intrinsic reference, or a previously performed training session.

The Degradation Category Rating (DCR) method, on the other hand, requires stimuli to be presented in pairs of (undegraded) reference and processed stimulus, where the latter is expected to be the result of sending the former through the system under test. As with the ACR method, a 10-second gray pause is used to let participants rate the perceived degradation on a 5-point ACR scale, which is labeled with regard to the perceived change in quality from "Imperceptible" to "Very annoying".

Like the DCR method, the Pair Comparison method (PC) requires stimuli to be shown in pairs. However, the reference is substituted for another processed version of a stimulus. Therefore, using PC, all possible combinations of processing parameters (i. e., the way the

[1] The study was conducted in collaboration with George Göksel as part of a bachelor thesis and was presented at the ITU-T at a meeting of Study Group 12 in June 2016 [62].



Ai Sequence A under test condition i
Bj Sequence B under test condition j
Ck Sequence C under test condition k

Fig. 7.1 Stimulus presentation method with ACR method standardized in ITU-T P.911[69].

various system configurations may influence the stimulus quality) can be put into relation to each other.

The last proposed method, Single Stimulus Continuous Quality Evaluation (SSCQE), is intended to be used on long-lasting stimuli (3-30 minutes). Here, subjects are supposed to use a physical slider with range from 0 to 100 ("perfect quality") to continuously rate their experience without being given a prior reference. Neither does ITU-T Rec. P.911 contain information on the meaning or label of the lower end of the scale, nor does it tell how frequently participants should update the slider position to reflect their experience.

In the study discussed in this section, the ACR method was employed as a passive tool, which is motivated in the next section.

7.1.2 Methodology

As a prerequisite for the study, a reliable cloud gaming setup to create visually degraded image in real time and a method to record the system's output on the client side as video were needed. Since two state-of-the art games were intended to be used, in order not to prematurely limit ratings to the lower end of the quality scales due to aged games' technically outdated visual output, the open-source GamingAnywhere platform [55] could not be used: For contemporary games, this platform lacks adapters to efficiently grab their visual output and feed back player commands. As an alternative, the commercial and closed-source Steam In-Home Streaming[2] cloud gaming system was employed. This platform is intended to allow players to stream a game from one computer in a home LAN to another possibly less powerful device. In contrast to GamingAnywhere, Steam features connectors for obtaining

[2]http://store.steampowered.com/streaming/ (last accessed: 2016-05-01)

rendered images and audio from recent games and has an optimized and fast encoding pipeline minimizing the delay between player input and system response. In the software's user interface, it is possible to set the transmission bandwidth. However, it was found that the offered granularity (Automatic, 3 Mbit/s, 5 Mbit/s, 10 Mbit/s, 15 Mbit/s, 20 Mbit/s, 25 Mbit/s, 30 Mbit/s, or unlimited) is not fine enough as already a stream with 10 Mbit/s showed virtually no visible compression artifacts. Through direct manipulation of a configuration file (*localconfig.vdf*), arbitrary bitrates could be defined nevertheless. The set transmission bit rate only affected the video compression bit rate. The audio processing remained unchanged. As part of a pre-test, suitable configurations of server and client were evaluated. A powerful ASUS G751JY notebook (Intel Core i7 4720HQ with 2.6 GHz, 16 GB RAM, NVidia Geforce GTX 980M) with the then latest beta version of Stream In-Home Streaming as of December 2015 comprised the server component, whereas a DELL Precision T1500 (Intel Core i5, 2.67 GHz, 6 GB RAM, NVidia Geforce GTX 950) with the same version of Steam constituted the client. To minimize the input delay of the setup, the Xbox controller to be used by the participants was directly connected to the server instead of the client. While this would be unrealistic in a real cloud gaming setup, the focus of this study on visual degradations makes this a helpful "short cut" to reduce a possibly interfering delayed system response. The client computer was equipped with a 26" ViewSonic VP2650wb[3] screen with a native pixel resolution of 1920x1200. Game sound was rendered through a pair of Fostex PM0.4 studio monitor loudspeakers, which were placed on the player's desk appropriately for stereo playback. The network connection between the server and the client consisted of a direct (i. e., without a network switch) Gigabit Ethernet connection.

It was found, that, despite sufficiently powerful server and client hardware, some combinations of frame capturing methods and encoders in Steam led to random bandwidth-independent frame losses. This was considered a bug in the beta software and mitigated by choosing a configuration which did not produce these errors. On the server side, game images were obtained using the "Game polled D3D11 NV12" method, which creates a copy of a game's invisible frame buffer when the game has finished rendering using Direct 3D 11 (D3D11) and switches it to be the player-visible front buffer (therefore game polled). This frame is copied to the computer's main memory (i. e., Random Access Memory (RAM)) in the *NV12*[4] YUV 4:2:0 chroma subsampled pixel format and subsequently compressed into a video stream using the libx264[5] video compression library. Although this technique of software-based video compression placed additional stress on the server's CPU, its computational power did not limit the system in any observable way. On the client side, however,

[3] http://ap.viewsonic.com/me/products/lcd/VP2650wb.php (last accessed: 2016-05-01)
[4] http://www.fourcc.org/yuv.php#NV12 (last accessed: 2016-05-01)
[5] https://www.videolan.org/developers/x264.html (last accessed: 2016-05-01)

the use of the hardware-accelerated codec "DXVA H.264 Decoder" (DXVA: DirectX Video Acceleration) was found to perform best without interfering with the software used to create video recordings for the study's passive rating task. For these video captures, *NVidia Shadowplay*[6] was used. This tool uses hardware capabilities of an NVidia GPU to grab frames from the video memory and directly compress them using a hardware codec on the GPU to an H.264 stream with 4:2:2 chroma subsampling. Although this chroma subsampling in the recording removes perceptible information despite the use of a very high compressor bit rate, it was deemed to be imperceptible, as the streamed frames were already downsampled on the server side to an even lower resolution of YUV 4:2:0. To avoid scaling artifacts, all involved systems and the *Shadowplay* software were configured to use the same 1920x1200 resolution at 60 Hz, which, as state above, was also the native resolution and refresh rate of the used display.

Since, as also mentioned above, no visible compression artifacts were noticeable at 10 Mbit/s, the even less compressed 100 Mbit/s bit rate was considered undegraded and used as one of the bit rate conditions. Below that, three further rates were chosen at 3 Mbit/s, 4 Mbit/s, and 5 Mbit/s, as these bit rate's degradations were on the one hand not severe enough to make playing impossible, and on the other hand feature a highly noticeable improvement of quality with each step.

In order to adhere to the requirements for audiovisual quality assessments defined in ITU-T Recommendation P.911 [69], the test was conducted in a standard-compliant neutral room with thick sound-absorbing gray curtains and daylight-imitating lamps. In order to obtain data points in the passive test, which could be compared to ratings from the interactive test, the ACR method from ITU-T Rec. P.911 was chosen for the video assessment. While the DCR and PC methods are a promising means to investigate the deterioration caused by video compression, they can on principle not be used to rate an interactive scene since a comparison of the current stimulus with another quality level would require a repetition the same of actions in a very short time scale which is not feasible with games.

Questionnaires

To measure the perceived quality of stimuli, three continuous rating scales were used as in Figure 2.4 to rate overall quality, video quality, and audio quality. These scales use the same core set of labels as the 5- or 9-point scales proposed in ITU-T Rec. P.911, but add overflow items to both ends of the scale, which may be used by participants if they had already rated a previous element at an extreme end of the core scale and want to stress that the current stimulus is even more extreme (cf. Section 2.6.6). The interactive sessions

[6] http://www.geforce.com/geforce-experience/shadowplay (last accessed: 2016-05-01)

were furthermore rated using the shortened In-Game Experience Questionnaire module of the Game Experience Questionnaire (GEQ) (cf. Section 2.6.3) and the Self-Assessment Manikin (SAM) (cf. Section 2.6.4). In order to assess the participants' wakefulness, the Karolinska Sleepiness Scale (KSS) (cf. Section 2.6.5) was used at the beginning and the end of an experiment.

Selection of participants and games

Prior to the study, potential participants were asked about their gaming habits and preferred games as part of a web survey.

Since cloud gaming services are used mainly by casual gamers [101] and have been shown to be more positively experienced by this group of players [105], persons playing up to 10h per week were preferred for this study. This furthermore mitigated the potential problem, that highly experienced participants might rate stimuli excessively bad due to disappointed inordinately high expectations. Another criterion used in the selection of participants was their average played session length: Since the test was expected to require close to two hours of concentrated playing and rating, persons were preferred who had stated to typically play at least one to two hours continuously.

In contrast to the other studies portrayed in this thesis, the selection of games in this case was guided by potential participants' preferences to make it more likely that the arising gaming scenarios would be realistic and intrinsically motivating and natural for the particular group of participants. This approach was chosen to employ games, which would cause the players to exert effort beyond the sole fulfillment of their task, feel intrinsically "emotionally attached" to the game's outcome (cf. Section 2.2), and potentially concentrate more on the game's content rather than the displayed visual quality of it. That emotional attachment was considered to be easier to reach in an at least rudimentary familiar gaming environment. The selected games were Grand Theft Auto 5 (GTA V)[7] and Call of Duty: Black Ops III (CoD)[8].

GTA V is an open world action adventure game published by Rockstar Games, first released in September 2013 for the Xbox gaming console. It allows players to freely explore a fictional state called San Andreas and fulfill various missions, of which most require committing crimes, to proceed in one of the game's three main story lines. GTA V incorporates elements from various game genres, as it allows players to, e. g., race cars, fly different kinds of aircrafts, operate tanks, and shoot guns in a first- and third-person perspective (cf. Figure 7.2). It is both one of the most expensive games ever created[9] and

[7] http://www.rockstargames.com/V/info (last accessed: 2016-05-02)

[8] https://www.callofduty.com/blackops3 (last accessed: 2016-05-02)

[9] http://www.ibtimes.com/gta-5-costs-265-million-develop-market-making-it-most-expensive-video-game-ever-produced-report (last accessed: 2016-05-02)

one of the commercially most successful titles[10]. To mitigate changes of the daytime or

Fig. 7.2 In-Game Screenshot of GTA V during a fight scene played in third-person perspective.

weather situations in the game world and keep the scenarios constant and comparable, a game modification[11] was utilized.

CoD is a First-Person Shooter (FPS) placed in a fictional world in the year 2065. The game was released in November 2015 and received critical acclaim[12]. As is typical for games of the FPS genre, the player has to stand through swift battles and shoot enemy fighters and robots using a diverse arsenal of weapons (cf. Figure 7.3).

Both selected game were technologically state of the art at the time of the test and would therefore likely satisfy participants' expectations of game play and aesthetics. From both games, four scenes were selected, each of which could be resumed without significant delay in case of the character's death and were impossible to complete within the time limit of three minutes. While all selected scenes in GTA V included driving tasks, some also required the player to defend objects and follow instructions from other characters in the game. The scenes selected in CoD, on the other hand, all revolved around following a path and identifying and eliminating opponents.

To produce stimuli for the passive rating test, prolonged sequences were recorded from both games while playing different missions with each of the four selected bit rates. From these, 10 second segments were extracted, so that the individual stimuli had preferably little

[10]http://www.polygon.com/2013/10/9/4819272/grand-theft-auto-5-smashes-7-guinness-world-records (last accessed: 2016-05-02)

[11]https://de.gta5-mods.com/scripts/simple-trainer-for-gtav (last accessed: 2016-05-02)

[12]http://www.ign.com/articles/2015/11/06/call-of-duty-black-ops-3-review (last accessed: 2016-05-02)

Fig. 7.3 In-Game Screenshot of CoD during a fight scene against robot opponents with prominently visible weapon typical for FPS games.

resemblance to each other and therefore met the criterion defined in ITU-T Rec. P.911 that the stimuli should show different types of scenes in order not to bore participants.

7.1.3 Test procedure

For the test, subjects were recruited from participants of the preceding web survey following the criteria outlined in Section 7.1.2. The study was conducted using a within-subject design and the test runs were planned to last approximately 90 minutes.

After being welcome by the instructor, the participants read a written introduction, explaining the procedure of the experiment. After this, they had to sign an informed consent and rate their sleepiness using a Karolinska Sleepiness Scale (KSS).

The main part of the experiment consisted of three blocks:

- Passive Test
- Interactive Test: Grand Theft Auto 5 (GTA V)
- Interactive Test: Call of Duty: Black Ops III (CoD)

Although the two interactive parts were always conducted en bloc, their order was balanced for the participants. To prevent order effects also for the passive test, half of the participants started with the passive part, whereas the other half started with the interactive block. Between each block, a 5-minute break was inserted.

The passive test commenced with a series of four stimuli showing both games at the best (100 Mbit/s) and the lowest (3 Mbit/s) bit rate levels. This training phase without rating was followed by the actual assessment session, in which 16 prepared stimuli (two stimuli for each combination of the two games and four bit rates) were shown in random order. Following the Absolute Category Rating (ACR) method defined in ITU-T Rec. P.911 [69], each 10-second stimulus was followed by a short break, during which the participants had to rate the video. Other than the 'up to 10 seconds' guideline in ITU-T Rec. P.911 and in Figure 7.1, the participants were given 15 seconds for their appraisal, since they had to use not just one, but three ACR scales (overall quality, video quality, and audio quality) to rate.

Each of the interactive tests began with an undegraded 6-minute training session, in which the participants were allowed to freely interact with the game and get used to the control and the game play. After this introduction, four test sessions per game were played with different bit rates which each started with reading a written instruction on the respective mission and lasted for two minutes. After these two minutes, the instructor would inform participants that the time had passed, but that they could continue for another minute if they wanted. After finishing playing, the participants filled the questionnaire and proceeded with the next session. Whereas the order of the missions for each game was static, the applied bit rates were randomized.

After the passive and interactive parts of the experiment were finished, the participants rated their sleepiness using the Karolinska Sleepiness Scale (KSS) again.

Altogether 20 subjects (3 females and 17 males; mean age = 21.64 years; SD = 1.089; range = 20-24) participated in the study, of whom nearly all (19) were students. They played and rated a total of 160 interactive sessions and created another 320 data points when they passively viewed and rated the 16 preproduced stimuli.

7.1.4 Results

The ratings on the Karolinska Sleepiness Scale (KSS) were coded as 1 = "*extremely alert*" to 9 = "*Extremely sleepy-fighting sleep*". The continuous rating scales used for the overall quality, video, and audio MOS were mapped to the range from 0 = "*extremely bad*" to 6 = "*ideal*". Ratings on the SAM pictorial scales were coded to the range from 1 to 9. GEQ items were coded with 0 = "*not at all*" to 4 = "*extremely*". From the 14 items of the In-Game Questionnaire, the 7 Player Experience dimension were calculated following [100]. The error bar in all following figures indicates a confidence interval of 95 %.

The averaged ratings for overall quality (MOS_{AV}), video (MOS_V), audio quality (MOS_A) from both the interactive and the passive test setting are shown in Figure 7.4, Figure 7.5, and Figure 7.6.

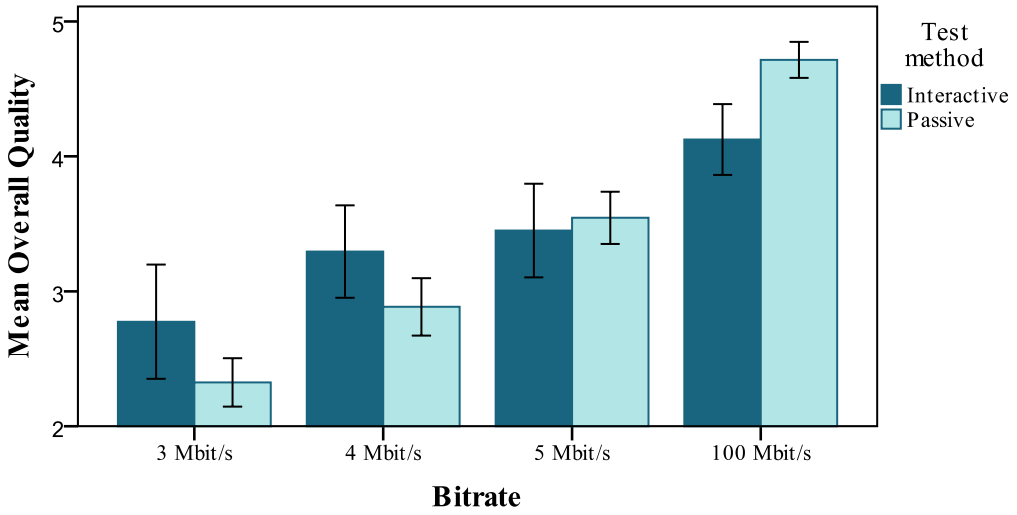


Fig. 7.4 Overall quality MOS ratings for GTA V and CoD scenarios in interactive and passive tests when transmitted at a bitrate of 3 Mbit/s, 4 Mbit/s, 5 Mbit/s, or 100 Mbit/s.

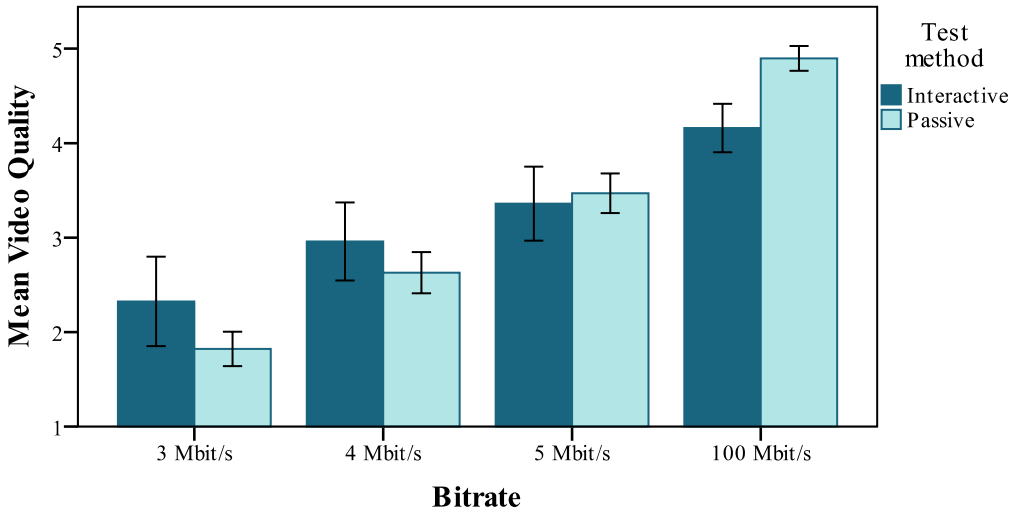


Fig. 7.5 Video quality MOS ratings for GTA V and CoD scenarios in interactive and passive tests when transmitted with a bitrate of 3 Mbit/s, 4 Mbit/s, 5 Mbit/s, or 100 Mbit/s.

The obtained mean ratings and the corresponding standard deviations for the overall quality item are compiled in Table 7.1. The standard deviations for ratings obtained using the passive test are considerably lower than those in the interactive test.

The participants' mean ratings on the Karolinska Sleepiness Scale (KSS) before ($M = 3.7$, $SD = 1.89$) and after the experiment ($M = 3.7$, $SD = 1.53$) did not show a clear effect and were even similar in the mean.

To analyze the obtained data, the distribution of the ratings for each condition was tested for normality using a Shapiro-Wilk test, which was preferred over a Kolmogorov–Smirnov

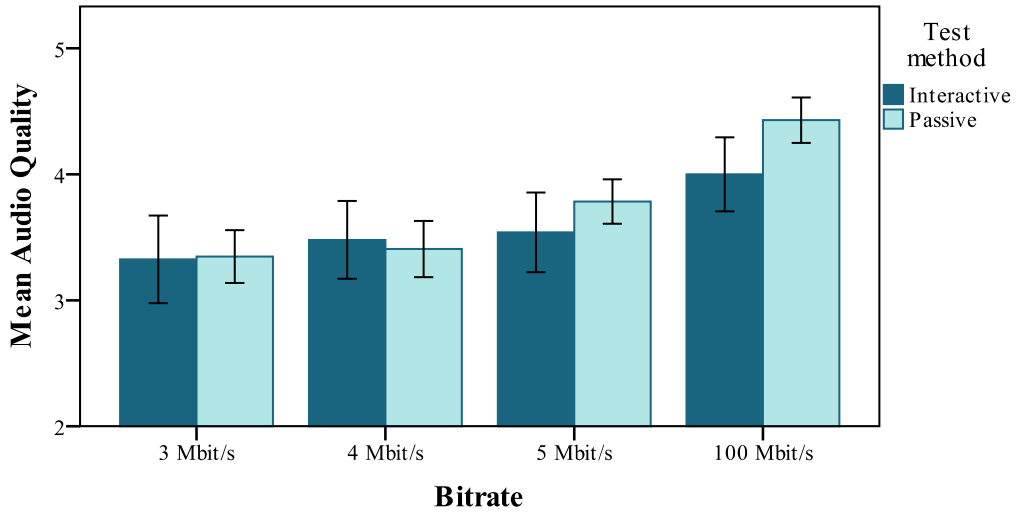


Fig. 7.6 Audio quality MOS ratings for GTA V and CoD scenarios in interactive and passive tests when transmitted with a bitrate of 3 Mbit/s, 4 Mbit/s, 5 Mbit/s, or 100 Mbit/s.

Table 7.1 Mean overall quality ratings (M) and standard deviations (SD) for both tested games for the interactive and passive test paradigms with all tested bit rates.

Bit rate	Test method	CoD		GTA V	
		M	*SD*	*M*	*SD*
3 Mbit/s	interactive	2.90	1.16	2.65	1.48
	passive	2.62	0.64	2.04	0.69
4 Mbit/s	interactive	3.14	1.13	3.18	1.02
	passive	3.10	0.75	2.67	0.72
5 Mbit/s	interactive	3.30	1.03	3.60	1.14
	passive	3.56	0.71	3.53	0.72
100 Mbit/s	interactive	4.23	0.71	4.02	0.93
	passive	4.80	0.50	4.63	0.59

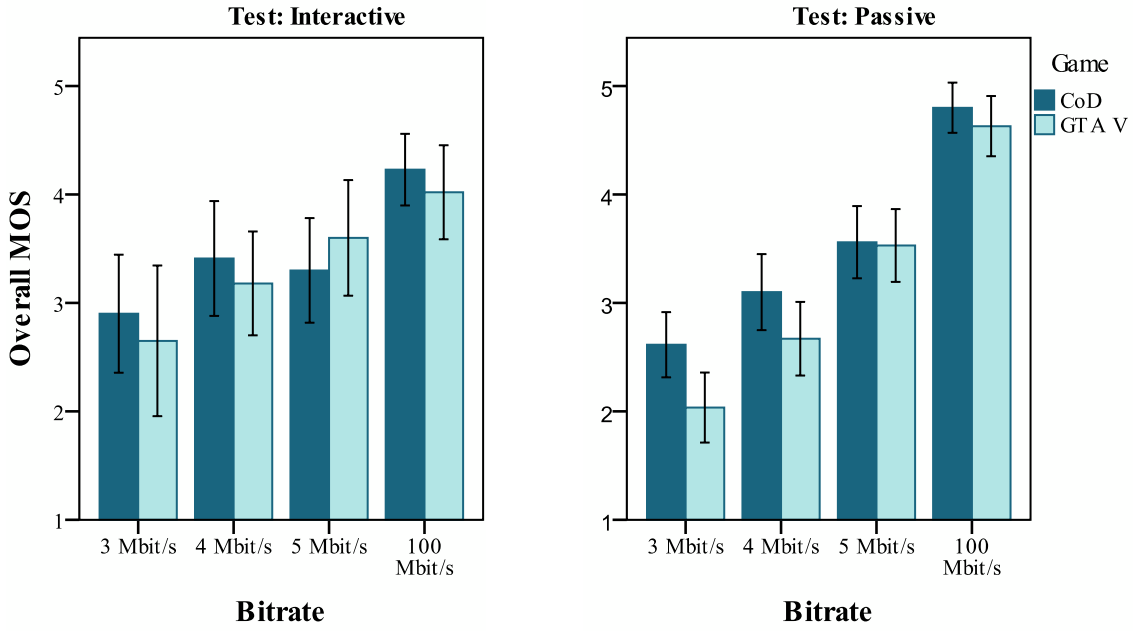
test due to the small sample size. To perform this test, the data was split into groups using the independent variables test method (interactive, passive), and bit rate. As this test revealed significant violations of the normality assumption in many items, non-parametric tests are used in the following.

To check if the applied test method caused the ratings for MOS_{AV}, MOS_A, and MOS_V to be significantly different (hypothesis H_0 is that they are similar), non-parametric Wilcoxon Signed-Rank tests [88] were performed. A significant result in this means that the medians of the compared sets of ratings differ and that this result is unlikely to be coincidental. The tests' results are compiled in Table 7.2.

Table 7.2 Results from non-parametric Wilcoxon Signed-Rank tests testing the median ratings from interactive and passive sessions with the displayed games GTA V and CoD for the overall quality (MOS_{AV}), video quality (MOS_V), and audio quality (MOS_A) items for similarity. A significant result ($p < .05$) means that the null hypothesis of both the passive and the interactive test yielding the same rating has to be discarded ($*$).

Bit rate	GTA V			CoD		
	MOS_{AV}	MOS_V	MOS_A	MOS_{AV}	MOS_V	MOS_A
3 Mbit/s	$p = .285$	$p = .046*$	$p = .775$	$p = .125$	$p = .156$	$p = .886$
4 Mbit/s	$p = .047*$	$p = .213$	$p = .294$	$p = .048*$	$p = .255$	$p = .420$
5 Mbit/s	$p = .984$	$p = .920$	$p = .618$	$p = .868$	$p = .948$	$p = .446$
100 Mbit/s	$p = .006*$	$p = .001*$	$p = .169$	$p = .001*$	$p = .000*$	$p = .008*$

In Figure 7.7 the overall quality ratings (MOS_{AV}) and the four used bit rates are differentiated by game and test method.



(a) Interactive test scenario.

(b) Passive test scenario.

Fig. 7.7 Overall quality (MOS_{AV}) ratings from interactive and passive test for both games (GTA V and CoD) with different bitrate settings (0: “extremely bad” - 6: “ideal”).

As the SAM and GEQ were only rated for the interactive scenario, they can only be examined in the light of the applied streaming bit rate change. A graph with the progression of their ratings is shown in Figure 7.8.

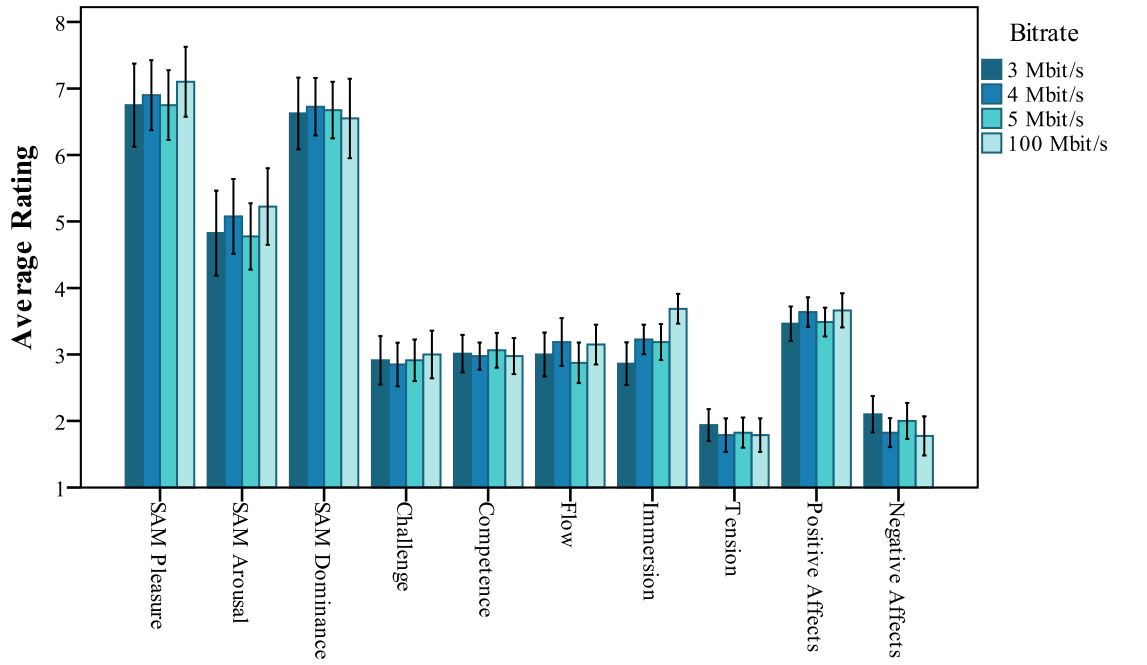


Fig. 7.8 Ratings for the SAM and GEQ dimensions averaged over both games for the four applied streaming bitrates in the interactive test.

7.1.5 Discussion

The analysis of the MOS results from the passive and interactive test shows that there is a great degree of similarity in the rating behavior in terms of improvement with rising bit rate, as can be seen in Figure 7.4 for overall quality, Figure 7.5 for video quality, and Figure 7.6 for audio. The visible rise of perceived audio quality with both test methods despite objectively unchanged parameters in that regard is surprising, but in line with previous research, e. g., [11]. Beerends *et al.* describe the influence of changed visual quality in an audiovisual stimulus on perceived audio quality as 1.2 points on a nine-point quality scale (i. e. 13.3 % of the scale's entire spread). In this study, the difference in MOS_A between the best and the worst bit rate condition was 0.675 (9.6 % of the scale) for the interactive case, and 1.083 (15.5 % of the scale), both on a 7-point scale, so the relative change in perceived audio quality is generally comparable to [11].

However, although the ratings from the passive and the interactive test both mirror the change in transmission bandwidth, the statistical test results in Table 7.2 attest, that in 4 out of 8 conditions (50 %), the overall quality was rated significantly differently, in 3 out of 8 conditions (38 %) video quality differed, and in one condition (13 %), the ratings for audio quality differed significantly.

In total, this means that passive tests cannot be used as simple replacement for interactive tests even if the independent test variable solely varies a visual aspect as in this case.

Generally, there seems to be an attenuating effect of the interactive game-play on ratings: In the passive test, the participants used a much greater range of the scale than in the interactive test, whereas in the latter case the ratings remained closer to the center of the scale. This effect was present for both tested games (cf. Figure 7.7). In retrospect, a flaw in the test design existed, in that participants were presented stimuli resembling the best and worst conditions as training in the passive test, but did only practice with the ideal condition in the interactive cases before starting to rate conditions. However, as the test design was balanced in a way that one half of the participants first performed the passive rating before starting to interactively play, it can be argued that at least the group which experienced the breadth of visual quality levels in the passive test, should be able to use the full scale in the interactive test. Unfortunately, a conclusive answer to that hypothesis is not possible with the obtained data set as the number of 5 persons per group does not allow a sound comparison of the groups, particularly in light of the high standard deviations observed in the interactive test.

Besides the different scale usage, the passive test appeared to be more sensitive to changes of transmission bit rate than the interactive test: Whereas the overall quality ratings for CoD are virtually the same in the 4 Mbit/s and the 5 Mbit/s conditions in the interactive test (cf. Figure 7.7a), they are clearly distinguishable in the passive test (cf. Figure 7.7b). Considering the substantially lower standard deviations of ratings obtained in the passive test (cf. Table 7.1), this method is able to discern quality variations much more sensitively than the interactive test.

Notwithstanding the participants' incomplete training in the interactive case, the ratings are interesting with regard to the effect of low and high bit rate on game-play: The 3 Mbit/s condition led to severe blockiness in the picture in both GTA V and CoD. Although this limited the players' ability to, e. g., identify small objects, which might be important for gaming decisions (e. g., to steer the car in time around an upcoming obstacle), such a handicap did not show up in the ratings in a prominent way. Instead, the games' contents and tasks seem to have driven attention away from the recognition of visual artifacts. This is corroborated by the results of SAM and GEQ in Figure 7.8: The participants seem to have been able to enjoy the games despite their low visual quality in the 3 Mbit/s condition.

The game selection process seems to have resulted in adequate titles for the participants of the study. Although Call of Duty: Black Ops III (CoD) was rated slightly better than Grand Theft Auto 5 (GTA V) in Figure 7.7a, the level is generally good (a MOS of four

is related to the label "Good") and confirmed by the high ratings for Pleasure in the SAM questionnaire seen in Figure 7.8.

Although the means of the KSS ratings from before and after the experiments remained the same, this is not true for the individual participants. While the experimental tasks exhausted some, they were apparently stimulating for others and made them feel more awake.

7.2 Assessing gaming experience with electroencephalography

As introduced in Section 2.7, physiological methods are a promising way to assess the quality of media consumption and particularly of gaming without the interruption inevitably caused by filling questionnaires or answering interview questions. In this section, a study[13] is presented, in which the quality variation caused by the change of one key parameter of a cloud gaming connection, the video streaming bandwidth, was assessed using self-assessment questionnaires and physiological measures using electroencephalography (EEG).

The contents of this section have previously been published in slightly different form in [19].

7.2.1 Methodology

To conduct the study, a cloud gaming test bed using the first-person shooter "Cube 2: Sauerbraten" and the open source platform GamingAnywhere [55] was built. The participants played two levels with two different video bit rates (low and high bit rate condition), of which one led to almost no perceptible visual degradation (high bit rate) whereas the other caused heavy blurring and blockiness (low bit rate).

To derive a feature from the EEG data to compare the different conditions and examine the degree of accordance with the subjective self-assessment, the main focus was on variations of the alpha frequency band power in the EEG signals. This can be used as an indicator of the player's cognitive state, as a higher power in this band corresponds to a reduced cognitive state. The rationale for using this as a feature is that prolonged playing of cloud gaming with very bad visual quality would cause additional cognitive strain and therefore lead to growing exhaustion and a reduced cognitive state. Therefore, the variation of the alpha band power between 9 and 11 Hz, (i. e., the center of the alpha band), due to the two video quality levels is analyzed.

[13] The study was conducted in collaboration with Richard Varbelow as part of a master thesis.

Fig. 7.9 Study setup with player seated at a desk and g.GAMMAcap with wiring in place.

As in all previously discusses laboratory studies, the study environment was set up according to ITU-T Recommendations P.910 [68] and P.911 [69] and was equipped with daylight-imitating lamps, and all walls were covered with thick neutral gray sound-absorbing curtains. Test participants were seated in a non-moving chair in front of a desk upon which the test client computer, a monitor, input devices and two loudspeakers were set up. Equipment of g.tec medical engineering GmbH was used to continuously record the EEG signal. The participants had to put on the g.GAMMAcap[2] containing 16 active ring electrodes located according to the international 10-20 system (Fz, F3-4, FP1-2, Cz, C3-4, Pz, P3-4, PO3-4, Oz, O1-2) [77]. Both the grounding and the reference electrodes were placed at the mastoids (bone structures behind the ear channel filled with air). The signal was amplified and digitized with the g.USBamp and recorded on a dedicated computer (Fujitsu Lifebook S761[14], Intel Core i7 2.7 GHz, 8GB RAM, Windows 7) using the software g.Recorder.

The hardware foundation for the cloud gaming server was provided by a DELL PowerEdge T420[15] server (2x Xeon E5-2430; 12 CPU cores at 2,2 GHz; 64 GB RAM) placed in a server cabinet with connection to the laboratory room through a switched Gigabit Ethernet network. For the study, the server was equipped with an Nvidia Quadro FX4800 graphics card. As in a realistic usage scenario, a virtualization platform was installed on the server, Citrix XenServer v6.2[16]. Within that virtualization a Windows 7 instance, equipped with 4 CPU cores and 4 GB RAM, was created. The physical Nvidia GPU was dedicated to this virtual machine, providing 3D OpenGL rendering capabilities to the game “Cube 2:

[14] http://sp.ts.fujitsu.com/dmsp/Publications/public/ds-LIFEBOOK-S761.pdf (last accessed: 2016-04-27)
[15] http://www.dell.com/us/business/p/poweredge-t420/pd (last accessed: 2016-04-27)
[16] http://xenserver.org (last accessed: 2016-04-27)

Sauerbraten"[17] running on the open-source cloud gaming platform GamingAnywhere[18][19] (v0.7.5) [55]. Being a first-person shooter, this game is particularly fast-paced and strongly depends on the player's ability to quickly discern visual features to recognize enemies and find his/her way through the virtual world. Two streaming configurations were created with the platform. Each transmitted the H.264-compressed video with a 1280x768 resolution at 50 fps and OPUS[20]-compressed audio with a 48 kHz sampling rate. In both cases, the OPUS audio compressor was configured to output 128 kbit/s. However, the video encoding bit rate differed and was set to 10 Mbit/s in the high quality (HQ) case and 1 Mbit/s in the low quality (LQ) case. Since the video compression was performed entirely in software (through FFMPEG[21]/x264[22]), its 'preset' was set to 'ultrafast' and the 'tune' parameter to 'zerolatency' to keep encoding latencies at bay. The provisioned CPU power was sufficient to avoid frame rate degradations due to processing bottlenecks, as the observed overall utilization of the cores stayed around 50 percent. As client, a DELL Latitude D630[23] laptop (Intel Core 2 Duo 2.5 GHz, 2 GB RAM, Windows 7) was used, which was connected to an external 22-inch screen.

Within the game, two levels ("Lost" and "Level9") were chosen based on their game mode being a campaign and the fact that the participants could not finish the level during the sessions. A campaign in "Sauerbraten" is a separately playable level, where the player has to defeat enemy monsters and progress linearly to reach the end. The participants were asked to get as far as possible which included finding buttons or computer terminals to open locked doors. The basic principle stayed the same for both levels, although "Lost" had some advanced capabilities as controlling a rail with a remote control. The overall interactive delay of the cloud gaming setup was observed to be about 110 ms using a high-speed (240 frames per second) camera recording.

7.2.2 Test procedure

Participants were recruited using a web portal for the management and acquisition of test subjects. Each experiment started with an introduction phase where the participants were informed about the test procedure, had to sign the consent form and complete the first questionnaire, collecting demographic data, gaming habits, and the emotional and wakefulness

[17] http://sauerbraten.org (last accessed: 2016-04-27)
[18] http://gaminganywhere.org (last accessed: 2016-04-27)
[19] https://github.com/chunying/gaminganywhere (last accessed: 2016-04-27)
[20] https://www.opus-codec.org (last accessed: 2016-04-27)
[21] https://ffmpeg.org (last accessed: 2016-04-27)
[22] http://www.videolan.org/developers/x264.html (last accessed: 2016-04-27)
[23] http://www.dell.com/us/dfb/p/latitude-d630/pd (last accessed: 2016-04-27)

state. Subsequently, the EEG equipment was set up while the participants played a training level to get familiar with the game. After the preparation of the EEG, a baseline was recorded during which the participants were asked to fixate a spot on the curtain in front of them for two minutes, and then to keep their eyes closed for the same period of time. Two gaming sessions followed, each 20 minutes long. To minimize learning effects as far as possible, instead of repeated sessions with short levels, the participants had to play both levels until they were interrupted when the time was up. The quality levels (HQ, LQ) served as random within-subject factor and the game levels were randomized to prevent order effects. After each session, a comprehensive questionnaire had to be completed gathering data in terms of quality ratings (MOS), game experience (GEQ), and again emotional (SAM) and wakefulness state (KSS). When all questionnaires were completed, the EEG equipment was removed and the test participants were offered an opportunity to wash their hair. Finally, they received financial compensation.

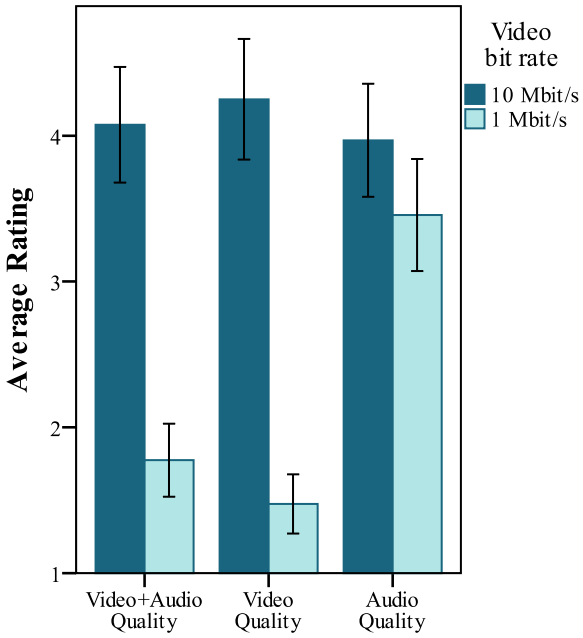
The experiments were conducted from 2015-09-01 to 2015-10-02 in a laboratory room at Technische Universität Berlin. Altogether 32 subjects (5 females and 27 males; mean age = 25.94 years; SD = 2.723; range = 19-31) participated in the study, of whom most (25) were students.

7.2.3 Results

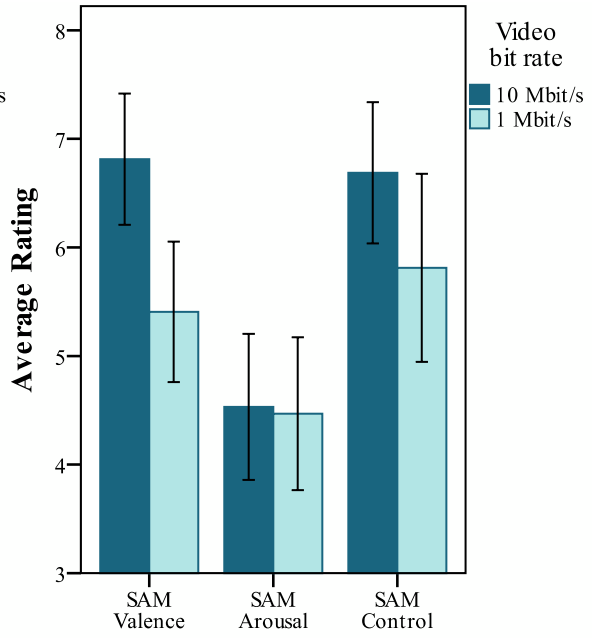
For the analysis multiple ANOVA for repeated measures were calculated. As independent variable the video quality level was used. The subjective scales and the alpha frequency band power served as dependent variables. The error bar in all figures indicates a confidence interval of 95 %.

Subjective results

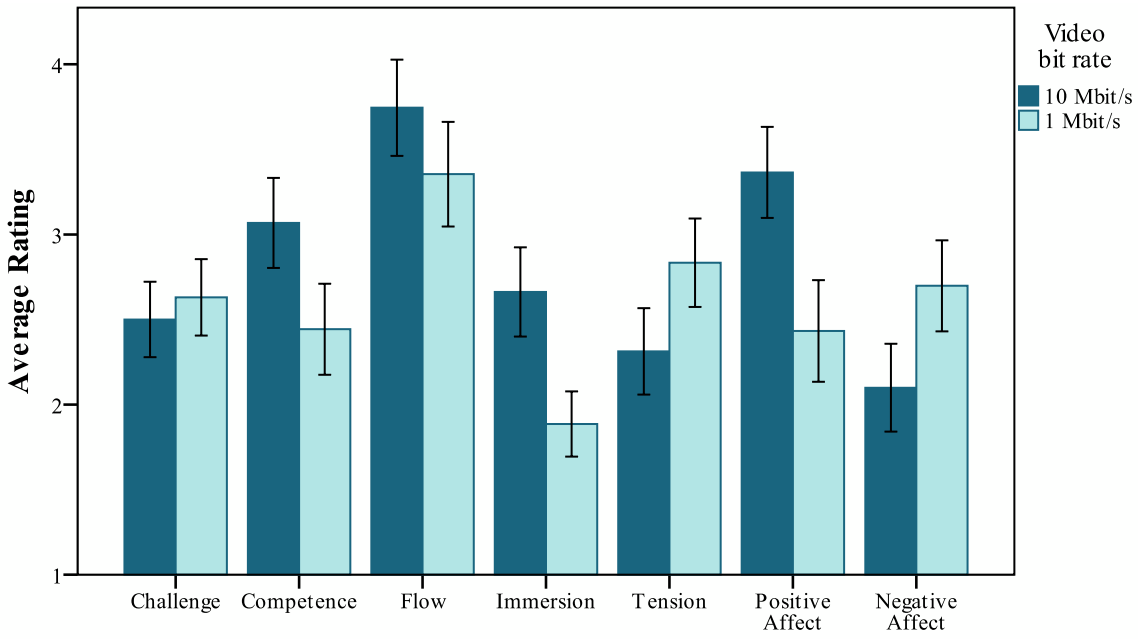
The MOS ratings (collected on a scale from 1 to 7 with a step size of 0.1, where 1 corresponds to “extrem schlecht” / “extremely bad” and 7 to “ideal”) for the video and audio quality show the expected difference in the subjects’ perception (Figure 7.10a). Although the audio quality was not changed, its rating is significantly affected by the video quality ($F(1,31) = 7.926$, $p < .01$, $\eta^2 = .204$) even if not as distinct as the video quality rating itself ($F(1,31) = 210.906$, $p < .01$, $\eta^2 = .872$), respectively the combined quality of audio and video ($F(1,31) = 132.517$, $p < .01$, $\eta^2 = .810$). For the emotional state (collected on scale from 1 to 9 with step size 1), a significant effect in the valence dimension of the self-assessment manikin (SAM) ($F(1,31) = 18.211$, $p < .01$, $\eta^2 = .370$) was found - test participants felt more pleasure when playing the high quality (HQ) condition (Figure 7.10b).



(a) MOS ratings of Audio+Video, Video, and Audio quality.



(b) Ratings on SAM pictorial scales for Valence, Arousal, and Control.



(c) Player Experience dimensions.

Fig. 7.10 Subjective self-assessment ratings for the high quality (10 Mbit/s) and low bitrate (1 Mbit/s) conditions.

There is also a tendency in the control dimension, implying a feeling of being more in

control during the HQ session, albeit this effect is not significant ($F(1,31) = 3.925$, $p < .1$, $\eta^2 = .112$).

The Karolinska Sleepiness Scale (KSS) (collected on a scale from 1 to 9 with step size 0.1, where 1 corresponds to "extremely alert" and 9 to "extremely sleepy – fighting sleep") reveals another significant effect ($F(1,31) = 5.859$, $p < .05$, $\eta^2 = .159$), namely that playing the low quality (LQ) condition leads to a slightly more tired state ($M = 3.96$, $SD = 1.86$) than the HQ session ($M = 3.46$, $SD = 1.50$).

Of the 7 dimensions of the Game Experience Questionnaire (GEQ) (coded on a scale from 1 to 5 with step size 1, where 1 corresponds to "not at all" and 5 to "extremely"), 6 showed significant effects (Figure 7.10c). When playing the HQ session, the subjects felt more competent ($F(1,31) = 14.235$, $p < .01$, $\eta^2 = .315$), were more in a flow state ($F(1,31) = 5.941$, $p < .05$, $\eta^2 = .161$), experienced stronger immersion ($F(1,31) = 25.207$, $p < .01$, $\eta^2 = .448$) in the game, felt less tense ($F(1,31) = 10.722$, $p < .01$, $\eta^2 = .257$), it affected them more positively ($F(1,31) = 24.255$, $p < .01$, $\eta^2 = .439$), and less negatively ($F(1,31) = 15.042$, $p < .01$, $\eta^2 = .327$) than the LQ session. Only the changes to the Challenge dimension were not significant, although there is a slight tendency towards being more challenged when playing at LQ.

Physiological results

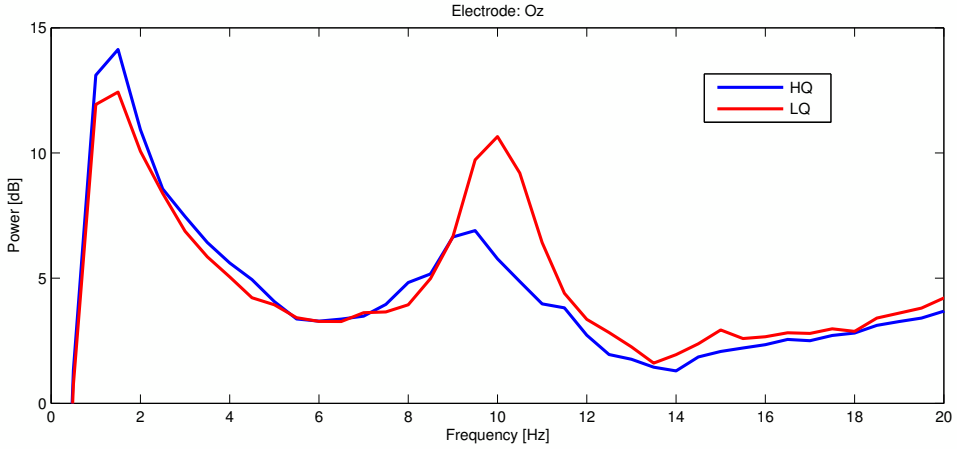


Fig. 7.11 Alpha frequency band power of the first half of the gaming sessions averaged over all participants for the data of electrode Oz and the two presented video quality levels.

In the EEG data, a significant effect for the alpha frequency band power of the electrode Oz ($F(1,27) = 4.34$, $p < .05$, $\eta^2 = .138$) was found for the first half of the sessions (cf. Figure 7.11). As the signals from two participants were overly noisy, and two more experienced

technical issues causing reoccurring recalibrations and jammed signals, four records were discarded. For the remaining participants, the power was calculated for the narrow alpha band in the interval 9-11 Hz. Fortunately, participants excluded from the physiological analysis are evenly distributed over the randomized quality order, so no unilateral influence could result. As can be seen in Figure 7.11, the power spectral density in the alpha frequency band in the range between 9 to 11 Hz is higher for the low video quality condition in comparison to the high video quality condition. All other occipital electrodes showed the same tendency but did not meet significance levels.

7.2.4 Discussion

The results show that the visual quality of the game is significantly reflected in nearly all tested measures. As expected, the MOS ratings for video quality were strongly influenced by the stimuli. However, the observed MOS levels also confirm that the chosen parameter sets were appropriate to create a high and a low quality condition. One surprising feature is the significant influence of video quality variations on audio quality ratings, even though audio quality remained unchanged throughout the study. This is, however, in line with the literature and was also observed previously in Section 7.1.

The SAM revealed a significant effect of the video quality on the valence of the participant's affective state, implying that they felt less pleasure after playing the LQ condition. This finding is consistent with the ratings for the Positive and Negative Affect dimensions in the GEQ. Besides Challenge, all other GEQ dimensions were significantly affected: Lower video quality caused less positive emotions (Positive Affect) and raised negative emotions (Negative Affect). It was less immersive and left players feeling less competent. However, the bad quality also heightened the tension and might also have caused the game to be more challenging although the latter effect was not significant. Considering the very bad quality the players had to endure in the LQ condition, the observed differences in the Player Experience dimensions are lower than expected. Apparently, even a very low level of visual quality does not completely break the underlying game principle, in that it is still tense and challenging and players could enter a state of flow.

The subjective data further showed a significant effect for the wakefulness state: The study participants felt more tired after the LQ session than after the HQ session.

This effect of tiredness was also observable in the physiological EEG data: Playing the LQ condition caused significantly higher spectral power in the alpha frequency band during the first half of that session compared to the HQ condition. While this effect was also observable in the second half of the sessions, it was less pronounced and did not reach significance level. This might imply that the longer a player played the game, the less

influence is exerted on the wakefulness state by the video quality. As a game is an interactive endeavor as opposed to mere passive video consumption, the player may over time adapt to the degraded visual quality, and the game's interactive content might dominate the perception.

7.3 Conclusions

In this chapter two test methodologies were investigated. The first part addressed a comparison of passive and interactive test paradigms. While passive (i. e., viewing and/or listening) tests are established in the assessment of audio, video, and audiovisual stimuli, gaming tests virtually always incorporate interactive playing. In the presented study, passive test methods were used to rate pre-produced recordings of gameplay which were compared to ratings from interactive game sessions using the same games. The comparison showed, that both methods were sensitive towards the applied changes in video transmission bit rate. However, ratings from the passive tests differed significantly from the interactive sessions in that they used a greater range of the available scale, and the values showed a lower standard deviation when compared to ratings obtained in the interactive test. This means, that passive tests may not be used as a replacement for interactive tests. However, they may be applicable as an extension to assess just visual aspects, and they would have the benefit of being both more efficient and sensitive in that scenario. The exact relation between assessments obtained this way and the overall quality opinion of the interactive gaming is, however, yet unexplored.

In the second part, electroencephalography was investigated with the goal of finding a physiological correlate of gaming experience in a cloud gaming setup with strongly varying streaming quality. It was found that the video quality influenced the overall quality MOS_{AV}, video quality MOS_V, audio quality MOS_A, GEQ player experience, the SAM valence rating, and the EEG alpha frequency band power in the first halves of the sessions. The observed rise in alpha frequency band power is likely related to a reduced mental state (i. e. tiredness) caused by prolonged playing under adverse streaming conditions, which is in line with previous works on the alpha-band effects of long-term exposure to strongly degraded audio material [5]. As such, physiological measures continue to be an interesting research field as they could one day reduce the dependency on subjective self-assessment in quality evaluations.

Comparing the bit rate variations' effects on the GEQ dimensions' ratings in Figure 7.8 and Figure 7.10c, the observed differences between the highest and the lowest bit rates differ strongly: Whereas almost no effect of the different video compression levels was noted in the first study except for the Immersion dimension, multiple dimensions show clear effects in the second study. However, in the latter, the degree of visual degradations was much

stronger in the lowest quality condition than in the first study, as can be seen by comparing the substantial drop in video quality ratings in Figure 7.10a to the smaller decrease seen in the interactive test in Figure 7.5. Consequently, the GEQ has to be considered quite insensitive to visual degradations as even the extreme quality level variation used in the latter study only caused modest effects to player experience according to its dimensions.

Chapter 8

Conclusion and future work

The subjective experience of gaming is the result of numerous factors. While the game itself sets the stage, it is influenced and limited in that by a great variety of further factors. Which of the many conceivable factors do actually meaningfully influence that experience is, however, largely unknown. This thesis attempts to fill that gap by selecting four major factors and examining and testing them with regard to the measurable influence they may exert.

8.1 Summary

After an introduction to existing definitions, measures, and measurement tools for gaming quality in Chapter 2, the subjectively perceivable effects of a set of influence factors on gaming quality were studied and discussed:

In Chapter 3, the differences between three mobile multi-player games were investigated. It was found, that the games were not only rated differently due to their differing contents and game tasks, but also because of the specific implementations, which reacted dissimilarly to the simulated network conditions. While many of these implementation-specific details are not problematic in a cloud gaming setup, because only the audiovisual output of the games is sent over the unpredictable Internet in that case and the games themself always run in a comparable environment, these implementation-specifics are very much a concern in non-cloud-gaming-based experiments and use cases: Since the freedom of game developers in the way they handle changing network behavior, different screen sizes, game interruptions, varying skill of players, etc. is almost unlimited, it is very unlikely that an accurate, yet generic mathematical quality model for all mobile games encompassing all relevant influence factors can ever be built. While this rules out a theoretical perfect model, approximations with a limited scope may well be possible. In the described experiment, network parameters were changed in a very wide range. While this posed to be a concern for two of the three games,

one game's quality ratings remained essentially immune to delay. When a narrower range of latencies was investigated, the differences between games and their implementations may likely not have been so extreme. As newer network technologies become more robust and cellular networks more reliable, maybe the intensity of network-induced quality-variations may decrease to such a degree, that simple approximations are possible and sufficient. The same may be true for other technology-induced variations: As the category of mobile games grows more mature, techniques will likely spread which handle interruptions gracefully, adapt to different displays intelligently, and adjust difficulty to the player wisely.

The study presented in Chapter 4 examined the effect of the device, and particularly its display size on gaming experience. It was found that an effect exists and that bigger screens are generally rated better. However, the only meaningful observed difference was, that a display can apparently be too small for enjoyable gaming, but that above a threshold somewhere between 3.27" and 5", the ratings leveled. A trend was observed for decreasing quality ratings on a very large tablet device (10.1"), yet this was not significant. Judging from results obtained in the study, a display size of around 7" was ideal for the games used in the test. The consequence of display size being an influence factor for gaming quality is that, in future studies, it has to be controlled.

A mobile cloud gaming setup was used to assess the influence of network variations on game playing in the study presented in Chapter 5. It could be shown, that the gaming experience of streamed mobile touch-based games is similarly sensitive to decreases of transmission bit rate as PC and console-based cloud gaming. However, in contrast to these more traditional non-mobile cloud gaming platforms, almost no delay influence was registered with the tested smartphone games. As the simulated network parameters lie well within the capabilities of current cellular networks, mobile cloud gaming is shown to be a suitable game delivery method from a technical perspective. Although the games were differently affected by lowered bit rate levels, these effects were much more homogeneous than those observed in the study with locally executed games in Chapter 3, making it likely feasible to model the effects. To support such an effort and facilitate collaboration, the ITU-T Study Group 12 Q13 has created a work item to create an opinion model with the name G.OMG.[1]

In Chapter 6 the influence of the context of playing was investigated. Participants played and rated the same games while riding a metro and while sitting in a quiet laboratory room at a desk. For both settings, the observed ratings were largely similar. Apparently, the perception of the game and its tasks dominates that of the environment. Although this finding

[1] http://www.itu.int/ITU-T/workprog/wp_item.aspx?isn=9999 (last accessed: 2016-06-21)

may come at a surprise, it is highly beneficial for the external validity of laboratory studies with games.

Finally, in Chapter 7, two different test methodologies were investigated: In Section 7.1, a test paradigm standardized for video quality assessment was used to rate sequences of recorded game-play. As a comparison, the very same games seen in the videos were played with equal visual degradations and rated by study participants. The obtained data showed, that both methods were individually adequate to assess the games, but when compared to each other, their results varied. Quality ratings for the interactively played sessions were more concentrated around the scale's center, whereas the video assessment was more sensitive and yielded less noisy ratings which were, however, spread more on the scale. As a consequence, quality ratings obtained in a passive video rating test do not resemble the quality perceived in interactive sessions. They may, nevertheless, be valuable to more efficiently and sensitively assess just the audiovisual output of a system without the intent to obtain ecologically valid data.

The other examined test paradigm used the physiological method EEG to monitor participants' scalp voltage differences while playing. Physiological methods may one day allow assessing gaming quality continuously in real time without the need to interrupt the player to obtain self-assessment ratings. However, the obtained results in the study were mixed. While the strongly different test conditions caused noticeable spectral variations in the EEG signals, the difference was only significant during the first half of the played sessions. However, further methods exist to analyze EEG signals and extract information. It is therefore possible, that other methods prove to yield more readily-usable metrics, which may then be used to infer perceived quality. Yet, another possible interpretation of the study results is a reducing perception of the visual degradations over time as the players get accustomed to the bad image quality.

Taken together, the obtained results from the studies demonstrate a surprising sensitivity of players towards changes of the system used for gaming. In all but one study, significant and meaningful differences in ratings were observed. This confirms, that gaming experience is in fact not merely the result of the game itself, but it is also greatly influenced by parameters outside the reach of game developers and publishers alike. Here, service and network providers have to understand the consequences of their design and operational decisions and may use these for both competitive and the consumers' benefit.

8.2 Limitations

A number of limitations arise from the studies and the way they were conducted. These are considered in the following.

First, the participants in the experiments were predominantly students of Technische Universität Berlin. The studied samples therefore may not have represented the whole population of mobile players appropriately as a significant number of elderly people also enjoy games [4]. Furthermore, the group of casual gamers was prevalent. This was done on purpose, as persons without any gaming experience are unlikely to have the competence to realistically judge the influence of parameter variations and would not play the games out of intrinsic motivation in the first place, and experts may be overly critical even concerning slight changes, which remain almost unnoticed for the majority of non-expert people.

Second, the games used in the test may not have been equally attractive and challenging to all test participants. This problem was partially remedied in some studies by running a pre-test survey and selecting games based on the majority's preferences. Generally, in all studies popular games were selected if possible, which is expected to increase the likeliness of having participants play games with a sufficient degree of intrinsic motivation.

Third, the session lengths of interactive playing may have been inappropriate to allow for, e. g., flow conditions to develop. This is a general problem of controlled gaming research and no good remedy has so far been found.

Fourth, the devices used in the test differed and did not have calibrated screens, which would be expected for proper visual quality assessments following ITU-T Recommendations P.910 and P.911. The screens may therefore have reproduced the games' output incorrectly and could have distorted the results. This concerns particularly the study in Chapter 4 where multiple different devices were compared.

Finally, most participants received a financial compensation for their effort in the studies. This leads to the situation that, strictly speaking, they did not play the games voluntarily and consequently without intrinsic motivation, but completed the tasks to earn the money.

8.3 Future work

In the research leading up to this thesis, a number of questions arose which shall be discussed in this section.

8.3.1 Standardized test methodology

In contrast to video, audio, and audiovisual quality assessments performed in accordance with ITU-T Recommendations P.910 and P.911, gaming quality tests currently lack comparability. This issue slows the scientific process as data points obtained in one laboratory as part of one study can rarely be put into relation to results from another laboratory due to a different methodology. As part of ITU-T Study Group 12, a standardization effort was begun to develop and recommend a common testing paradigm under the working title P.GAME.

Game selection

As discussed in Chapter 3, differences in games' contents and implementations makes it difficult to compare them. Since no established set of references games exists, study results can currently not be put in relation to each other. The selection and standardization of such a set of well-balanced games has the potential to significantly improve that situation.

Session duration

Game sessions in this thesis lasted between one and 20 minutes. Contrary to the strict recommendations for visual tests seen, e. g., in ITU-T Recommendation P.910 (approx. 10 seconds per stimulus), such a precise and strict rule is not useful for gaming as, depending on the games' contents and test equipment requirements (e. g., EEG), different session lengths are required. However, a scientifically-founded recommendation of a range, and particularly a minimum duration, would nevertheless be beneficial. Currently, it is not certain whether two minutes of gaming are sufficient to allow players to get genuinely immersed in the game and experience flow.

Technical setup

As shown in Chapter 4, the devices used for playing games influence a player's gaming experience. With smartphones and tablets being sold in highly different hardware quality classes, a series of minimum requirements could help comparability of results. While a clear recommendation of specific products is not favorable due to the industry's rapid update cycle, minimum requirements could ascertain that, e. g., too-low pixel-density, insufficient display color gamut, or too-high device input delay would not skew obtained results.

Test methodology

While currently virtually all gaming-related studies involve time-consuming interactive gameplaying, this may be dispensable for specific quality aspects with passive tests if the ratings obtained this way could successfully be brought into relation with interactive game ratings. It might furthermore be interesting to investigate the feasibility of adapted Degradation Category Rating (DCR) and Pair Comparison methods (cf. Section 7.1.1). These methods could be adjusted such as that the player has two screens while he plays and one display shows, e. g., the undegraded output, whereas the other is processed by the system under test (DCR). A similar setup is conceivable for Pair Comparison where participants could relate different configurations of the system to each other.

8.3.2 Effects of enhancements to cloud gaming technology

Cloud gaming is a rapidly developing technology. However, one of the great limitations of the concept is its physically limited minimum system response delay: A data center located in the United States cannot serve European customers in a way that allows the round-trip delay to be lower than 40ms[2] because causal influences and signals cannot travel faster than the speed of light [23]. However, concealment techniques on the client side are conceivable. Assuming that a cloud gaming system's video stream transported a wider angle of view than was being shown to the player in a First-Person Shooter game, then the client could react to player input requiring changes of the perspective by varying the displayed window from the more wide-angled streamed view. In the most extreme case, a full 360° view could be transmitted, from which the player would freely choose his desired perspective without any requirement for the server to cooperate. Comparable methods could be devised for many aspects of gameplay as long as the player input does not cause changes to the game state. Such improvements would likely effect the subjective experience of a streamed game significantly and cause prediction models of conventional cloud gaming to become imprecise.

Efficient video compression is the core technology supporting the cloud gaming paradigm. With improvements to it (e. g., through the upcoming HEVC [115]), the subjective experience considering a constant bandwidth is likely to improve. Similarly, enhanced methods of error correction, error concealment, and particularly the application of forward error correction (FEC) would cause visual effects traditionally linked to packet loss to completely disappear. This would have strong implications for a model of cloud gaming quality which would either have to be repeatedly adapted to ever improving coding techniques, or have to be designed

[2] http://royal.pingdom.com/2007/06/01/theoretical-vs-real-world-speed-limit-of-ping/ (last accessed: 2016-05-19)

so general, that these improvements would just require adjusting coefficients. In contrast to the long-term stability of models used in predicting voice call quality, these changes would likely come in short intervals due to the rapid progression in the involved fields.

8.3.3 Setup complexity

With the complexity of gaming testbeds rising rapidly (particularly so in cloud gaming), it may be more efficient to emulate effects of the system rather than to fully implement it. Visual distortions comparable to video compression artifacts could, e. g., be created in real time through an FPGA (cf. [106]) placed between computer and screen, and input delay could be created, e. g., by delaying the forwarding of commands on the USB level between input device and computer or console (such a prototype has been developed successfully[3]). Compared to the full implementation of cloud gaming testbeds, these cheap approaches may save significant time and effort, and might furthermore be usable in other research areas like quality assessments of video telephony. However, they are, unfortunately, not easily applicable in the research of *mobile* gaming, as input device, computer, and display form a unit without the possibility of easily interfering with signal-forwarding in-between the components.

8.3.4 Quality of gaming

Finally, several subjective measures for gaming quality have been presented in Chapter 2, but their relationship remains unknown. How is the overall quality perception (and hence the MOS) derived from these dimensions and how is acceptance formed? Furthermore, the question remains how all of these vary over time, as the study employing EEG in Section 7.2 may be interpreted in the way that players adapted to the bad visual quality in the course of a 20-minute session.

[3] https://github.com/justusbeyer/USBLatencyInjector

References

[1] E. Aarseth, S. M. Smedstad, and L. Sunnanå, "A multi-dimensional typology of games," in *Proceedings of the 2003 DiGRA International Conference: Level Up*, Utrecht, The Netherlands: Universiteit Utrecht, 2003, pp. 48–53.

[2] H. Ahmadi, S. Zad Tootaghaj, M. R. Hashemi, and S. Shirmohammadi, "A game attention model for efficient bit rate allocation in cloud gaming," *Multimedia Systems*, vol. 20, no. 5, pp. 485–501, 2014.

[3] T. Akerstedt and M. Gillberg, "Subjective and objective sleepiness in the active individual," *The International Journal of Neuroscience*, vol. 52, no. 1-2, pp. 29–37, 1990.

[4] J. C. Allaire, A. C. McLaughlin, A. Trujillo, L. A. Whitlock, L. LaPorte, and M. Gandy, "Successful aging through digital games: Socioemotional differences between older adult gamers and Non-gamers," *Computers in Human Behavior*, vol. 29, no. 4, pp. 1302–1306, 2013.

[5] J.-N. Antons, "Neural Correlates of Quality Perception for Complex Speech Signals," PhD thesis, Technische Universität Berlin, 2015.

[6] J.-N. Antons, J. O'Sullivan, S. Arndt, P. Gellert, J. Nordheim, A. Kuhlmey, and S. Möller, "PflegeTab: Enhancing Quality of Life Using a Psychosocial Internet-based Intervention for Residential Dementia Care," International Society for Research on Internet Interventions (ISRII), Seattle, USA, Tech. Rep., 2016.

[7] G. Armitage and G. Annitage, "An Experimental Estimation of Latency Sensitivity In Multiplayer Quake 3," in *The 11th IEEE International Conference on Networks (ICON2003)*, 2003, pp. 137–141.

[8] G. Armitage, *Lag over 150 milliseconds is unacceptable*, 2001. [Online]. Available: http://gja.space4me.com/things/quake3-latency-051701.html.

[9] S. Arndt, J.-N. Antons, R. Schleicher, S. Möller, and G. Curio, "Using Electroencephalography to Measure Perceived Video Quality," *IEEE Journal on Selected Topics in Signal Processing*, vol. 8, no. 3, pp. 366–376, 2014.

[10] R. Bartle, "Hearts, clubs, diamonds, spades: Players Who Suit MUDs," *Journal of MUD research*, vol. 1, no. 1, p. 19, 1996.

[11] J. G. Beerends and F. E. De Caluwe, "The Influence of Video Quality on Perceived Audio Quality and vice versa," *Journal of the Audio Engineering Society*, vol. 47, no. 5, pp. 355–362, 1999.

[12] T. Beigbeder, R. Coughlan, C. Lusher, J. Plunkett, E. Agu, and M. Claypool, "The Effects of Loss and Latency on User Performance in Unreal Tournament 2003," in *Proceedings of 3rd ACM SIGCOMM workshop on Network and System Support for Games*, ACM, 2004, pp. 144–151.

[13] H. Berger, "ÜBER DAS ELEKTRENKEPHALOGRAMM DES MENSCHEN," *European archives of psychiatry and clinical neuroscience*, vol. 87, no. 1, pp. 527–570, 1929.

[14] S. Bertman, *Handbook to life in ancient Mesopotamia*. Oxford University Press, 2005.

[15] J. Beyer, V. Miruchna, and S. Möller, "Assessing the impact of Display Size, Game Type, and Usage Context on Mobile Gaming QoE," in *6th International Workshop on Quality of Multimedia Experience (QoMEX 2014)*, Singapore: IEEE, 2014, pp. 69–70.

[16] J. Beyer and S. Möller, "Gaming," in *Quality of Experience: Advanced Concepts, Applications and Methods*, S. Möller and A. Raake, Eds., Springer Berlin Heidelberg, 2014, pp. 367–381.

[17] J. Beyer and S. Möller, "Assessing the Impact of Game Type, Display Size and Network Delay on Mobile Gaming QoE," *PIK - Praxis der Informationsverarbeitung und Kommunikation*, vol. 37, no. 4, pp. 287–295, 2014.

[18] J. Beyer and R. Varbelow, "Stream-A-Game: An open-source mobile Cloud Gaming platform," in *International Workshop on Network and Systems Support for Games (NetGames)*, Zagreb, Croatia, 2015, pp. 1–3.

[19] J. Beyer, R. Varbelow, J.-N. Antons, and S. Möller, "Using Electroencephalography and Subjective Self-Assessment to Measure the Influence of Quality Variations in Cloud Gaming," in *7th International Workshop on Quality of Multimedia Experience (QoMEX)*, Costa Navarino, Greece: IEEE, 2015, pp. 26–29.

[20] J. Beyer, R. Varbelow, J.-N. Antons, and S. Zander, "A Method For Feedback Delay Measurement Using a Low-cost Arduino Microcontroller," in *Proc. 7th Int. Workshop on Quality of Multimedia Experience (QoMEx 2015)*, Costa Navarino, Greece: IEEE, 2015, pp. 1–2.

[21] J. Blow, "Game Development: Harder Than You Think," *Queue - Game Development*, vol. 1, no. 10, pp. 29–37, 2004.

[22] M. Bodden and U. Jekosch, "Entwicklung und Durchführung von Tests mit Versuchspersonen zur Verifizierung von Modellen zur Berechnung der Sprachübertragungsqualität," Ruhr-Universität, Bochum, Tech. Rep., 1996.

[23] M. Born, *Die Relativitätstheorie Einsteins*, 7. Ausgabe. Springer Berlin Heidelberg, 2003.

[24] M. Bradley and P. J. Lang, "Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. I, pp. 49–59, 1994.

[25] M. Bredel and M. Fidler, "A Measurement Study regarding Quality of Service and its Impact on Multiplayer Online Games," in *Proceedings of the 9th Annual Workshop on Network and Systems Support for Games*, Taipei, Taiwan: IEEE, 2010, pp. 1–6.

[26] E. Brown and P. Cairns, "A Grounded Investigation of Game Immersion," in *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, ACM, 2004, pp. 1297–1300.

[27] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Boredom, Engagement and Anxiety as Indicators for Adaptation to Difficulty in Games," in *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era - MindTrek '08*, Tampere, Finland: ACM, 2008, pp. 13–17.

[28] J. Chen, "Flow in Games (and everything else)," *Communications of the ACM*, vol. 50, no. 4, pp. 31–34, 2007.

[29] K.-T. Chen, Y.-C. Chang, P.-H. Tseng, C.-Y. Huang, and C.-L. Lei, "Measuring the latency of cloud gaming systems," in *Proceedings of the 19th ACM international conference on Multimedia - MM '11*, New York, New York, USA: ACM, 2011, pp. 1269–1273.

[30] K.-T. Chen, P. Huang, and C. L. Lei, "Effect of network quality on player departure behavior in online games," *IEEE Transactions on Parallel and Distributed Systems*, vol. 20, no. 5, pp. 593–606, May 2009.

[31] K.-T. Chen, P. Huang, and C. Lei, "How Sensitive are Online Gamers to Network Quality?" *Communications of the ACM*, vol. 49, no. 11, pp. 34–38, 2006.

[32] K.-T. Chen and C.-L. Lei, "Are all games equally cloud-gaming-friendly? An electromyographic approach," in *11th Annual Workshop on Network and Systems Support for Games (NetGames)*, Venice, Italy: IEEE, Nov. 2012, pp. 1–6.

[33] M. Claypool, "The effect of latency on user performance in Real-Time Strategy games," *Computer Networks*, vol. 49, no. 1, pp. 52–70, 2005.

[34] M. Claypool, "Motion and scene complexity for streaming video games," in *Proceedings of the 4th International Conference on Foundations of Digital Games - FDG '09*, Port Canaveral, Florida, USA: ACM, 2009, p. 34.

[35] M. Claypool and D. Finkel, "The effects of latency on player performance in cloud-based games," in *13th Annual Workshop on Network and Systems Support for Games*, Nagoya, Japan: IEEE, 2014, pp. 1–6.

[36] M. Csikszentmihalyi, *Beyond Boredom and Anxiety: The Experience of Play in Work and Games.* Jossey-Bass Publishers, 1975.

[37] S. Dahlskog and A. Kamstrup, "Mapping the game landscape: Locating genres using functional classification," in *Proceedings of the 2009 DiGRA International Conference: Breaking New Ground: Innovation in Games, Play, Practice and Theory*, London, UK: DiGRA, 2009.

[38] DIN 55350-11, *Begriffe zu Qualitätsmanagement und Statistik - Teil 11*. Berlin, Germany: Beuth Verlag, 2005.

[39] H. Dixon, V. Mitchell, and S. Harker, "Mobile phone games: Understanding the user experience," in *Proceedings of 3rd International Conference on Design and Emotion*, Loughborough, UK, 2002, pp. 1–6.

[40] C. C. Duncan, R. J. Barry, J. F. Connolly, C. Fischer, P. T. Michie, R. Näätänen, J. Polich, I. Reinvang, and C. Van Petten, "Event-related potentials in clinical research: Guidelines for eliciting, recording, and quantifying mismatch negativity, P300, and N400," *Clinical Neurophysiology*, vol. 120, no. 11, pp. 1883–1908, 2009.

[41] C. Elverdam and E. Aarseth, "Game Classification and Game Design: Construction Through Critical Analysis," *Games and Culture*, vol. 2, pp. 3–22, 2007.

[42] S. Engl, "mobile gaming – Eine empirische Studie zum Spielverhalten und Nutzungserlebnis in mobilen Kontexten," Magisterarbeit, Universität Regensburg, 2010.

[43] T. Fullerton, *Game Design Workshop: A Playcentric Approach to Creating Innovative Games*. Elsevier, 2008, pp. 1–470.

[44] B. J. Gajadhar, Y. De Kort, and W. A. Ijsselsteijn, "Shared Fun Is Doubled Fun: Player Enjoyment as a Function of Social Setting," in *Proceedings of the Second International Conference on Fun and Games*, Eindhoven, Netherlands: Springer Berlin Heidelberg, 2008, pp. 106–117.

[45] M. García-Valls, T. Cucinotta, and C. Lu, "Challenges in Real-Time Virtualization and Predictable Cloud Computing," *Journal of Systems Architecture*, vol. 60, no. 9, pp. 726–740, Aug. 2014.

[46] K. Goldhammer, A. Wiegand, D. Becker, and M. Schmid, "Goldmedia Mobile Life Report 2012," Goldmedia GmbH, Berlin, Tech. Rep., 2008. [Online]. Available: https://www.bitkom.org/Publikationen/2009/Studie/Mobile-Life-2012/081009-BITKOM-Goldmedia-Mobile-Life-20121.pdf.

[47] A. Gurtov and J. Korhonen, "Measurement and Analysis of TCP-Friendly Rate Control for Vertical Handovers," *ACM Mobile Computing and Communications Review*, vol. 8, no. 3, pp. 73–87, 2004.

[48] S. Hemminger, "Network Emulation with NetEm," in *Proceedings of the 6th Australian National Linux Conference (LCA 2005)*, Canberra, Australia, 2005, pp. 1–9.

[49] T. Henderson, "Latency and User Behaviour on a Multiplayer Game Server," in *Networked Group Communication*, Springer Berlin Heidelberg, 2001, pp. 1–13.

[50] H.-J. Hong, D.-Y. Chen, C.-Y. Huang, and K.-T. Chen, "Placing Virtual Machines to Optimize Cloud Gaming Experience," *IEEE Transactions on Cloud Computing*, vol. 3, no. 1, pp. 42–53, 2015.

[51] T. Hoßfeld, F. Metzger, and M. Jarschel, "QoE for Cloud Gaming," *Multimedia Communications Technical Committee IEEE Communications Society E-Letter*, vol. 10, no. 6, pp. 26–29, 2015.

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

[53] T. Hoßfeld and T. Zinner, "QoE Management for Cloud Applications with Software Defined Networking," in *NMI 2014 - Virtuell und doch zuverlässig: Cloud für sichere Anwendungen*, Berlin, Germany, 2014.

[54] J. Hou, Y. Nam, W. Peng, and K. M. Lee, "Effects of screen size, viewing angle, and players' immersion tendencies on game experience," *Computers in Human Behavior*, vol. 28, no. 2, pp. 617–623, 2012.

[55] C.-Y. Huang, C.-H. Hsu, Y.-C. Chang, and K.-T. Chen, "GamingAnywhere: An Open Cloud Gaming System," in *Proceedings of the 4th ACM Multimedia Systems Conference (MMSys)*, Oslo, Norway: ACM, 2013, pp. 36–47.

[56] Y. Ida, Y. Ishibashi, N. Fukushima, and S. Sugawara, "QoE assessment of interactivity and fairness in First Person Shooting with group synchronization control," in *Proceedings of the 9th Annual Workshop on Network and Systems Support for Games*, Taipei, Taiwan: IEEE, 2010, pp. 1–2.

[57] IEEE Standard 802.11n-2009, *Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, Amendment 5: Enhancements for Higher Throughput*. IEEE, 2009, pp. 1–536.

[58] IEEE Standard 802.15.1-2005, *Part 15.1: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specification*. IEEE, 2005, pp. 1–721.

[59] W. IJsselsteijn, Y. De Kort, K. Poels, A. Jurgelionis, and F. Bellotti, "Characterising and Measuring User Experiences in Digital Games," in *International Conference on Advances in Computer Entertainment Technology*, vol. 2, 2007, p. 27.

[60] ISO 9000:2000, *Quality Management Systems: Fundamentals and Vocabulary*. International Organization for Standardization, 2000.

[61] ITU-T Contribution COM12-166, "QoE and perceptive quality of video game in passive mode," ITU-T Study Group 12, Geneva, Switzerland, Tech. Rep. Source: Orange Labs, 2016, pp. 1–7.

[62] ITU-T Contribution COM12-390, "Comparison of interactive and passive test methodologies to measure Gaming Quality of Experience (QoE)," ITU-T Study Group 12, Geneva, Switzerland, Tech. Rep., 2016, pp. 1–12.

[63] ITU-T Recommendation E. 800, *Definition of terms related to quality of service*. Geneva, Switzerland: Internation Telecomunication Union, 2008, pp. 1–30.

[64] ITU-T Recommendation E.800, *Terms and definitions related to quality of service and network performance including dependability*. Geneva, Switzerland: International Telecommunication Union, 1994, pp. 1–57.

[65] ITU-T Recommendation G.107, *The E-model: a computational model for use in transmission planning*. Geneva, Switzerland: International Telecommunication Union, 2014, pp. 1–25.

[66] ITU-T Recommendation P.800, *Methods for subjective determination of transmission quality*. Geneva, Switzerland: International Telecommunication Union, 1996, pp. 1–37.

[67] ITU-T Recommendation P.851, *Subjective quality evaluation of telephone services based on spoken dialogue systems*. Geneva, Switzerland: Internation Telecomunication Union, 2003, pp. 1–38.

[68] ITU-T Recommendation P.910, *Subjective video quality assessment methods for multimedia applications*. Geneva, Switzerland: Internation Telecomunication Union, 2009, pp. 1–42.

[69] ITU-T Recommendation P.911, *Subjective audiovisual quality assessment methods for multimedia applications*. Geneva, Switzerland: Internation Telecomunication Union, 1998, pp. 1–46.

[70] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hoßfeld, "An Evaluation of QoE in Cloud Gaming Based on Subjective Tests," in *5th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*, Seoul, Korea: IEEE, 2011, pp. 330–335.

[71] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hoßfeld, "Gaming in the clouds: QoE and the users' perspective," *Mathematical and Computer Modelling*, pp. 1–27, 2011.

[72] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs, and A. Walton, "Measuring and defining the experience of immersion in games," *International Journal of Human-Computer Studies*, vol. 66, no. 9, pp. 641–661, 2008.

[73] S. Jumisko-Pyykkö and M. M. Hannuksela, "Does context matter in quality evaluation of mobile television?" In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services MobileHCI 08*, Amsterdam, Netherlands: ACM, 2008, pp. 63–72.

[74] J. Juul, *Half-Real: Video Games Between Real Rules and Fictional Worlds*. MIT Press, 2005.

[75] K. Kaida, M. Takahashi, T. Åkerstedt, A. Nakata, Y. Otsuka, T. Haratani, and K. Fukasawa, "Validation of the Karolinska sleepiness scale against performance and EEG variables," *Clinical Neurophysiology*, vol. 117, no. 7, pp. 1574–1581, 2006.

[76] K. J. Kim, S. S. Sundar, and E. Park, "The effects of screen-size and communication modality on psychology of mobile device users," in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11*, Vancouver, BC, Canada, 2011, pp. 1207–1212.

[77] G. H. Klem, H. O. Lüders, H. H. Jasper, and C. Elger, "The ten-twenty electrode system of the International Federation," *Electroencephalography and Clinical Neurophysiology*, vol. 52, no. 3, pp. 3–6, 1999.

[78] H. O. Knoche, J. D. McCarthy, and M. a. Sasse, "Can small be beautiful? assessing image resolution requirements for mobile TV," in *Proceedings of the 13th annual ACM international conference on Multimedia*, Singapore: ACM, 2005, pp. 829–838.

[79] H. C. Koerper and N. A. Whitney-Desautels, "Astragalus Bones: Artifacts Or Ecofacts?" *Pacific Coast Archaeological Society Quarterly*, vol. 35, no. 2-3, pp. 69–80, 1999.

[80] H. Korhonen and E. M. I. Koivisto, "Playability heuristics for Mobile Games," in *Proceedings of the 2nd international conference on Digital interactive media in entertainment and arts*, Perth, Australia: ACM, 2007, pp. 28–35.

[81] K. Kumar, J. Liu, Y. H. Lu, and B. Bhargava, "A survey of computation offloading for mobile systems," *Mobile Networks and Applications*, vol. 18, no. 1, pp. 129–140, 2013.

[82] S. Kumar, L. Xu, M. K. Mandal, and S. Panchanathan, "Error resiliency schemes in H.264/AVC standard," *Journal of Visual Communication and Image Representation*, vol. 17, no. 2, pp. 425–450, 2006.

[83] N. Lazzaro and K. Keeker, "What's My Method?: A Game Show on Games," in *CHI'04 extended abstracts on Human factors in computing systems*, Vienna, Austria: ACM, 2004, pp. 1093–1094.

[84] P. Le Callet, S. Möller, and A. Perkis, *Qualinet white paper on definitions of quality of experience*. European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), 2013. [Online]. Available: http://www.qualinet.eu/images/stories/QoE%7B%5C_%7Dwhitepaper%7B%5C_%7Dv1.2.pdf.

[85] Y. Liu and H. Li, "Exploring the impact of use context on mobile hedonic services adoption: An empirical study on mobile gaming in China," *Computers in Human Behavior*, vol. 27, no. 2, pp. 890–898, 2011.

[86] R. Lopes and R. Bidarra, "Adaptivity challenges in games and simulations: A survey," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 3, no. 2, pp. 85–99, 2011.

[87] N. Maniar, E. Bennett, S. Hand, and G. Allan, "The effect of mobile phone screen size on video based learning," *Journal of Software*, vol. 3, no. 4, pp. 51–61, 2008.

[88] J. H. McDonald, *Handbook of biological statistics*. Sparky House Publishing Baltimore, 2009, vol. 2.

[89] S. Möller, C. Kühnel, K.-P. Engelbrecht, I. Wechsung, and B. Weiss, "A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction," in *International Workshop on Quality of Multimedia Experience, QoMEx 2009*, San Diego, California, USA: IEEE, 2009, pp. 7–12.

[90] S. Möller, "Skalierung [scaling]," in *Quality Engineering: Qualität kommunikationstechnischer Systeme [Quality engineering: quality of communication technology systems]*, Heidelberg: Springer, 2010, pp. 41–55.

[91] S. Möller, J.-N. Antons, J. Beyer, S. Egger, E. N. Castellar, L. Skorin-Kapov, and M. Sužnjevic, "Towards a New ITU-T Recommendation for Subjective Methods Evaluating Gaming QoE," in *7th International Workshop on Quality of Multimedia Experience (QoMEX)*, 2015, pp. 1–6.

[92] S. Möller, S. Schmidt, and J. Beyer, "Gaming taxonomy: An overview of concepts and evaluation methods for computer gaming QoE," in *5th International Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt, Austria: IEEE, 2013, pp. 236–241.

[93] A. Morello and V. Mignone, "DVB-S2: The second generation standard for satellite broad-band services," *Proceedings of the IEEE*, vol. 94, no. 1, pp. 210–226, 2006.

[94] L. Nacke, "Affective Ludology, Flow and Immersion in a First-Person Shooter: Measurement of Player Experience," *Loading...: The Journal of the Canadian Game Studies Association*, vol. 3, no. 5, pp. 1–21, 2009.

[95] L. E. Nacke, M. N. Grimshaw, and C. A. Lindley, "More than a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game," *Interacting with Computers*, vol. 22, no. 5, pp. 336–343, 2010.

[96] L. E. Nacke, A. Nacke, and C. A. Lindley, "Brain Training for Silver Gamers: Effects of Age and Game Form on Effectiveness, Efficiency, Self-Assessment, and Gameplay Experience.," *CyberPsychology & Behavior*, vol. 12, no. 5, pp. 493–499, 2009.

[97] J. Nakamura and M. Csikszentmihalyi, "The Concept of Flow," in *Handbook of positive psychology*, Oxford University Press, 2002, pp. 89–105.

[98] J. Pace, "The Ways We Play, Part 2: Mobile Game Changers," *Computer*, vol. 46, no. 4, pp. 97–99, 2013.

[99] L. Pantel and L. Wolf, "On the impact of delay on real-time multiplayer games," in *NOSSDAV '02 Proceedings of the 12th international workshop on Network and operating systems support for digital audio and video*, Miami, Florida, USA: ACM, 2002, pp. 23–29.

[100] K. Poels, Y. D. Kort, and W. IJsselsteijn, "The fun of gaming: Measuring the human experience of media enjoyment," Eindhoven University of Technology, Tech. Rep., 2009, pp. 1–46.

[101] PricewaterhouseCoopers AG, *Media Trend Outlook 2015 Cloud Gaming: Vielseitiger Einfluss auf die Videospiel-Industrie*, 2015. [Online]. Available: https://www.pwc.de/de/technologie-medien-und-telekommunikation/assets/pwc-media-trend-outlook%7B%5C_%7Dcloud-gaming.pdf.

[102] Z. Qi, J. Yao, C. Zhang, M. Yu, Z. Yang, and H. Guan, "VGRIS: Virtualized GPU Resource Isolation and Scheduling in Cloud Gaming," *ACM Transactions on Architecture and Code Optimization*, vol. 11, no. 2, pp. 1–25, 2014.

[103] A. Raake, M. Garcia, and S. Möller, "T-V-MODEL : PARAMETER-BASED PREDICTION OF IPTV QUALITY," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, Nevada, USA: IEEE, 2008, pp. 1149–1152.

[104] F. Rheinberg, R. Vollmeyer, and S. Engeser, *Die Erfassung des Flow-Erlebens*. Institut für Psychologie, Universität Potsdam, 2003. [Online]. Available: http://psych-server.psych.uni-potsdam.de/people/rheinberg/messverfahren/Flow-FKS.pdf.

[105] O. K. B. Richstad, "User Preferences for Video Game Delivery - A Case Study of Cloud Gaming," Master thesis, Norwegian University of Science and Technology (NTNU), 2015.

[106] F. Roth, "Using low cost FPGAs for realtime video processing," Master thesis, Masaryk University, 2011.

[107] M. D. Rugg and M. G. H. Coles, *Electrophysiology of mind: Event-related brain potentials and cognition.* Oxford University Press, 1995.

[108] C. Schaefer and T. Enderes, "Subjective quality assessment for multiplayer real-time games," in *NetGames '02 Proceedings of the 1st workshop on Network and system support for games*, Braunschweig, Germany: ACM, 2002, pp. 74–78.

[109] E. Schmider, M. Ziegler, E. Danay, L. Beyer, and M. Bühner, "Is It Really Robust?: Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption," *Methodology*, vol. 6, no. 4, pp. 147–151, 2010.

[110] D. K. Schoenenberg, "The Quality of Mediated-Conversations under Transmission Delay," PhD thesis, Technische Universität Berlin, 2015.

[111] R. Schreier and A. Rothermel, "Motion adaptive intra refresh for the H.264 video coding standard," *IEEE Transactions on Consumer Electronics*, vol. 52, no. 1, pp. 249–253, 2006.

[112] I. Slivar, L. Skorin-Kapov, and M. Suznjevic, "Cloud Gaming QoE Models for Deriving Video Encoding Adaptation Strategies," in *Proceedings of the 2016 ACM Multimedia Systems Conference*, Klagenfurt, Austria: ACM, 2016, pp. 1–12.

[113] I. Slivar, M. Suznjevic, and L. Skorin-Kapov, "The impact of video encoding parameters and game type on QoE for cloud gaming: A case study using the steam platform," in *7th International Workshop on Quality of Multimedia Experience, QoMEX 2015*, Costa Navarino, Greece, 2015, pp. 1–6.

[114] I. Slivar, M. Suznjevic, L. Skorin-Kapov, and M. Matijasevic, "Empirical QoE study of in-home streaming of online games," in *13th Annual Workshop on Network and Systems Support for Games (NetGames)*, Nagoya, Japan: IEEE, 2014, pp. 1–6.

[115] G. J. Sullivan, J. R. Ohm, W. J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[116] M. Suznjevic, J. Beyer, L. Skorin-Kapov, S. Möller, and N. Sorsa, "Towards understanding the relationship between game type and network traffic for cloud gaming," in *IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, Chengdu, China: IEEE, 2014, pp. 1–6.

[117] L. Vangelista, N. Benvenuto, S. Tomasin, C. Nokes, J. Stott, A. Filippi, M. Vlot, V. Mignone, and A. Morello, "Key technologies for next-generation terrestrial digital television standard DVB-T2," *IEEE Communications Magazine*, vol. 47, no. 10, pp. 146–153, 2009.

[118] S. Wang and S. Dey, "Cloud mobile gaming: modeling and measuring user experience in mobile wireless networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 16, no. 1, 2012.

[119] S. Wang and S. Dey, "Modeling and characterizing user experience in a cloud server based mobile gaming approach," in *Global Telecommunications Conference (GLOBECOM)*, Honolulu, HI, 2009, pp. 1–7.

[120] A. F. Wattimena, R. E. Kooij, J. M. van Vugt, and O. K. Ahmed, "Predicting the perceived quality of a First Person Shooter: the Quake IV G-model," in *NetGames '06 Proceedings of 5th ACM SIGCOMM workshop on Network and system support for games*, Singapore: ACM, 2006, pp. 1–4.

[121] D. Weibel and B. Wissmath, "Immersion in computer games: The role of spatial presence and flow," *International Journal of Computer Games Technology*, vol. 2011, 2011.

[122] D. Weibel, B. Wissmath, S. Habegger, Y. Steiner, and R. Groner, "Playing online games against computer- vs. human-controlled opponents: Effects on presence, flow, and enjoyment," *Computers in Human Behavior*, vol. 24, no. 5, pp. 2274–2291, 2008.

[123] T. Westermann, "User Acceptance of Mobile Notifications," PhD thesis, Technische Universität Berlin, 2016.

[124] M. J. P. Wolf, "Genre and the video game," *The medium of the video game*, pp. 113–134, 2002. [Online]. Available: http://www.robinlionheart.com/gamedev/genres.xhtml.

[125] S. Zander and G. Armitage, "Empirically measuring the QoS sensitivity of interactive online game players," in *Proceedings of ATNAC*, Sydney, Australia, 2004, pp. 1–8.