Assessing the Quality of Experience of Cloud Gaming Services

vorgelegt von
M. Sc.
Steven Schmidt
ORCID: 0000-0001-6620-1997

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. Jan Nordholz
Gutachter: Prof. Dr.-Ing. Sebastian Möller
Gutachterin: Prof. Dr. Lea Skorin-Kapov
Gutachter: Dr. Raimund Schatz

Tag der wissenschaftlichen Aussprache: 07.07.2021

Berlin 2021

Zusammenfassung

Die Videospielebranche hat den größten Unterhaltungssektor unserer Zeit geschaffen, der schnell wächst und das Leben vieler Generationen weltweit beeinflusst. Besonders hochkomplexe Spiele erfordern jedoch, dass Spieler über leistungsstarke Hardware verfügen. So wurde in den letzten Jahren ein neues Konzept namens Cloud Gaming entwickelt, mit dem Nutzer diese Spiele, die auf einem Cloud-Server ausgeführt werden, fernsteuern können. Während der Ansatz verschiedene Vorteile mit sich bringt, stellt die zusätzliche Netzwerkverbindung und Videoverarbeitung viele neue technische Herausforderungen für Netzbetreiber und Dienstanbieter. Um diese Hindernisse zu überwinden und die Zufriedenheit ihrer Kunden zu gewährleisten, haben Unternehmen ein starkes Interesse daran, das Nutzungserleben (Quality of Experience, QoE) der Spieler zu untersuchen und vorherzusagen.

Traditionelle QoE-Evaluierungen von Multimediadiensten verwenden kontrollierte subjektive Experimente, bei denen die Teilnehmer gebeten werden, ihre Meinung zu präsentierten Stimuli, z. B. einer Netzwerkbedingung, unter Verwendung validierter Fragebögen nachträglich zu äußern. Um diesen Beurteilungsprozess zu beschreiben, schlagen Möller und Kollegen (2013) eine Taxonomie der Qualitätsaspekte von Cloud-Gaming-Systemen vor. Die Anwendbarkeit etablierter Bewertungsmethoden zur Messung des hochgradig mehrdimensionalen Konstrukts der Gaming-QoE ist jedoch sehr begrenzt. Da sich das Gebiet der Gaming-QoE noch in der Anfangsphase befindet, fehlen validierte Bewertungsmethoden, die speziell für Cloud-Gaming-Dienste entwickelt wurden.

Ziel der vorgestellten Forschung ist es daher, eine umfassende, zuverlässige und validierte Bewertungsmethode für die Gaming-QoE für Cloud-Gaming-Dienste zu entwickeln und zu evaluieren. Die Methode ermöglicht die Gestaltung subjektiver Tests zur Erstellung von Datensätzen für Qualitätsvorhersagemodelle und ermöglicht ein detailliertes Verständnis der Beziehungen zwischen einer Vielzahl von Qualitätsaspekten.

Als erster Schritt in Richtung eines einheitlichen Evaluierungsansatzes wurden verfügbare Fragebögen auf kompakte Weise zusammengefasst. Da kein Messinstrument zur Bewertung der Interaktionsqualität identifiziert wurde, wurde ein neuer Fragebogen, die Gaming Input Quality Scale (GIPS), entwickelt. Darüber hinaus wurde für dessen Erstellung ein neues Framework zur Bewertung der Gaming-QoE mithilfe eines Crowdsourcing-Ansatzes entworfen. Zuletzt wurde die Evaluierungsmethode basierend auf einem großen Datensatz, der dominante Netzwerk- und Codierungsbedingungen umfasst, unter Verwendung der Strukturgleichungsmodellierung untersucht.

Die Forschung ergab, dass das Crowdsourcing-Framework vergleichbare Ergebnisse wie Laborstudien liefern kann. Zudem wurde gezeigt, dass die entwickelte Evaluierungsmethode zuverlässige und gültige Benutzerbewertungen liefert und gleichzeitig ein Testdesign ermöglicht, das zur Entwicklung von Qualitätsvorhersagemodellen geeignet ist. Zusammenfassend sind die Hauptbeiträge der Arbeit (1) eine empirische Validierung einer Taxonomie von Qualitätsaspekten von Cloud-Gaming-Diensten sowie der angewandten Bewertungsmethoden, (2) ein neues Framework für die Durchführung von

Qualitätsbewertungsstudien in einer häuslichen Umgebung, die möglicherweise die ökologische Validität von Studienergebnissen erhöht, (3) ein psychometrisch validiertes und zuverlässiges Instrument zur Messung der Interaktionsqualität und (4) zahlreiche Beiträge zu Standardisierungsaktivitäten in Bezug auf Einflussfaktoren (ITU-T Rec. G.1032), subjektive Bewertungsmethoden (ITU-T Rec. P.809) und ein Planungsmodell zur Vorhersage der QoE von Cloud Gaming (ITU-T Rec. G.1072), die in Zukunft zu zuverlässigeren, validen und vergleichbaren Forschungsergebnissen führen können.

Abstract

The video gaming industry created the largest entertainment sector in our times that is rapidly growing and impacts the life of many generations worldwide. However, especially highly complex games demand players to possess powerful hardware. Thus, in the recent years a new concept called Cloud Gaming evolved that allows players to remotely control these games running on a cloud server. Whereas the approach results in various advantages, the additional network connection and video processing provokes many new technical challenges for network operators and service providers. To overcome those obstacles, and to ensure the satisfaction of their costumers, companies have a strong interest in evaluating and predicting the Quality of Experience (QoE) of players.

Traditional QoE evaluations of multimedia services make use of controlled subjective experiments in which participants are asked to retrospectively express their opinion on presented stimuli, impacted by e.g., a network condition, using validated questionnaires. Aiming to describe this judgement process, a taxonomy of quality aspects of cloud gaming systems was suggested by Möller and colleagues (2013). However, the applicability of established assessment methods to measure the highly multi-dimensional construct of gaming QoE is very limited. Also, as the field of gaming QoE is still in a nascent phase, there is a lack of validated assessment methods specifically developed for cloud gaming services.

Thus, the aim of the presented research is to develop and evaluate a comprehensive, reliable, and validated assessment method of gaming QoE for cloud gaming services. The method will allow the design of subjective tests to create datasets for quality prediction models and enable a detailed understanding of the relationships between a broad range of quality aspects.

As a first step towards a unified evaluation approach, available questionnaires were combined in a concise way. As no measurement tool to assess the interaction quality was identified, a new questionnaire, the Gaming Input Quality Scale (GIPS), was developed. Furthermore, for its creation, a new framework to assess Gaming QoE using a crowdsourcing approach was designed. Lastly, based on a large dataset including dominant network and encoding conditions, the evaluation method was investigated using structural equation modelling.

The research revealed that the crowdsourcing framework yielded comparable results to lab studies. Additionally, it was shown that the developed assessment method provided reliable and valid user ratings while allowing a test design suitable to develop quality prediction models. In summary, the main contributions of the thesis are (1) an empirical validation of a taxonomy of quality aspects of cloud gaming services as well as applied assessment methods, (2) a new framework to conduct quality assessment studies in a home environment which potentially increases the ecological validity of study results, (3) a psychometrically validated, and reliable instrument to measure the interaction quality, and (4) numerous contributions to standardization activities regarding influencing factors (ITU-T Rec. G.1032), subjective evaluation methods (ITU-T Rec. P.809), and an opinion model predicting cloud

gaming QoE (ITU-T Rec. G.1072) that can lead to more reliable, valid, and comparable research results in the future.

Acknowledgements

The path to finalizing a dissertation is long and full of obstacles, but I have met numerous people who helped me overcome all these challenges over the past years, and I am truly thankful.

Foremost, I would like to express my deepest thanks to my supervisor Prof. Dr.-Ing. Sebastian Möller for providing me with guidance throughout all these years. Thank you very much for your continuous support, for all the time we spent with discussions, for your recommendations about various decisions I had to make, and for enabling exciting opportunities for my personal growth.

My time at the Quality and Usability Lab was not only joyful due to an interesting research topic but first and foremost due to all my great colleagues, of whom some turned into good friends.

Among these, my greatest thanks go to Saman Zadtootaghaj. You were an inspiration in many aspects of our close collaboration. Thank you for being such a good friend, for your reliable and professional support, for always being honest and open-minded, and for the many interesting discussions we had to progress in our research. I want to give special thanks to Babak Naderi. Thank you for always being there, sharing your experiences, introducing me to the challenges of crowdsourcing, and the many interesting discussions about statistics we had. You helped me not lose myself in my thought processes and are truly a great team player. Furthermore, I would like to express my gratitude to Saeed Shafiee Sabet for the delightful and constructive collaboration. You showed me different perspectives on our complex topic, and I could always count on you. A great thank you also goes to Nabajeet Barman. It was a pleasure to work with you, and I am thankful for all the great advice you gave me on the academic and personal levels. Thank you, Justus Beyer, for initiating the gaming QoE research at the QU Lab, introducing me to your work and the interesting research proposal you wrote.

My travel to Sydney with Benjamin Weiss, Thilo Michael, and Stefan Uhrig was one of the most remarkable experiences I had in the last years. I always like to think back and this time and also want to thank you three for your scientific support. Thank you, Benjamin, for your help with the development of my questionnaire, thank you, Thilo, for your support with our daily teaching life, and thank you, Stefan, for the many fruitful discussion we had about user engagement and statistical methods as well as with proofreading my dissertation. And also, thank you to Prof. Andrew Johnston for making this visit possible by hosting us at the UTS.

Thank you to Jan-Niklas Voigt-Antons, Tanja Kojic, Falk Schiffner, Gabriel Mittag, Martin Burghart, Maija Poikela, Patrick Ehrenbrink, Friedemann Köster and many more for being such a great team at the QU Lab, for many interesting discussions, advise, and encouragements. And of course, also a big thank you to Irene Hube-Achter and Yasmin Hillebrenner for your amazing administrative work and for always being supportive, pragmatic, and for creating a pleasant atmosphere in our office. You did more than one can hope for. Thank you also to Tobias Jettkowski for your technical support, which enabled various user studies I conducted.

I will cherish the memories of all our joint travels, I will miss the cheese fondue in Geneva with you, and I wish you all the best in your future life.

I would also like to thank my committee members, Prof. Dr. Lea Skorin-Kapov and Dr. Raimund Schatz. I highly appreciate your effort and help in co-examining my dissertation. Also outside of the QU Lab, I would like to thank colleagues within the ITU-T Study Group 12 and Qualinet members for their great work.

Lastly, I want to express my dearest gratitude to my family and friends for your understanding and ongoing support throughout these years. You allowed me to develop myself, to work carefree, and encouraged me at any time.

Table of Contents

List of Abbreviations

AC Service Acceptability

ACR Absolute Category Rating

ACR-HR ACR with hidden reference

ANOVA Analysis of Variance

AP Appeal

APM Actions per minute

AQ Audio Quality

AVE Average Variance Extracted

AWS Amazon Web Services

CFA Confirmatory Factor Analysis

CFI Comparative Fit Index

CG Crowdgaming

CH Challenge

CMIN Minimum Discrepancy

CN Controllability

CO Competency

CR Composite Reliability

CS Crowdsourcing

DMOS Differential MOS

DoF Degrees of Freedom

EC-ACR extended continuous ACR

ECG Electrocardiography

EDA Electrodermal activity

EEG Electroencephalography

EFA Exploratory Factor Analysis

EMG Electromyography

FL Flow

fps Frames per second

FSS Flow State Scale

GEQ Game Experience Questionnaire

GIPS Gaming Input Quality Scale

GPU Graphics Processing Unit

GTA Grand Theft Auto

HCI Human-Computer Interaction

HDR High-dynamic-range

HDTV High Definition Television

HEVC High Efficiency Video Coding

HIT Human Intelligence Task

HMD Head-Mounted Display

HR Heart rate

HUD Heads-Up Display

IC Intuitive Controls

IEQ Immersive Experience Questionnaire

IF Immediate Feedback

iGEQ in-game Game Experience Questionnaire

IM Immersion

IR Item Reliability

ITU-T Telecommunication Standardization Sector of the International Telecommunication Union

KMO Kaiser-Meyer-Olkin

LE Learnability

MaxRH Maximal Reliability

MDA Mechanics-Dynamics-Aesthetics

ML Maximum-Likelihood

MMO Massively Multiplayer Online

MOS Mean Opinion Score

MSV Maximum Shared Variance

NA Negative Affect

NPX Negative Player Experience

PA Positive Affect

PAF Principal Axis Factoring

PCA Principal Component Analysis

PClose p of Close Fit

PENS Player Experience and Need Satisfaction

PI Performance Indication

PLCC Pearson Linear Correlation Coefficient

PLEXQ Playful Experiences Questionnaire

PPX Positive Player Experience

PR Playing Performance

PX Player Experience

PXI Player Experience Inventory

QoE Quality of Experience

QoS Quality of Service

RE Responsiveness

Rec. Recommendation

RMSE Root Mean Square Error

RMSEA Root Mean Square Error Of Approximation

RQ Research Question

RTCP Real-time Control Protocol

RTP Real-time Transport Protocol

RTSP Real Time Streaming Protocol

RTT Round-Trip Time

SE Standardized Effect estimates

SEM Structural Equation Modeling

SG12 Study Group 12

SI Spatial Information index

SRCC Spearman's Rank Correlation Coefficient

SRMR Standardized Root Mean Square Residual

SVQ Spatial Video Quality

TCP Transmission Control Protocol

TE Tension

TI Temporal Information index

TVQ Temporal Video Quality

UDP Datagram Protocol

UESz User Engagement Scale

URL Uniform Resource Locator

UX User Experience

VD Video Discontinuity

VF Video Fragmentation

VIF Variance Inflation Factor

VL Suboptimal Video Luminosity

VM Virtual Machine

VQ Video Quality

VU Video Unclearness

WebRTC Web Real Time Communication

Chapter 1

Introduction

1.1 Motivation

In 1972, Bushnell, Dabney, and Alcorn at Atari created the world-famous arcade game PONG. Only five years later, popular and affordable gaming consoles such as the Atari 2600, the Intellivision, and the Color TV-Game by Nintendo enabled players to enter enjoyable virtual gaming worlds in their home environments [1]. Meanwhile, modern digital games became an impressive form of art and entertainment. Not only do steadily new game concepts evoke new challenges to players, some games become even life-like due to a richness of graphical details that most likely nobody would have ever thought of. Modern games allow rich storytelling based on novels, include orchestra music particularly composed for games, and have highly realistic avatar representation allowing players to identify themselves with their characters.

The gaming industry has unprecedented managed to intrinsically motivate users to interact with their services. According to the latest report of Newzoo, the leading provider of market intelligence covering the global gaming industry, there will be a total of 2.7 billion players across the globe by the end of 2020. The global games market will generate revenues of $159.3 billion in 2020, an increase of 9.3 % compared to the year before [2]. This surpasses the movie industry (box offices and streaming services) by a factor of four and almost three times the music industry market in value [3]. One of many reasons for the success of the gaming industry is also its wide audience. On the contrary to what many people may believe, according to the latest report of the Entertainment Software Association, the trade association of the video game industry in the United States, over 40 % of players are older than 35 years and only 21 % are younger than 18. Additionally, 41 % are female players and 65 % of players play together and feel connected with others [4]. Even though the Asia-Pacific region generated the most revenue (49 %), the success of the gaming industry is a global phenomenon, as also in Europe (19 %) and North America (25 %) [3] millions of gaming enthusiasts spend their money to relax, escape reality, and connect with friends and family.

Over the last decade, applications using internet connectivity and cloud computing have extended tremendously. They now include not only services like the transmission of files, web browsing, speech, and audio-visual communication, but also to a significant percentage purely entertainment-related uses. Here, apart from video streaming services (i.e., IPTV, YouTube, Netflix), the rapidly growing domain of online gaming emerged in the late 1990s and early 2000s allowing social relatedness to a great number of players. During traditional online gaming, typically, the game logic and the game user interface are locally executed and rendered on the player's hardware. The client device is connected via

the internet to a game server to exchange information influencing the game state, which is then shared and synchronized with all other players connected to the server. However, in 2009 a new paradigm comparable to the rise of Netflix for video consumption and Spotify for music consumption emerged: CLOUD GAMING.

Inspired by new possibilities due to technological advancements in network infrastructures and video coding, Steve Perlman promoted the first known cloud gaming service called OnLive at the Game Developers Conference. On the contrary to traditional online gaming, cloud gaming is characterized by the execution of the game logic, rendering of the virtual scene, and video encoding on a cloud server, while the player's client is solely responsible for video decoding and capturing of client input. While cloud gaming is still a new type of service and has to further establish, Newzoo estimates that cloud gaming will generate revenues of $585 million by the end of 2020 and most likely will be worth $4.8 billion by 2023 [5]. This is evident by more and more cloud gaming services being released these days. Google Stadia, Sony's PlayStation Now, and GeForce Now are arguably the most popular cloud gaming services, but also other solutions such as Microsoft's Project xCloud, Amazon's Luna, Shadow, Paperspace and Parsec, Telekom's MagentaGaming, and LiquidSky are on the rise.

"The reason games moving to the cloud is exciting, is that there's a tremendous amount of people who want to play games, but don't have the equipment to do it."

— Andrew Fear, senior product manager of GeForce Now [6] —

The above statement of Andrew Fear points out one of the biggest benefits of cloud gaming: the reduction of hardware expenses for users due to the very low-performance requirements of the players' client device, which are otherwise potentially unable of running high-quality games. However, there is a variety of other promising advantages of cloud gaming. Cloud gaming enables device and platform independence, i.e., a game developed for a PC running a Windows operating system can be played on an Android phone as the client device only has to decode the video and allow in return some inputs of the player. This also improves accessibility as games can be played on mobile devices and an installation of the game on each user device is not required anymore. Finally, due to the mandatory access to the cloud server, cloud gaming prevails piracy.

However, while with the cloud gaming paradigm, the complexity and execution requirements of the game are entirely decoupled from the capabilities of the user's device, the quality of cloud gaming systems depends primarily on the additionally added Internet connection. Apart from conflicting business ideas with game publishers by offering full title games in their service, the pioneer of cloud gaming OnLive was subject to various criticism. The service did not reach acceptance due to being not reliable in providing a satisfying video quality [7] particularly for very demanding games, and due to interaction delays beyond acceptable thresholds, e.g., 100 ms for action-based games such as First-person shooters. Shea et al. measured the interaction delay and image quality of the OnLive system and conclude that the encoding process alone added about 100 to 120 ms latency to the network transmission delay [8]. Evidently, due to its interactivity and video-based presentation, cloud gaming places great demands on the technical system and service components. Even though apparently the network infrastructure and encoding performance were not sufficient at that time, OnLive, which was shut down in 2015 after being bought by Sony, did not consider one of the most important aspects of their service properly enough: THE USER.

Despite the technical challenges and aspiration to build a functional system, it is obvious that a system which is designed to be used by humans must satisfy their needs and create a great user experience. To be competitive under these conditions, it is highly important for service providers to offer an optimal Quality of Experience (QoE), which is described by the degree of delight or annoyance of the user [9], to their customers. This can be achieved through proper planning of network infrastructures and system configurations based on planning models as well as by continuous and automated monitoring of the users' QoE to allow an optimized resource allocation and quality control. However, whereas the quality of speech, audio-visual communication services, and of video streaming services has seen thorough investigation in both science and industry for several decades, the quality of interactive video games has only recently been addressed. Here, two fundamentally different approaches are present in recent gaming research. On one hand, the Quality of Service (QoS) mostly linked to network-engineering is investigated from a technical perspective analyzing the impact of individual network characteristics and encoding settings on user satisfaction or on game performance characteristics. Apart from many other factors that influence a gaming experience, the most dominant network bottlenecks identified are limited bandwidth, latency, and packet loss which are all interconnected and have an impact on the video stream and interactions of players. On the other hand, the human factors perspective is taken and features experienced during the gaming activity are subject of the research. In this context, especially the User Experience (UX) research traditions have identified and evaluated concepts such as immersion and flow but also various emotional responses as well as player characteristics (cf. [10]).

Yet, the link between QoS and QoE with respect to cloud gaming services has only been addressed for individual use cases in the past. Thus, planning and monitoring models for gaming QoE prediction, which capture this relationship in a generalizable way for different games and technical scenarios and consequently allow service providers to offer an optimized QoE to their users, are still missing.

To develop such models, datasets containing user ratings collected under various relevant conditions are required. Therefore, two major requirements must be considered. First, a wide knowledge about influencing factors, referred to as quality factors in the following, must be available to sufficiently cover all relevant aspects when planning empirical studies. Second, reliable and valid quality evaluation methods must be used in these studies allowing users to reflect upon their quality judgement. By the time the work on the present dissertation started, especially the assessment methods used in the research community were still very limited. Many researchers used self-developed tools for vaguely defined concepts without following standardizes procedures. Standardized assessment method of gaming QoE would be of great use for researchers and industries as they allow a comparison between conducted studies and offer means to assess valid and reliable ratings with respect to gaming QoE. For this reason, Study Group 12 (SG12) of the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) has decided during the 2013-2016 Study Period to start work on three new work items:

1. Definition of subjective methods for evaluating the quality experienced during gaming activities (P.GAME)
2. Identification of factors affecting QoE in gaming applications (G.QoE-gaming)
3. Predicting gaming QoE with the help of an opinion model for gaming applications (G.OMG)

Each work item should result in a corresponding Recommendation (Rec.), as part of the P-Series (Telephone transmission quality, telephone installations, local line networks) or G-series (Transmission systems and media, digital systems and networks) of Recommendations, as indicated by the preliminary abbreviation of the respective work item [11]. From a bird's-eye view, the aim of the present research is to contribute to these activities and further reduce the gap between QoS and QoE research for cloud gaming services. A special focus will be given to the assessment methods of gaming QoE, as they form a fundamental basis for future research.

1.2 Scope and Research Questions

The very high complexity of cloud gaming systems with respect to their technical components, a multitude of possible influencing factors, and the multi-dimensionality of gaming QoE requires to some extent a reduction of the conceivable scope of research.

For the presented research, the perspective of network operators and cloud gaming service providers is taken who have an economic interest to optimize their service to improve the subjective experience and ultimately the service acceptance of their customers. The research should enable them to improve the QoE of their customers but does not focus on the design of a game but rather on the impact of network and encoding parameters. Consequently, as these providers only have limited access to players' motivations and preferences, nor an impact of the design of the played games, these aspects are beyond the scope of the research. However, some information about the game content itself is of relevance as this is an important influencing factor on gaming QoE.

With respect to the technologies considered, the research will adhere to the current state of cloud gaming services but it is to be expected that soon updates regarding higher resolutions, frame rates, and other technical aspects are required. Virtual reality gaming requiring 3D rendering devices, mobile input and output devices, as well as input devices other than keyboard and mouse are not within the scope of the presented research. However, even though the focus of this work is on cloud gaming services using a desktop PC, some findings may also apply to such systems as well as to online gaming, where the game is primarily executed on a client, or passive gaming video streaming applications, where only video content is streamed to passive viewers of the game.

With regards to test participants for subjective studies, primarily non-expert gamers are considered as it is reasonable to assume that expert gamers are not the target group of cloud gaming services since they are very sensitive towards technical impairments. Hence, it will be subject to further research if the findings of the presented work are also accurate for highly experienced gamers. Lastly, the influence of social factors, which are arguably important especially for multiplayer games, will not be covered by the research. Thus, only one participant at the same time will take part in subjective user studies.

To place the presented work in the frame of the state-of-the-art, it must be noted that the research will build up upon previous work in the domain of cloud gaming. To be highlighted here is a taxonomy about QoS and QoE aspect of cloud gaming presented by Möller, Schmidt, and Beyer in 2013 [12]. The cloud gaming taxonomy includes a categorization of influencing factors as well as interaction performance metrics, and quality aspects, and organizes them along three layers. The taxonomy will serve as a basis for the presented work. Quality aspects represent categories of recognizable characteristics relevant to the quality of a service (cf. Section 2.3).

As a consequence of the motivation behind the work and the scope described above, in the following the aim of the thesis and derived Research questions (RQ) will be summarized.

Aim of Thesis

The overarching aim of the thesis is to provide and evaluate a comprehensive assessment method of gaming QoE for cloud gaming services. Therefore, the completeness of the theoretical foundation - the cloud gaming taxonomy - will be investigated empirically and if necessary improved in order to provide a holistic overview of the most important gaming QoE aspects of a cloud gaming service. As a result, guidelines for the design of subjective tests and measurement tools for assessing a broad range of gaming QoE aspects will be provided. Lastly, the importance and interplay of concepts such as immersion and interaction quality for the overall gaming QoE judgement will be analyzed. As a consequence, the presented work will ultimately provide the necessary means for the development of gaming QoE prediction models.

In order to reach this aim, the following approachable research questions are derived:

RQ1 Is the cloud gaming taxonomy representing all relevant quality aspects?
RQ2 How can the broad range of quality aspects of gaming QoE be measured?
RQ3 How should a subjective test for assessing gaming QoE be designed?
RQ4 Is there an alternative to traditional laboratory studies for gaming QoE assessment?
RQ5 How relevant are the individual quality aspects for the overall gaming QoE?

1.3 Structure of Thesis

Following on this chapter, in Chapter 2 an introduction to research about Quality of Experience, the concept and challenges of cloud gaming, and fundamental aspects of the cloud gaming taxonomy will be provided.

Chapter 3 summarizes the state-of-the-art of assessment methods in the gaming domain, and introduces achievements of the work on standardization activities with respect to assessment methods as well as subjective test designs. Furthermore, a study comparing an interactive and passive test paradigm will be described, and an example of a standardized test procedure will be given.

In Chapter 4 a new methodology using a crowdsourcing approach for gaming QoE assessment will be introduced. A comparison of results obtained from traditional laboratory tests and the new crowd-based method will be performed. Finally, a large dataset of subjective ratings of various network and encoding conditions will be created using this method.

Chapter 5 describes the development of the Gaming Input Quality Scale (GIPS). The previously build dataset will be used for the development of the psychometrically validated, and reliable instrument, which aims at measuring the input quality of cloud gaming system.

Next, Chapter 6 presents research about the game content as an important influencing factor on gaming QoE. First, a comparison of different game scenarios with respect to their sensitivity towards network delays will be made. Following, a classification of game scenarios with respect to their sensitivity to network and encoding impairments will be presented.

In Chapter 7, a large-scale dataset developed in the scope of the ITU-T standardization activities will be presented. The dataset will be used to empirically validate the cloud gaming taxonomy. This step will allow the research community to obtain insight into the reliability and validity of the developed and used assessment methods throughout the present work, and to gain knowledge about the relationships of the various quality aspects such as video quality, input quality, and immersion.

Finally, in Chapter 8, a summary of the key contributions of the previous chapters, and answers to the research questions will be provided. The thesis closes with limitations of the presented work and an outlook on future work.

1.4 Publications the Thesis Is Based On

The following section describes the author's publications which form the basis of the presented work *and Where to Find Them* in the thesis.

- S. Möller, S. Schmidt, and J. Beyer, "Gaming Taxonomy: An Overview of Concepts and Evaluation Methods for Computer Gaming QoE", in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, 2013, pp. 236–241. DOI: 10.1109/QoMEX.2013.6603243.

This paper proposes a taxonomy of cloud gaming QoE aspects and forms a fundamental part of the whole thesis. The paper is especially part of Section 2.3. Sebastian Möller and Justus Beyer were mainly responsible for the paper writing process while the content was mainly based on the Bachelor thesis of the author of the present thesis [13].

- S. Schmidt, S. Zadtootaghaj, and S. Möller, "Towards the Delay Sensitivity of Games: There Is More Than Genres", in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–6. DOI: 10.1109/QoMEX.2017.7965676.

For this paper investigating the importance of specific game scenarios with respect to the influence of network delay on QoE, the author was responsible for all necessary processes including the study design, implementation of rating tools and setups, analysis of the results, paper writing as well as conduction of subjective user studies. The latter was supported by Saman Zadtootaghaj whereas Sebastian Möller contributed to the study design. The paper is a part of Section 6.1 in the thesis.

- S. Schmidt, S. Möller, and S. Zadtootaghaj, "A Comparison of Interactive and Passive Quality Assessment for Gaming Research", in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2018, pp. 1–6. DOI: 10.1109/QoMEX.2018.8463417.

For this paper in which two different test paradigms are compared, the author was responsible for all necessary steps during its creation including the study design, implementation of rating tools and setups, analysis of the results, paper writing as well as conduction of subjective user studies. The study design and conduction of subjective tests was supported by Saman Zadtootaghaj whereas Sebastian Möller contributed to the study design. The paper is a part of Section 3.2 in the thesis.

- S. Schmidt, B. Naderi, S. S. Sabet, *et al.*, "Assessing Interactive Gaming Quality of Experience Using a Crowdsourcing Approach", in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2020, pp. 1–6. DOI: 10.1109/QoMEX48832.2020.9123122.

In this paper, a crowdsourcing framework for gaming QoE assessment was described. The author was responsible for all necessary processes including the study designs, development of games, implementation of the crowdsourcing survey, analysis of the results, paper writing as well as running the crowdsourcing tests. The implementation of means to communicate between the games and the

server and design choices were supported by Babak Naderi. Saman Zadtootaghaj and Saeed Shafie Sabet assisted the paper writing process. Sebastian Möller supported the test designs. The paper is a fundamental part of Chapter 4.

- S. Zadtootaghaj, S. Schmidt, N. Barman, *et al.*, "A Classification of Video Games Based on Game Characteristics Linked to Video Coding Complexity", in *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*, IEEE, 2018, pp. 1–6. DOI: 10.1109/NetGames.2018.8463434.
- S. S. Sabet, S. Schmidt, S. Zadtootaghaj, *et al.*, "Delay Sensitivity Classification of Cloud Gaming Content", in *Proceedings of the 12th ACM International Workshop on Immersive Mixed and Virtual Environment Systems*, ser. MMVE '20, Istanbul, Turkey: Association for Computing Machinery, 2020, pp. 25–30, ISBN: 9781450379472. DOI: 10.1145/3386293.3397116.

The above co-author publications are related to this thesis with respect to the importance of the game scenarios for gaming QoE research presented in Section 6.1 and 6.2, but were mostly omitted from it for the sake of brevity and focus. The author of the present thesis was involved in the study design, including the implementation of rating tools and games, and analysis of the results for a majority of the publications. Additionally, the author assisted the conduction of subjective user studies and publication writing, but was not involved in the implementation of systems, machine learning approaches, and video material encoding.

The author of the thesis was strongly involved in activities at the ITU-T SG12, especially for the work items G.QoE-gaming (cf. Section 2.2.2), P.GAME (cf. Section 3.3), and G.OMG (cf. Section 3.3.2 and Section 7.1 and Section 7.4). These activities led to three corresponding ITU-T recommendations, ITU-T Rec. G.1032, ITU-T Rec. P.809, and ITU-T Rec. G.1072, respectively. Additionally, the work about crowdsourcing in the context of gaming QoE was published in the scope of the work item P.CrowdG (cf. Chapter 4).

ITU-T Contributions related to the work item G.QoE-gaming (ITU-T Rec. G.1032):

- S. Schmidt, S. Zadtootaghaj, and S. Möller, "Updates on the first draft of Influence Factors in Gaming Quality of Experience (QoE)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.41, 2017.
- S. Zadtootaghaj, S. Schmidt, and S. Möller, "Influence Factors on Gaming Quality of Experience (QoE)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.104, 2017.

ITU-T Contributions related to the work item P.GAME (ITU-T Rec. P.809):

- S. Schmidt, S. Zadtootaghaj, and S. Möller, "Update on the Proposal for a Draft New Recommendation on Subjective Evaluation Methods for Gaming Quality (P.GAME)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.17, 2017.
- S. Schmidt, S. Zadtootaghaj, S. Möller, *et al.*, "Update on the Proposal for a Draft New Recommendation on Subjective Evaluation Methods for Gaming Quality (P.GAME)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.98, 2017.
- S. Schmidt, S. Zadtootaghaj, S. Möller, F. Metzger, M. Hirth, and M. Sužnjević, "Subjective Evaluation Methods for Gaming Quality (P.GAME)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.205, 2018.

ITU-T Contributions related to the work item G.OMG (ITU-T Rec. G.1072):

- S. Schmidt, S. Zadtootaghaj, S. Möller, *et al.*, "Requirement Specification and Possible Structure for an Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.200, 2018.
- S. Schmidt, S. Zadtootaghaj, F. Schiffner, *et al.*, "Data Assessment for an Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.293, 2018.
- S. Schmidt, S. Zadtootaghaj, M. Utke, *et al.*, "First Draft for an Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.387, 2019.

- S. Schmidt, S. Zadtootaghaj, S. Möller, *et al.*, "Proposal for an Opinion Model Predicting Gaming QoE for Mobile Online Gaming", ITU-T Study Group 12, Geneva, ITU-T Contribution C.441, 2019.
- S. Schmidt, S. S. Sabet, S. Zadtootaghaj, *et al.*, "Proposal of a Content Classification for Cloud Gaming Services", ITU-T Study Group 12, Geneva, ITU-T Contribution C.444, 2019.
- S. Schmidt, S. Zadtootaghaj, S. Möller, *et al.*, "Performance Evaluation of the Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.445, 2019.
- S. Schmidt, S. Zadtootaghaj, S. Möller, B. Nabajeet, M. G. Martini, S. S. Sabet, and C. Griwodz, "Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.446, 2019.
- S. Schmidt, S. Zadtootaghaj, and S. Möller, "Corrigendum for ITU-T Recommendation G.1072: Opinion Model Predicting Gaming QoE ", ITU-T Study Group 12, Geneva, ITU-T Contribution C.511, 2020.

ITU-T Contributions related to the work item P.CrowdG:

- S. Schmidt, B. Naderi, S. Zadtootaghaj, *et al.*, "Guidelines for the Assessment of Gaming QoE Using Crowdsourcing", ITU-T Study Group 12, Geneva, ITU-T Contribution C.376, 2019.
- B. Naderi, S. Schmidt, S. Zadtootaghaj, *et al.*, "Draft text for P.CROWDG Recommendation Subjective Evaluation of Gaming Quality with a Crowdsourcing Approach", ITU-T Study Group 12, Geneva, ITU-T Contribution C.437, 2019.
- S. Schmidt, S. S. Sabet, B. Naderi, *et al.*, "Evaluation of Interactive Test Paradigm for P.CROWDG Recommendation", ITU-T Study Group 12, Geneva, ITU-T Contribution C.485, 2020.

Chapter 2

Quality Factors and Feature Space of Cloud Gaming Services

2.1 Quality of Experience Research

The evaluation of the quality of multimedia services has attracted the attention of researchers for many years. Whereas traditional telecommunication and broadcasting services have always been designed following quality considerations, perceived quality has only recently been a major topic for interactive internet services, facing a paradigm shift from QoS towards QoE [35]. In parallel to this, the UX related to interactive services has been addressed by Human-Computer Interaction and Human Factors experts with qualitative and quantitative methods, and processes for service design have been developed. These efforts have led to stable concepts of QoE and UX, as well as applications of these concepts mostly to media transmission services (for QoE) and human-computer interaction services (for UX).

With the scope set for this work, especially with taking the perspective of cloud gaming service providers, it was motivated in the beginning why the users' Quality of Experience of a cloud gaming service is fundamentally important. To allow also readers without expertise in this field to follow the research presented in this thesis, some essential concepts and terms must be introduced.

Providing a definition for the term *quality* is more complicated than it appears, even though most people use the word regularly in everyday life. A search in the Cambridge Dictionary would, for example, result in some of the following definitions: quality means "how good or bad something is", quality relates to "a high standard" and to "a characteristic or feature of someone or something," as well as "the degree of excellence of something." Juran argues in his Quality Handbook that many suppliers in the past defined quality as conformance to the specification of their product. In contrast, most customers consider if the features of a product respond to their needs. Consequently, he introduced a more comprehensive definition of quality as the "fitness for use" [36]. Over the years, the concepts' definition has been slowly evolving and refined. Something noticeable about quality is that a judgment of quality can change over a certain period of time, or more concretely depending on previous *experiences* and current expectations. While someone did consider watching a movie on one of the first High Definition Television (HDTV) devices in 1998 as very good quality, the same may not hold true anymore for someone who is frequently watching movies on a TV with a 4k resolution using High-dynamic-range (HDR). Alternatively, disruptive lighting or noise condition in the user's environment can significantly impact the experience, i.e., the context of using a service or product is essential.

This example already implies that the terms quality and experience are somehow connected and are, considered separately, not solely a result of product/service features or of a users' achievement. It also becomes clear that, from a user's perspective, the QoS, which is a well-established research domain for many years, is not sufficient enough. According to the ITU Rec. E.800 [37], QoS is defined as:

Quality of Service (QoS)

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

To provide an intuitive and complete definition of the holistic concept Quality of Experience, the members of the COST Action QUALINET significantly advanced a common understanding of the above-mentioned concepts in a whitepaper [9]. Based on the work of Jekosch [38], the authors defined experience as follows:

Experience

An experience is an individual's stream of perception and interpretation of one or multiple events.

Whereas an *event* is defined as "an observable occurrence [...] which is [...] determined in space (i.e., where it occurs), time (i.e., when it occurs), and character (i.e., what can be observed)" and *perception* is defined as "a process which involves the recognition and interpretation of stimuli which register our senses (responses of sensory receptors and sense organs to environmental stimuli)." The authors further defined the term quality as [9]:

Quality

The outcome of an individual's *comparison and judgment process*. It includes perception, reflection about the perception, and the description of the outcome. In contrast to definitions that see quality as "qualitas", i.e., a set of inherent characteristics, we consider quality in terms of the evaluated excellence or goodness, of the degree of need fulfillment, and in terms of a "quality event".

To be highlighted in this definition is that the quality judgment process, for which a slightly modified version of a well-known judgement model based on Jekosch [38] and Raake [39] is illustrated in Fig. 2.1, includes an individual's comparison of a "perceived composition of an entity with respect to its desired composition" [38]. Thereby, the perceived composition refers to the "totality of features of an entity," and the desired composition to the "totality of features of individual expectations and/or relevant demands and/or social requirements" [38]. Consequently, and that is what makes QoE a strongly subjective and multi-faceted concept, these so-called quality features are key components for a quality judgement as different features might be considered and weighted by different users. A quality feature, for which immersion or the smoothness of the video would be an example in the context of cloud gaming, is defined as [38]:

Quality Feature

A recognizable and nameable characteristic of an entity that is relevant to the entity's quality.

In recent research about identifying relevant components of quality, so-called perceptual (quality) dimensions were investigated [40],[41], which represent orthogonal quality features. Thus, QoE is often described as a multidimensional concept. Additionally, as illustrated in Fig. 2.1, the quality judgement process is affected by several types of influencing factors (e.g., human, system/content, and context factors), also called quality elements. These elements are often objective and instrumentally measurable parameters of the system or transmission channel for multimedia applications. A quality element, for which a network delay or a characteristic of a game would be an example in the context of cloud gaming, is defined as [38],[35]:

Quality Element

A contribution to the quality of a material or immaterial product... in one of the planning, execution or usage phases.

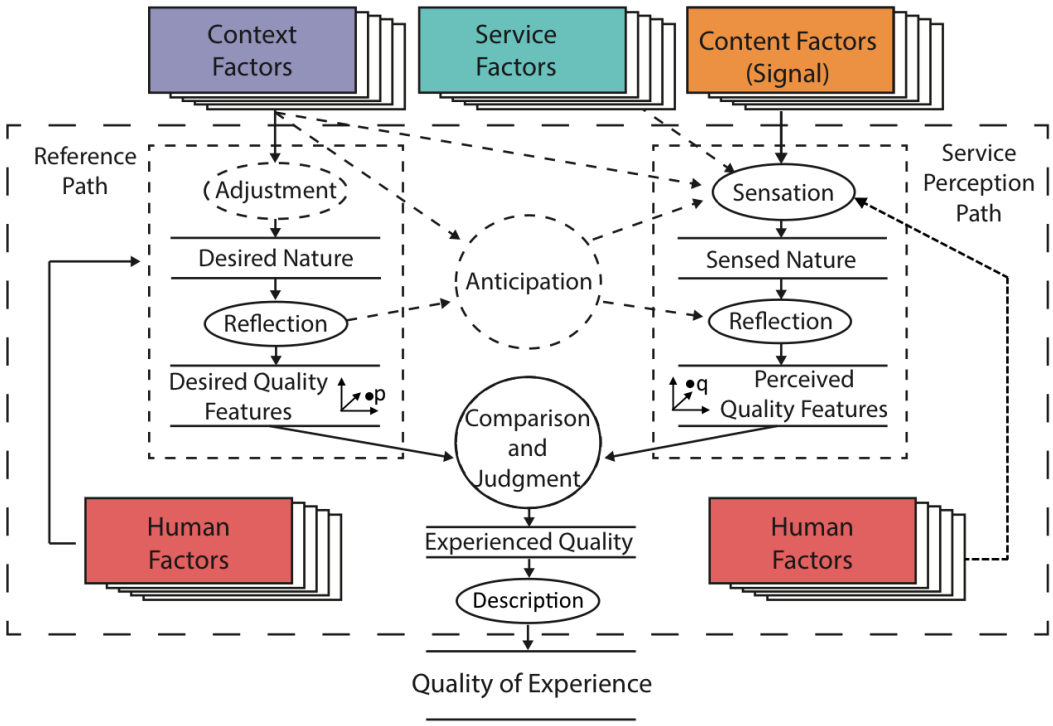


Figure 2.1: Quality judgement process derived from Jekosch [38] and Raake [39] with context and system influencing factors split as separate factors based on Hines [42] and quality features based on Côté [43].

To deeper understand the comparison and judgement importance of the illustrated process, the separated paths, i.e., the reference and service perception path, must be further examined.

The starting point for the service (quality) perception path is a physical event or signal, which consists of or is influenced by several service or content factors. The signal reaches the sensory organs, which will be processed through a low-level perceptual procedure into a sensed nature of the signal mediated by context factors such as light conditions or a viewing-distance or human factors such as visual capabilities. This is followed by a reflection process using cognitive processing to derive "nameable" perceived quality features q based on relevant signal characteristics. These quality features form a so-called quality feature space. Next, these perceived features are compared to the desired quality features p resulting from the reference path, which are strongly influenced by human factors

such as the user's current emotional state, motivation, or expectations. To compare the points q and p, two possible approaches are commonly considered. In the vector model approach, the influence of each quality feature on the overall quality is described by a linear combination of all features, where the overall quality is defined according to [44] as the "perceived quality of the system that is judged upon the totality of quality features that the user considers for the judgment." Thus, the better, i.e., higher, the quality features, the better the quality is, or the reverse. Alternatively, the point p can be considered as an ideal point, and the overall quality is described by the Euclidean distance between q and p. The relation between the overall quality and its underlying perceptual quality features is also called a quality profile [41].

After the judgement, i.e., the aggregation of all features into a single quality value [43], based on the positioning of the points in the quality feature space, the experienced quality, also named quality of experiencing in [35], is derived during the process of experiencing in the light of the person's context, personality and current state [9].

Finally, the user describes the experienced quality, e.g., by means of a questionnaire. Here, the question of how one can measure and quantify quality is a central element of the presented research. The outcome of the rating process is a judgement of the QoE, which is defined in [9], and partially adopted in the ITU-T Rec. P.10 [45] as:

Quality of Experience (QoE)

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.

When referring to "gaming QoE" in the remainder of the thesis, this considers the general definition of QoE but with quality elements and features that are important in the context of gaming, more specifically cloud gaming.

A second, arguably even more popular research domain is user experience (UX), which has its origin in Human-Computer Interaction (HCI). UX research has historically developed from usability research, which often aims at enhancing the efficiency and effectiveness of a system [46]–[48], but was initially focusing on the prevention of negative emotions when using a technology [46], [49]. For the latter, pragmatic aspects of the investigated systems have been identified in usability research. These days, UX often targets the understanding of HCI as a particular emotional experience (e.g., pleasure). The pragmatic aspects are only considered as requirements for positive experiences but not as direct contributors to those [50]. Similarly, the concept of positive or hedonic psychology (cf. [51]) has been used in HCI and UX research [46]. Consequently, the related research community has mainly focused on the hedonic aspects of experiences as described by Dieffenbach et al. [52] and as critically outlined by Mekler et al. [53].

With respect to the assessment of user experiences, the devotion towards hedonic psychology also comes with the need for measuring emotional responses (or experiential qualities) [54]. In contrast to QoE, where the assessment of the experienced (media) quality of a multimedia system is in focus, the assessment of experiential qualities in UX calls for the assessment of a range of additional aspects; e.g., Bargas et al. list affect, emotion, fun, aesthetics, hedonic and flow as qualities that are assessed in the context of UX [55]. Hence, this assessment approach often considers a substantially broader range of quantified qualities or experience dimensions. However, the UX domain has currently developed

towards design-based UX research that steers away from quantitatively measurable qualities and mostly focuses on qualitative research approaches. This trend has marginalized the assessment or model-based UX research camp in recent UX developments, as denoted by Law [56]. Therefore, it lacks novel approaches for quantitative, multidimensional assessment approaches.

In the context of the present thesis, certain aspects of these assessment-based UX approaches will be merged with existing QoE approaches. However, the latter will be the main focus of the presented research as it is primarily technology-driven and technology-centered, focuses on the quality formation, and finally enables the optimization of technical parameters at different layers (cf. [35]). Inspired by the work of Hammer et al. in [46], one could also argue the research targets the assessment of the subjectively perceived Quality of the User eXperience (QUX) or even Quality of the Player eXperience (QoPX) of cloud gaming services.

2.2 An Introduction to Cloud Gaming

To comprehensively understand the presented work of this thesis, the research of QoE must be set into the context of cloud gaming. Thus, an introduction to the concept of cloud gaming from a technical perspective is given, which by no means aims to be complete in every detail. However, the complexity of the service and with that also consequences for the list of possible system influencing factors on gaming QoE should be illustrated sufficiently.

2.2.1 Components of a Cloud Gaming System

Before going into the cloud gaming architecture itself, it is helpful to know how the executed game application can be defined. Juul proposed the following definition, which is also embedded in the ITU-T Rec. G.1032:

Game

A game is a rule-based system with a variable and quantifiable outcome, where different outcomes are assigned different values, the player exerts effort in order to influence the outcome, the player feels emotionally attached to the outcome, and the consequences of the activity are optional and negotiable [57].

Consequently, a digital game is substantially different from productivity-oriented applications. For the use of a speech dialog system or a navigation app, the focus is on completing a task without a high mental effort, challenges, or the need to learn. However, the goal of a game is to provide enjoyment to players. While the motivations, and with those ways to achieve enjoyment, are plentiful, a gaming session would not result in enjoyment without challenges in beating obstacles and learning about the game. Lastly, contrary to productivity-oriented applications, users of game systems are (usually) intrinsically motivated and rewarded [58], and are often highly immersed in their activity. Thus, the traditional view of usability, which is composed of efficiency, effectiveness, and satisfaction, does not apply to a game. While the game itself is a fundamental part of a cloud gaming system, the whole system is very complex from a technical point-of-view. It provokes many multi-disciplinary challenges in the areas of resource allocation, distributed architectures, data compression, adaptive transmission [59]. An abstract version of a typical framework of a cloud gaming service is illustrated in Fig. 2.2. In

line with this framework is the definition of cloud gaming based on the ITU-T Rec. G.1032 [60] stating that:

Cloud Gaming

Cloud gaming is characterized by game content delivered from a server to a client as a video stream with game controls sent from the client to the server. The execution of the game logic, rendering of the virtual scene, and video encoding is performed at the server, while the client is responsible for video decoding and capturing of client input.

Cloud gaming - also called Gaming as a Service or Games-On-Demand - combines the successful concepts of Cloud Computing and Online Gaming [61]. In sum, the key concept behind cloud gaming is that the game application, including the game logic, storage as well as the rendering of the game scene, are executed on a cloud gaming server instead of a player's client device, such as for traditional gaming. The client, also named thin-client, therefore does not require high processing power and only serves as an interface to the user. It must be noted that, additionally, communication over a network of the cloud gaming server to an online game server is possible if the executed game does not run encapsulated locally on the cloud server but instead require the exchange of game states with the online game server. This very often is the case for multiplayer games.

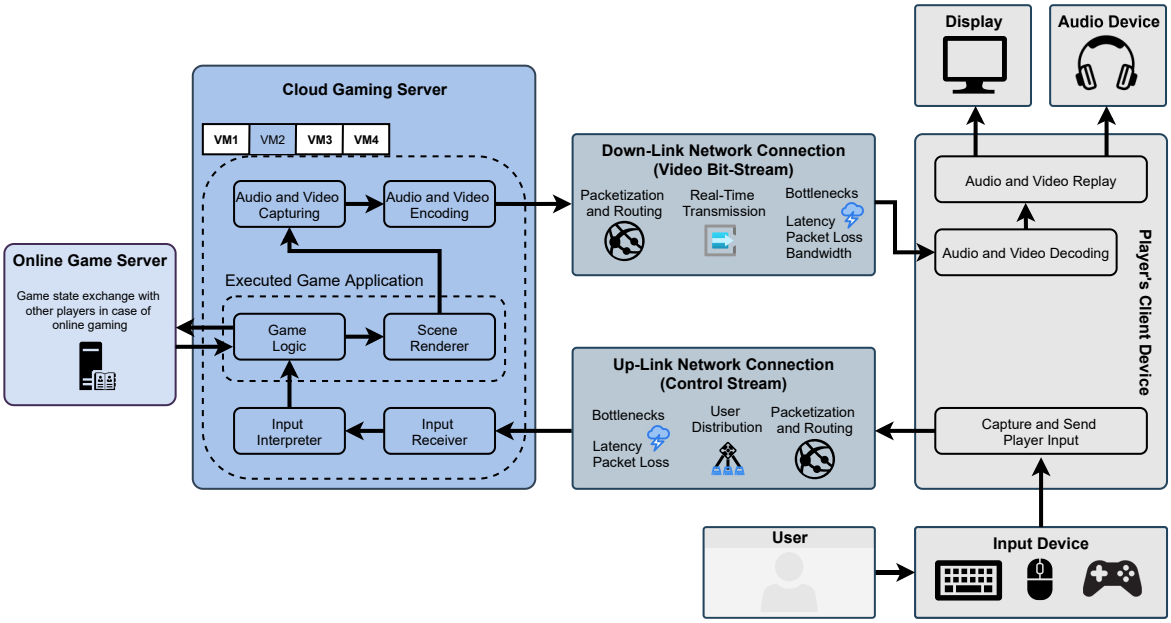


Figure 2.2: Abstract framework of a cloud gaming system based on [8] and [62].

To allow this concept to function, two data streams are necessary: a control stream including the input events of the player transmitted from the client to the cloud server (up-link), as well as a video bit-stream including the rendered game scene, which is transmitted from the cloud server to the client (downlink). Especially the latter adds a major challenge to the system, as the significantly larger packets containing the game scene (compared to input events) [63] must be encoded and transmitted in real-time to allow a smooth interaction with the game. Compared to the size of control event packets, the size of video packets is significantly larger and thus provokes a high load on the used network. Current services such as Stadia recommend an available bandwidth of at least 10 Mbps and up to 35 Mbps to

stream in 4K resolution at 60 Frames per second (fps). Especially in the case of multi-user households, this can lead quickly to bandwidth issues.

For the video encoding, typically hardware-accelerated encoding is used which reduces the memory read time and thus the overall system delay. There are numerous codecs with different compression efficiencies and encoding speeds. Within this thesis, the NVEnc codec, a hardware-accelerated version of the H.264/MPEG-4 AVC codec, is used. However, new encoders already advance, such as High Efficiency Video Coding (HEVC), also known as H.265, and the open source codec AV1, which is considered the successor to Google's VP9 codec. At the time of the presented research, those were not available or resulted in significantly higher encoding times as shown in [64].

Another major challenge is the added network connections, which are prone to packet loss and cause a delay due to the various processing steps. These do not only include the transmission of packets *per se* but also processes such as segmentation, fragmentation, and encryption but also the routing possibly leading to network congestion and forward error correction. While for a real-time stream, the User Datagram Protocol (UDP) would be a preferred protocol over the Transmission Control Protocol (TCP) due to the absence of packet retransmission, various network protocols are possible for cloud gaming. Parsec's cloud gaming system, for example, is using a peer-to-peer networking protocol called Better User Datagrams, which is a reliable UDP approach with custom congestion control algorithm to achieve low-latency video delivery[1]. As network control protocol, servers often use the Real Time Streaming Protocol (RTSP) together with the Real-time Transport Protocol (RTP) in conjunction with Real-time Control Protocol (RTCP). While RTP performs the data streaming, RTCP is used for quality control and synchronization of streams. Lately, also Web Real Time Communication (WebRTC) used in Google's Stadia service [65], is becoming more popular for handling the data flows between the server and the client.

Huang et al. [66] categorized the components causing a reduced responsiveness of a cloud gaming system. The authors define a response delay (RD) as the time difference between a user submitting a command and the corresponding in-game action appearing on the screen, which is further divided into a processing delay (PD), a playout delay (OD), and to a network delay (ND), also known as Round-trip time (RTT). PD, i.e., time for the server to receive and process a player's command, can be even further divided into (1) memory copy of raw game images, (2) format (color-space) conversion, (3) video encoding, and (4) packetization. OD, i.e., the time required for the client to receive, consists of: (1) frame buffering for all packets belonging to the current frame, (2) video decoding, and (3) screen rendering, i.e., displaying the decoded frame. It should be noted that when referring to delay in this thesis, an additional (symmetric) network RTT is considered. Furthermore, with regards to the resulting waiting times, it must be considered that a cloud gaming system is a discrete system in which components such as the game logic and video processing run at fixed (but different), input-independent rates. A user's input is only included if it is also available.

The placement of servers also plays a crucial role for the system performance as long distances to the end user cause high propagation delays. As a server, services such as Amazon Elastic Compute Cloud (EC2) provided by Amazon Web Services (AWS) would be one of many options. Also adding a small number of edge servers as part of a content distribution network (CDN) is a lately investigated approach, which promises reduced latency [67] and more reliable data transmission.

[1] https://parsecgaming.com/game-streaming-technology/

Lastly, it must also be mentioned that on the cloud server multiple Virtual machines (VM) are used to allow a more resource-friendly execution of the service. In a traditional gaming system, a single Graphics Processing Unit (GPU) is used for a fast creation of graphics. The GPU pass-through technique allows the VM to access a physical GPU card to exploit the hardware's acceleration features and thus to handle GPU-intensive applications such as games [68]. Additionally, a GPU can be virtually divided for multiple users, and multiple GPU cores allow a high scalability of the GPU powered VMs. An example of such a technology is NVIDIA virtual GPUs (vGPUs) software[2]. More information about the architecture of cloud gaming services can be found in [8], [59], [65], [69].

2.2.2 Influencing Factors on Gaming QoE

The previous chapter described the complexity of cloud gaming services from a technical perspective. Knowledge about these technical parameters is very important for the assessment of gaming QoE as they either represent parameters under test or must be controlled in order to achieve reliable and valid experimental results. Furthermore, they are important for building instrumental quality prediction models. However, as illustrated in Section 2.1, gaming QoE is subject of various other influencing factors, which will be addressed in this section. According to [9], an influencing factor is defined as:

Influencing 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 little information existed on influencing factors of gaming QoE and their interactions, ITU-T SG12 decided to start the work item called G.QoE-gaming during the 2013-2016 Study Period. Members of the Qualinet group, more specifically its Gaming QoE Task Force, identified several factors that have shown an impact on Gaming QoE based on reports in the literature, as well as possible other candidates. This effort resulted in a first ITU-T contribution for the G.QoE-gaming work item [70]. In line with the categorization from Reiter et al. in [35], the following factors were distinguished: user (or human) factors, system factors, and context factors. As shown in Fig. 2.1, also a split of system and content factors, or game factors in the case of cloud gaming, is possible. Based on two additional contributions by Schmidt et al. [19] and Zadtootaghaj et al. [20], the work item G.QoE-gaming was finalized in 2018 and resulted in the ITU-T Rec. G.1032 [60]. In Fig. 2.3 an overview of the identified influencing factors is provided.

For many of these factors, a detailed description can be found in [12], [35], [60], [71]. As an in-depth summary of all possible influencing factors and their state-of-the-art would be beyond the frame of this thesis, in the following, the focus will be on those factors which are most relevant to the presented research. To begin with, it must be mentioned that influencing factors must not be regarded as isolated as they are often interwoven with each other [11]. A characteristic of a game maybe results in an increase of required encoding bitrate to avoid video quality artefacts, or the impact of strong audio compression might only become perceivable if a game contains narrative audio elements. Thus, very often, some influencing factors do not impact a gaming QoE on their own but have a mediating influence on network or encoding impairments. As a combination of all possible factors, which is to some extent desirable to create a quality prediction model, would lead to an unfeasible amount of test

[2]https://www.nvidia.com/en-us/design-visualization/solutions/multi-virtual-gpus/

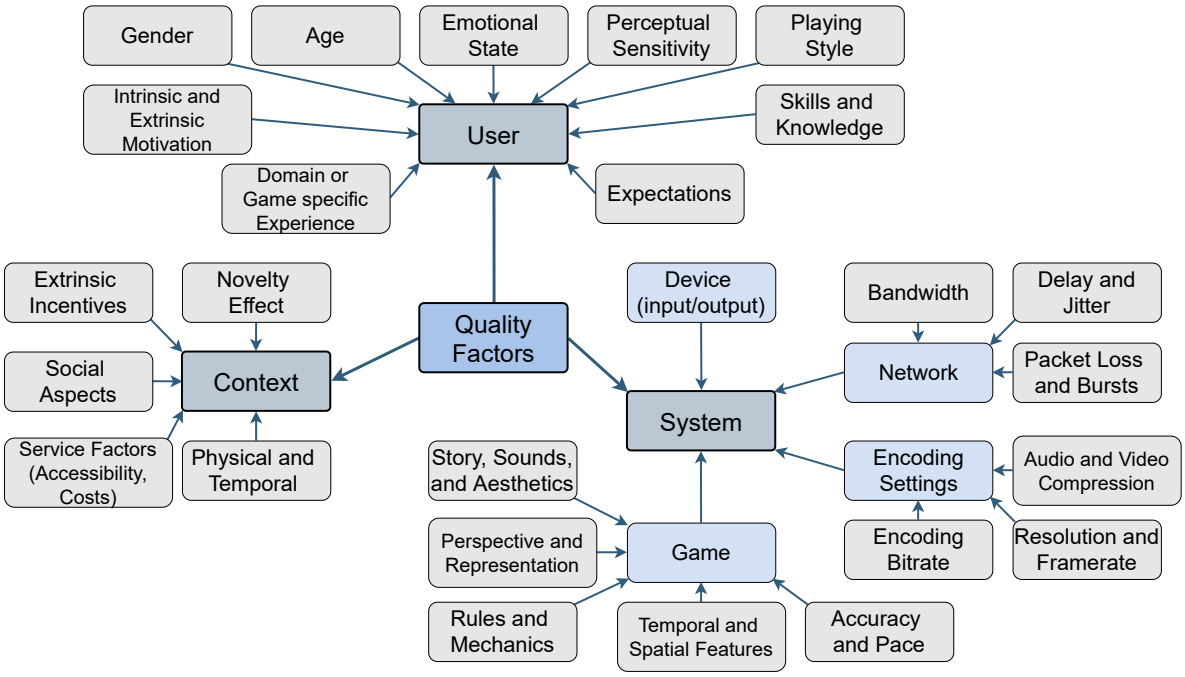


Figure 2.3: Overview of quality factors on gaming QoE.

parameters, research often focuses on one independent factor to investigate its impact of gaming QoE to potentially narrow down the necessary parameter space. This makes research about influencing factors very important.

With respect to user factors, i.e., any variant or invariant property or characteristic of a human user describing the demographic and socio-economic background, the physical and mental constitution, or the user's emotional state [9], it is important to consider the quality judgement process (cf. Section 2.1).

On the one hand, experiences with gaming in general or with a specific game do not only result in skills and knowledge for or about the used cloud gaming system, but they can also influence the expectations, i.e., "the desired composition of an entity." Two of the most known classifications of users are to distinguish between "hardcore" and "casual" gamers depending on the average time of playing per time period, and to group them into "newbie" and "pro" gamer (or experts) based on the experience with a particular game or type of game, which in many cases is also linked to a player's skills. Slivar et al. showed that there is a significant impact of players' previous gaming experience, as experienced players (based on self report) generally gave lower overall quality ratings for conditions with a very low encoding bitrate, and higher ratings for less impaired conditions, as opposed to novice players [72]. These findings were also confirmed in a study about traditional online gaming in [73] where the gaming QoE was assessed for various network conditions.

On the other hand, the abilities of a user, in terms of perceptual sensitivity (e.g., visual and auditory acuity), to "perceive the composition of an entity" is also important for the quality judgement process. While these abilities might be related to the expertise and knowledge of a player, also the age or even the emotional state, e.g., due to being so excited to play that impairments in a video are not noticed, could influence them. While studies in [74] could not show a direct impact of age and gender on player's state of immersion, these factors may influence other factors, such as expectations (for users having previously used a certain technology) or motivation. While the latter is arguably of high interest for research about the engagement of players to games, to the best of the author's knowledge no studies are available showing a mediating effect of the motivation for playing on the gaming QoE for cloud gaming

services. Thus, this factor was not considered for the scope of the thesis. Vermeulen et al. investigated whether there is a mediating effect of previous experience with certain game genres on game design preferences of male and female players [75]. The authors found that women are generally more annoyed by violence and less attracted to complex gameplay than men, as well as that female players who do not play core genres such as fighting, action-adventure, sports, survival horror, racing, role-playing, strategy, or massively multiplayer online (MMO) have generally different game preferences.

Only limited knowledge is currently available about context influencing factors, i.e., a factor that embrace any situational property to describe the user's environment [9]. This might be caused by the circumstance that many of these factors, such as service costs or novelty effects, cannot be investigated well in a laboratory study. Thus, they will not be in the focus of the present thesis but should be controlled to carry out subjective user studies, e.g., keeping the light and noise conditions of test room constant. Regarding social aspects, Vermeulen et al. were able to demonstrate in [76] that female players perceive more stress and rate their own gaming skills lower when assuming to play against a male opponent. Suznjevic et al. also showed in their study presented in [73] an impact of players' social context on gaming QoE. Novice players reported higher QoE ratings in case of playing together in a team with higher skilled players than in a homogeneous group. To be highlighted is a work presented by Beyer et al. in [77]. The authors investigated whether participants playing the same games in a simulated metro environment compared to a quiet laboratory room would influence the gaming QoE ratings. However, no significant effect was found. This is a first indication that the physical environment is not a dominant influencing factors, as potentially the involvement in the playing and rating tasks dominate the environment impact. In Chapter 4, further research on the comparison of physical environments will be presented.

Lastly, some dominant system influencing factors, i.e., properties and characteristics that determine the technically produced quality of an application or service [9] should be summarized. A great amount of research was been carried out for traditional online gaming but also recently for cloud gaming to investigate the influence of network and encoding settings. Thereby, the end-to-end delay was identified as one of the most dominant influencing factors on gaming QoE [59], [78]. Jarschel et al. [61], [79] evaluated user perceived QoE and identified delay as one of the key influencing factors for cloud gaming QoE. The authors showed that participants' user experience was not influenced by a delay of 80 ms for a role play game and for a soccer game but a degradation was present for a fast-paced racing game. Also, a study by Sackl et al. [80] showed an influence of delay for a platform and racing game. The authors additionally showed differences in users' ability to detect latency. While generally, participants able to perceive delay reported lower gaming QoE, surprisingly, also some quality reduction was reported by players who were unable to detect the delay for a racing game. Raaen et al. [81] investigated the sensitivity of the human perceptual system when encountering motor-visual delays in a user study in which participants had to quantify the smallest noticeable response delay. The results showed that about 25 % of the participants were able to perceive a delay of less than 40 ms, whereas half of the players perceived delays shorter than 100 ms. In addition to the gaming experience, a delay can also degrade a player's gaming performance as shown by Sabet et al. in [82] who showed a mediating effect of gaming performance on the impact of delay on gaming QoE, i.e., if the performance remains the same, QoE degradation is lower than when the performance is reduced. Clincy and Wilgor showed that a delay could also mediate the effect of packet losses and revealed a very strong impact of packet loss at rates of 1 % on player's experience, making a game nearly unplayable [83]. Jarschel et al. [79] demonstrated

that the transmission direction, i.e., downlink or uplink, in which packet loss occurs, is more important in cloud gaming than in conventional gaming, as the video stream is influenced significantly stronger than the control stream. Slivar et al. showed that the influence of delay and packet loss is much stronger for cloud gaming compared to traditional online gaming, and that there is an approximately linear relationship between packet loss and QoE [84].

Hong et al. revealed in their research that at a fixed encoding frame rate, a higher bit rate always leads to higher quality. However, at a fixed bitrate, increasing the frame rate does not necessarily improve the gaming QoE [85]. Slivar et al. [72], [86] also investigated the influence of encoding parameters on gaming QoE. The authors conducted two experiments, one with levels of frame rate of 30 fps and lower, and another test with frame rate conditions of 25 fps and higher, but both with the same bit rate conditions (3, 5, 10 Mbps). The results showed, in line with Hong's finding, that when playing a complex game (e.g., Serious Sam 3), at a low level of bit rate (e.g., 3 Mbps), increasing the frame rate from 25 fps to 60 fps reduces the overall quality ratings. As a cloud gaming system using a higher framerate, or also a higher resolution, requires more data throughput, a too low bitrate would result in strong spacial video artefacts, i.e., blockiness, while a too low framerate would reduce the smoothness of the video stream. Such a trade-off between these parameters is a typical example of the need for a proper resource allocation of a cloud gaming service. Slivar et al. also showed that in case of an adequate level of bit rate, there are no significant differences in quality ratings for 35, 45 and 60 fps. Claypool et al. inspected the effect of frame rate on overall quality and player performance for first-person shooting games. Five levels of framerate, 3, 7, 15, 30, and 60 fps, were selected as independent variable. An analysis showed a significant difference between the five levels of frame rate for performance as well as quality ratings. However, no significant difference was found between 30 and 60 fps [87], [88]. Beyer et al. [89] showed in their study using a first-person shooter game in a cloud gaming setup, that a very bad video quality caused by a low encoding bitrate of 1 Mbps at a framerate of 50 fps influenced the perceived gaming QoE, player experience (PX) aspects such as flow, immersion, and competence, valence, and the alpha frequency band power of test participants. A more detailed overview of encoding related research is provided in [72].

With respect to device characteristics, Beyer et al. investigated the gaming QoE for four different screen sizes, namely 3.27", 5", 7", and 10.1". The results show that there is an acceptance threshold somewhere between 3.27" and 5". If a screen size is larger than this threshold, gaming QoE does not further increase significantly within the range of tested screen sizes [71], [77]. Huang et al. performed cloud gaming experiments using both mobile and desktop clients. The authors revealed that players are more satisfied with the graphics quality, defined by the authors as the visual quality of the game screen, on mobile devices and the control quality, defined by the authors as the quality of the control mechanism, on desktops [90]. Furthermore, in recent years one can witness many advancements with respect to output devices leading to steadily increased display resolutions, higher and adaptive (e.g., Nvidia's G-Sync and AMD's FreeSync) refresh rates of displays, and also better color representation through HDR monitors, enabling new rendering technique such as ray tracing to its full potential.

Throughout the majority of studies summarized before [73], [79], [83], [91], the executed game was identified as one of the most dominant influencing factors mediating the effect of other factors on gaming QoE. However, a game classification targeting these effects is yet still missing. This makes the generalizability of many conclusions drawn from research very difficult. To address this issue, more research regarding this matter is presented in Chapter 6.

While the presented overview of influencing factors is by no means complete, even when including the factors summarized in the ITU-T Rec. G.1032, it covers the majority of factors that can be investigated or should be controlled when evaluating an existing cloud gaming system. However, there are many other elements in this complex system that are not considered, such as client capabilities with respect to the CPU, GPU, and memory, cheating protections, buffer sizes, congestion controls, concealment and error correction algorithms, [92] and more. Also, the possible connection to the online game server poses many problems. Metzger et al. showed that a) the framerate has a larger impact on the end-to-end delay than the server tick rate, i.e., the rate with which game states are updated, and b) that only if both rates are high enough, the network delay will have a dominant influence, i.e., there is a masking effect of these rates on the delay [93]. Also, audio quality-related aspects were in the context of cloud gaming so far not subjected to the same degree of scientific scrutiny as video compression [60]. Lastly, new challenges invoked by virtual reality gaming using head-mounted displays (HMDs), as well as other interaction modalities such as tactile or kinaesthetic feedback, smell, and other sensory experiences, are not yet addressed [11].

2.3 Gaming QoE Taxonomy for Cloud Gaming Services

After the complexity of a cloud gaming system and the broad range of influencing factors was described in the previous sections, in the following the multi-dimensionality of the construct gaming QoE will be examined. Concretely, a multi-layered gaming QoE taxonomy offering an overview of currently available and relevant concepts will be presented. This section is mostly taken from the initial publication of the taxonomy in [12] whereby, in addition, some adaptations based on the ITU-T Rec. P.809 are included. The taxonomy, which was inspired by a taxonomy about multimodal human-machine interaction developed in [94], forms the theoretical foundation of the present thesis and the related research regarding the ITU-T gaming activities.

To follow a common terminology, to begin with, the terms quality factor and quality aspects should be explained. The terms are linked to the quality judgement process described in Section 2.1 and were introduced by Möller in his research about the quality of transmitted speech [95].

Quality Factor

Individual categories of properties of the service that are relevant to its quality. They each comprise one or more quality elements.

Quality Aspects

Individual categories of the quality of the service under investigation. They each include one or more quality features.

The gaming QoE taxonomy, which is depicted in Fig. 2.4, is composed of multiple layers. Taking the perspective of the technical system (top layer) and of the human user (bottom layer), quality elements and quality features are grouped together to quality factors and quality aspects, which in turn integrate to QoS and QoE (overall quality), respectively. Those of the factors and features which are relevant for gaming are put into a logical relationship to highlight their dependencies. At the top of the taxonomy, one can find the classification of quality factors into user, system, and context factors, which were already explained in detail in the previous section. It must be noted that these factors do not only

influence the QoS but also various QoE aspects. Strongly linked to the QoS, the taxonomy considers the *interaction performance* in the next layer.

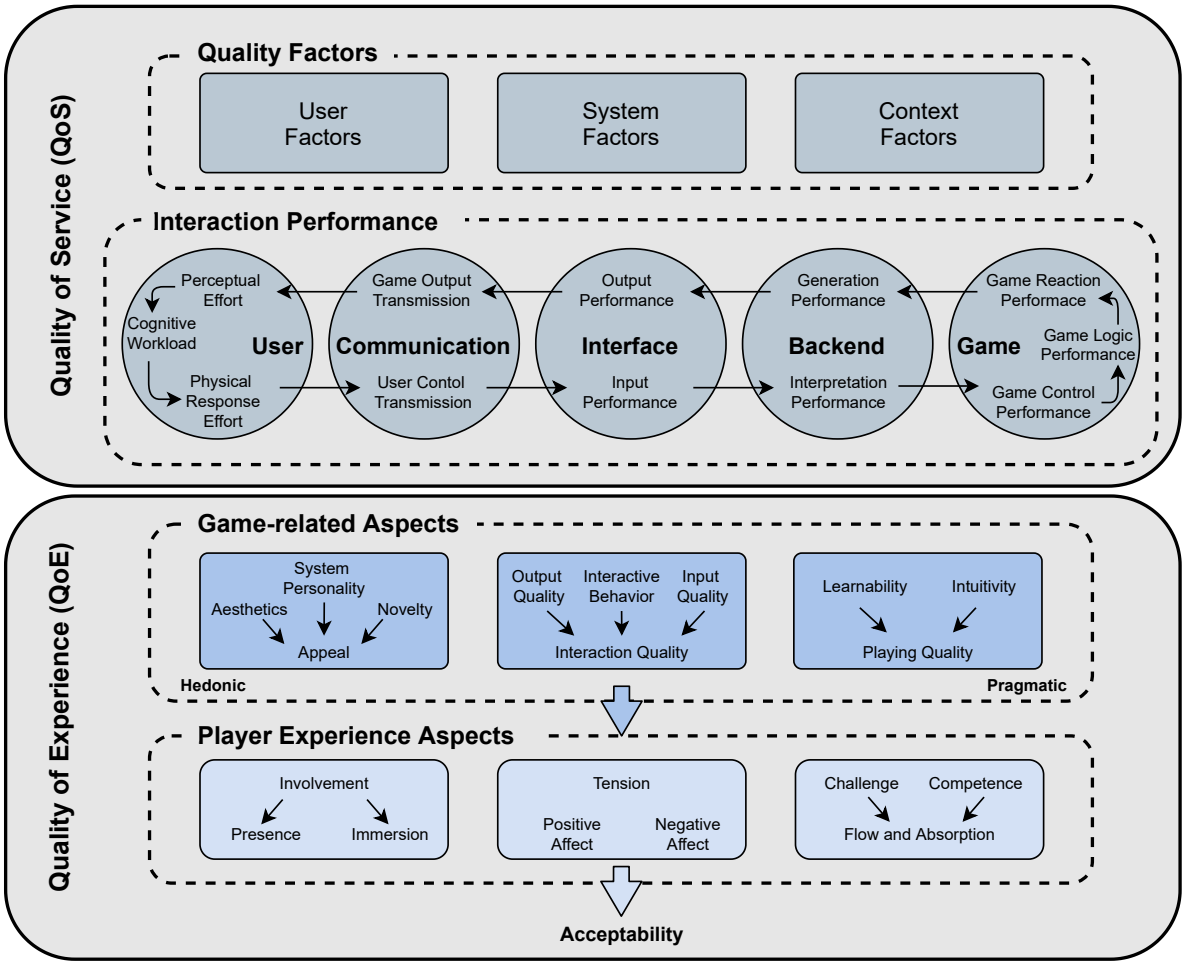


Figure 2.4: Taxonomy of gaming QoE factors and aspects adapted but based on [12] and [96].

The interaction performance, describing the behavior of and processes within the user and the system during the interaction, is separated into *user performance* as well as *system performance*. The latter can be further divided into performance aspects of the user interface (device and software), performance aspects of the backend platform, and performance aspects of the game, i.e., the story behind the interaction. Between these blocks, there may be communication channels: Physical channel between the user and the user interface; IP-based channels between the user interface and the backend platform (e.g., in cloud gaming), and between the platform and the game (e.g., in multi-player games where the game of one user is influenced by another user and this information is exchanged on a game-level).

System Performance:

- User interface performance: includes the input and the output performance of the user interface.
- Backend platform performance: can be subdivided e.g., into the performance of the interpretation of user input by the platform and the performance of generating corresponding output.
- Game performance: mostly influenced by the control the user has over the interaction in the game, the game rules, and the game reaction. Can be expressed e.g., in terms of game success, time-on-game, game errors, or alike.

- Communication channel performance: includes all aspects of any transmission channels involved, i.e., the effectiveness and efficiency of forwarding user controls to the game and the performance of forwarding game output to the user.

User Performance:

- Perceptual effort: the effort required decoding the system feedback, as well as understanding and interpreting its meaning.
- Cognitive Workload: commonly specifies the costs of task performance (e.g., necessary information processing capacity and resources). As gaming does mostly no have a “task”, obtain an “outcome” is considered as a gaming task. It should be noted that it is actually the task of the game to put a certain load on the user; thus, keeping the load low does not always result in a good gaming experience.
- Physical response effort: Physical effort required to interact with the game. It may largely be influenced by the device used for the interaction.

The bottom layer representing the QoE is subdivided into *game-related aspects* and *player experience aspects* which will be described in the next sections. A similar split can also be found in the work by Abelle et al. inc[97]. The authors made use of the Means-End theory for the development of their Player Experience Inventory (PXI) tool which aims to provide insight into how specific game design choices are experienced by players. The PXI distinguish between *functional consequences*, i.e., the immediate experiences as a direct result of game design choices, such as audiovisual appeal or ease-of-control, and how these lead to specific emotional responses called *psychosocial consequences*, i.e., the second-order emotional experiences, such as immersion or mastery [97]. The Means-End theory has its origin in the consumer psychology domain [98] pointing out that consumers choose a product not only based on specific attributes ("means") but based on benefits or desired "consequences" aligned with personal values ("ends"). Interestingly, Abelle et al. also illustrated that the relationship between functional consequences and game enjoyment is mediated by psychosocial consequences [97].

However, the gaming QoE taxonomy does not only target game enjoyment but also the overall perceived QoE, which can influence the acceptability of a cloud gaming service. Following the general definition, acceptability describes how readily a user will actually use the system. Acceptability may be represented by a purely economic measure, relating the number of potential users to the quantity of the target group. Acceptability is influenced by PX, but also by other factors such as costs, accessibility, service conditions, etc. [12].

2.3.1 Game-related Quality Aspects

The game-related aspects are categorized into aesthetic aspects, interaction quality aspects, as well as playing quality aspects, which will be described in more detail below.

Interaction Quality: When looking into the quality of computer games, a common quality feature is the playability of a game. However, there seems to be no agreement on the definition of playability. As an example, Sánchez et al. [99] define playability as “the set of properties to describe the player’s experience with a particular game system, that the principal goal is fun/entertainment to the player in a satisfactory and credible way, playing alone or with other players. Playability reflects the player’s pleasure, experience, sensations, and feelings when he/she is playing the videogame”. Along a similar track, playability is defined as “the degree to which a game is fun to play and usable, with an emphasis

on the interaction style and plot-quality of the game; the quality of gameplay. Playability is affected by the quality of the storyline, responsiveness, pace, usability, customizability, control, intensity of interaction, intricacy, and strategy, as well as the degree of realism and the quality of graphics and sound" by the Foraker Labs[3]. In contrast to this, Engl [100] defines playability as the degree to which all functional and structural elements of a game (hardware and software) enable a positive PX for the player. This definition considers playability as a prerequisite of positive PX (like usability can be considered as a prerequisite of user satisfaction, see definitions of usability in [95]), or as a technical and structural basis for this, but not the PX itself. For the gaming QoE taxonomy and present thesis, the latter narrower definition was or will be adopted, and is called "interaction quality" of a game for the remainder of this work. This meaning is in line with the general definition of this term for multimodal interactive systems and includes input quality (player to system), output quality (system to player, e.g., in terms of graphics quality, video quality, audio quality), as well as the interactive behavior (in task-oriented the term "cooperativity" is typically used, but as the game storyline is not designed to be cooperative to the user, the general term interactive behavior is preferred at this point).

Playing Quality: Playing quality, which considers pragmatic aspects of a gaming service, can be considered as a kind of game usability. Rajanen and Tapani investigated the views and practices of 50 North American game companies regarding game usability, and identified the following aspects related to it: intuitiveness, immersiveness, minimal frustration, logic, transparent interface, understandability, learnability, memorability, and efficiency [101]. Furthermore, Pinelle et al. defined game usability in [102] as "the degree to which a player is able to learn, control, and understand a game. [...] Game usability does not address issues of entertainment, engagement, and storyline, which are strongly tied to both artistic issues (e.g., voice acting, writing, music, and artwork) and technical issues (graphic and audio quality, performance issues)." However, the definition of usability is based on effectiveness and efficiency. These concepts are more difficult to define for a game (it is actually the game's task to make the user spend resources). As a consequence, the term "playing quality" over "game usability" is preferred here and specified by the sub-aspects learnability and intuitivity (or intuitive controls), leaving out effectiveness and efficiency of the taxonomy presented in [94].

Aesthetic Aspects: In line with general multimodal interaction [94], the aspects aesthetics, system personality, and appeal are considered and describe the hedonic aspects of a gaming service. Aesthetics is the sensory experience the system elicits, and the extent to which this experience fits individual goals and spirit [103]. The system's personality refers to the users' perception of the system characteristics originating from technical and game characteristics. The appeal is a result of the aesthetics of the product, its physical factors, and the extent to which the product inherits interesting, novel, and surprising features [104], [105].

2.3.2 Player Experience Aspects

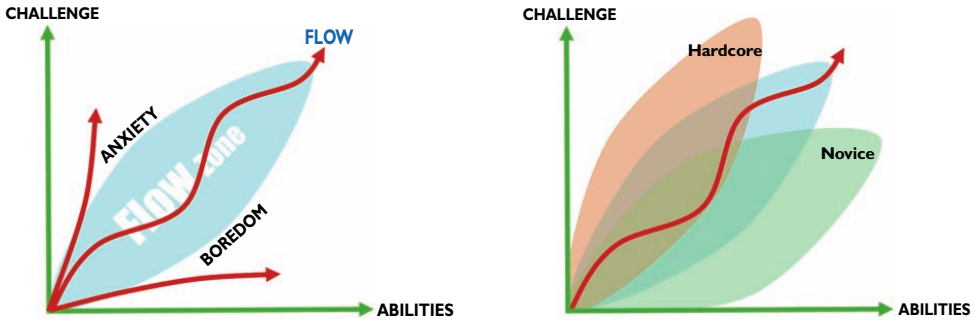
Player experience (PX) is a broad concept that covers a large set of sub-aspects. As mentioned before, the "degree of delight or annoyance of the user" is considered as a key aspect of QoE which should be reflected in PX as well. Poels et al. [106] defined a comprehensive taxonomy of PX which was adopted for the gaming QoE taxonomy. According to their definition, PX consists of the sub-aspects challenge, control, flow, tension, immersion, competence, positive affect, and negative affect.

[3] http://www.usabilityfirst.com/glossary/playability

In the following paragraphs, some insight into these concepts will be provided.

Positive and negative affect: Positive affect can come in many different forms, and it is usually the target of all gaming activities. According to Murphy [107], fun is "the positive feelings that occur before, during, and after a compelling flow experience. [...] It is not perfect, but it is concrete. The list of positive feelings associated with this definition of fun is quite long and includes: delight, engagement, enjoyment, cheer, pleasure, entertainment, satisfaction, happiness, fiero, control, and mastery of material"; negative ones might be frustration and boredom. Applied to computer games, Lazzaro [108] investigated emotions and classified them into four types of fun: Hard fun (linked e.g., to computer games; typical is a constant change between frustration and fiero), easy fun (linked e.g., to curiosity, mostly covered by immersion), serious fun (linked e.g., to relaxation from stress), and people fun (linked to social interaction). The fun types may be linked to the playing style user types from Bartle, e.g., an achiever mostly searches for hard fun, an explorer for easy fun, a socializer for people fun, and a killer for hard and people fun [109].

Flow, challenge, control: According to Csikszentmihalyi [110], flow is an equilibrium between boredom and fear, between requirements and abilities, and it is a dynamic experience of complete dissolution in the activity of an acting person. The activity itself constantly poses new challenges, so there is no time for boredom or sorrows. Intrinsic motivation is important for flow, as well as control over the game [111]. Hassenzahl [112] relates flow to user experience: "Briefly, flow is a positive experience caused by an optimal balance of challenges and skills in a goal-oriented environment. In other words, flow is the positive UX derived from fulfilling the need for competence (i.e., mastery); it is a particular experience stemming from the fulfillment of a particular be-goal". The concept was adopted to gaming by many researchers. Chen describes differing ideal zones of the flow phenomenon based on the capabilities and challenges for different user groups [113]. The flow channel, visualized in Fig. 2.5a, represents the diagonal line on the abilities-challenge plane, which can also fluctuate slightly due to difficulty changes for different game scenarios. However, if the challenge is too high, the flow experience is destroyed, leading to anxiety of losing, whereas a too easy game would end in boredom. In general, everybody can experience flow, but there seem to be factors that reduce flow in games, like age, reaction time, abilities, exposure to computers (digital natives vs. newbies), see e.g. [114]. Consequently, the zones might be different for experienced hardcore players, i.e a shift upwards on the challenge axis, in contrast to less experienced novice players as shown in Fig. 2.5b based on [113].



(a) Boundaries of flow zone. **(b)** Flow zones per player group.

Figure 2.5: Visualization of the flow zones based on Chen [113].

Absorption A concept closely related to flow is cognitive absorption which is a multidimensional construct describing a "State of deep involvement with software." It is based on three closely interrelated concepts: the personality trait of absorption, the state of flow, and the notion of cognitive engagement [115]. Being absorbed thereby refers to being in a state of deep attention with the event experienced [116]. The notion of cognitive engagement can be described by the three dimensions attention focus, curiosity, and interest [117]. Webster and Ho argue that absorption is "identical to flow, just without the dimension of control". Thus, one may be passively engaged, e.g., in watching TV, while experiencing "passive flow" is impossible" [116], [117].

Presence Presence is a psychological state of "being there" mediated by an environment that engages one's senses, captures attention, and fosters active involvement. The degree of presence experienced in this environment depends on the fidelity of its sensory components, the nature of the required interactions and tasks, the focus of the user's attention/concentration, and the ease with which the user adapts to the demands of the environment [118]. However, it must be noted that the concept of presence is mostly in the focus of virtual reality applications.

Immersion One of the most well-known concepts in the gaming domain, but also recently in other multi-media communities, is immersion. However, up to this point, the research community has not yet agreed upon a conclusive definition of "immersion" [119] and thus, immersion tends to be confused with presence or is considered synonymous [120]. However, it appears probable that immersion represents a multidimensional subjective phenomenon that includes other phenomena like involvement and engagement [121], place and plausibility illusions as well as social presence [122], [123]. A basic distinction can be made between system immersion, which comprises all objective elements of the multimedia system, and mental immersion, which comprises various aspects of subjective user experience that are affected by those system factors. Similar dichotomies were proposed between (system) immersion and presence by Slater [124]. Especially with a focus on games, immersion is used to describe the degree of involvement with a computer game and has been classified by Brown and Cairns into three phases as "engagement", "engrossment", and "total immersion" which build upon each other. The authors state that to enter the level of engagement, the player has to overcome the barrier of preferences, invest time as well as effort and have the attention to learn how to play the game. To enter the stage of engrossment, the player needs to combine game features and master the control of the game in order to become emotionally attached. While players in this state are less aware of their surroundings and themselves, they might reach a state of total immersion by overcoming the barriers of empathy and atmosphere. In total immersion, players described a sense of presence and of being cut off from reality to such an extent that the game was all that mattered [125], [74]. Inspired by the taxonomy of immersion presented in [119], for the present thesis when speaking of immersion, the term mental immersion (during gameplay) might be best suitable and in line with the PX model the gaming QoE taxonomy is based on. Mental immersion can be defined as follows:

Mental Immersion

Mental immersion describes a state of the user linked to mental absorption, mediated by the game content including narrative elements and challenges [126].

Relationships between engagement concepts

Some of the above-introduced concepts appear to be strongly overlapping and related to one another. Something they have in common is that they describe user engagement in a gaming experience. O'Brien

and Toms [127] have qualitatively defined engagement as “a value of user-experience that is dependent on numerous dimensions, comprising aesthetic appeal, novelty, usability of the system, the ability of the user to attend to and become involved in the experience and the user’s overall evaluation of the salience of the experience” [128]. Brockmyer et al. summarize that a "continuum of deepening engagement from presence to flow to absorption" may exist for some individual experiences. Jennett states that immersion can be seen as a precursor for flow, whereas flow describes an optimal and, therefore, extreme experience. A game could be considered to provide a highly immersive experience but it does not necessarily meet the requirements for perceiving flow [129]. Jennett further argues that "Immersion is an experience in time and that even though games with simple graphics such as Tetris do not involve presence (i.e., it is unlikely you will feel like you are in a world of falling blocks) they can still be immersive, leading to time loss, not noticing things around you.", [129]. Finally, cognitive absorption is seen as "an attitude towards information technology in general whereas immersion is the actual experience of a particular occasion of playing a videogame" [115], [129]. Thus, in the frame of the present work, the concepts of flow and immersion are considered, whereas presence and absorption appear to be less important in the context of cloud gaming services on a 2D-screen.

2.4 Summary

This chapter provided an introduction to the concept of quality followed by the research domains of QoS, QoE, and UX. Essential for evaluating cloud gaming services is the understanding that quality is the result of a comparison between desired and expected quality features by the user of a service leading to a judgement process. It was highlighted that considering QoS alone is not sufficient and that instead, QoE in combination with UX should be considered. Both concepts are part of the cloud gaming taxonomy introduced in this chapter that aims to explain which quality factors, i.e., influencing factors such as a network delay, impact various quality aspects, and how these aspects form the QoE judgement. Throughout the upcoming chapters, the taxonomy will be investigated with respect to its completeness and possible causal relationships between the quality aspects. Next, the components of a cloud gaming system were summarized. The high complexity of the service results in many quality factors that can be grouped into user, system, and context factors. A vital concept of a proper assessment of the gaming QoE is that the impact of these factors must be individually investigated while controlling for the influence of the remaining factors. Thus, knowledge about the importance of the factors is mandatory. While a summary of the state-of-the-art with this regard was provided, the quantitative and qualitative impact of various factors is still unknown. Many of these factors are difficult to categorize, as precise definitions and parameters are still missing. In Chapter 6, the present dissertation will contribute to this matter with a special focus on the game content as an important influencing factor. Finally, an overview of the quality features, also called quality aspects, which are considered to be relevant for the quality judgment process was provided. The summary showed that gaming QoE is a highly multi-dimensional construct composed of game-related factors such as the input and output quality, as well as player experience aspects such as flow and immersion. This makes a coherent measurement of all these quality aspects a highly challenging task. Providing methods for their assessment will be the topic of the next chapter.

Chapter 3

Methods for Assessing Gaming QoE

3.1 Overview of Assessment Methods

Traditional paradigms to evaluate the QoE of multimedia services, as well as the UX of interactive services, make use of subjective experiments. In these experiments, participants are confronted with the service to be evaluated, e.g., in terms of media stimuli for media delivery services, or in terms of prototypes of the interactive service, and are asked to retrospectively express their opinion on the presented stimuli or services, e.g., on a rating scale. To limit the impact of confounding factors, such experiments are commonly carried out in dedicated laboratory environments which are controlled regarding room acoustics and background noise conditions (for speech and audio services), or lighting conditions (for video services). Consequently, the ITU-T Rec. P.10/G.100 [45] defines the assessment of quality as follows:

Quality Assessment

The process of measuring or estimating the QoE for a set of users of an application or a service with a dedicated procedure, and considering the influencing factors (possibly controlled, measured, or simply collected and reported). The output of the process may be a scalar value, multidimensional representation of the results, and/or verbal descriptors. All assessments of QoE should be accompanied by the description of the influencing factors that are included. The assessment of QoE can be described as comprehensive when it includes many of the specific factors, for example, a majority of the known factors.

To this end, researchers make use of the field of psychophysics, which deals with the relationships between physical quantities and their perception by humans. However, the assessment of gaming QoE is a very challenging task, as gaming in a broader sense can be seen as a multimodal interaction. In contrast to a service such as telephony, the physical quantities are not solely limited to sound waves but also electromagnetic waves resulting in a visual perception, and even pressure levels of an input device causing a tactile sensation might be considered. Even though haptic feedback and audio quality will only play a minor role in the presented research, the performed assessment of gaming QoE is also highly complex due to its multidimensionality as well as the very high number of potential influencing factors (cf. Section 2.2.2). This makes the decision about which quality aspects should be measured and which quality factors investigated, very complicated. Also, the motivations and expectations of players, which are very important for the quality judgement process, may change over time. The gaming market

is rapidly changing and also newer hardware components offering higher resolutions and framerate have an influence on this matter.

3.1.1 Classification of Assessment Methods

With respect to available methods to evaluate gaming systems, the complex and interactive nature of gaming services as well as their special goals (cf. [58]) leads to the fact that traditional usability measures such as task completion time, error tracking, and heuristics as well as behavioral assessment and psycho-physiological assessment, alone, can only be used to a (so far) rather limited extent to examine gaming QoE. To motivate the selection of assessment methods used in the present thesis, which addresses the RQ2, an overview of various assessment methods will be given in the following.

Behavioral Assessments

Behavioral assessments are characterized by using observations and tracking of (mostly non-intentional) user behaviors with regard to physical movements, social interaction as well as in-application actions [120]. While the former methods typically aim to assess the emotional state of users such as excitement, anger, sadness, and calmness, the latter can be used to derive various measures of interactions with a game such as the Actions per minute (APM) or even performance metrics. Performance metrics like scores, deaths, or progress information might be an indicator of how well a player was performing compared to other gaming sessions, but they do not necessarily reflect the self-judgement of a player's performance nor the feeling of competency. In contrast to productivity applications, in games, such metrics have no intrinsic sense of good or bad [71]. Furthermore, these metrics are not generalizable as concepts such as scores vary strongly between game scenarios and are not applicable to every type of game [130]. Also, touch information with respect to the pressure level can be used to detect the emotional states of players. Gao et al. used players' finger stroke behavior to measure four emotional states (excited, relaxed, frustrated, and bored) as well as two levels of arousal and of valence. Valence describes the positive or negative affectivity, whereas arousal measures how calm or excited a user is. Their analysis showed that pressure features can be used to distinguish between frustration and the other three emotional states. Additionally, arousal states were discriminated by stroke speed and directness features, whereas stroke length features showed differences between boredom and relaxation [131]. Behavioral assessments seem to be especially used to evaluate the playing quality, also called game usability. In their survey of 50 North American game companies regarding game usability, Rajanen and Tapani revealed that companies mainly use playtesting (cf. [10]), observation of live gameplay, usability testing, and focus groups [101].

Psycho-physiological Assessments

Some of the behavioral methods, e.g., facial expressions or eye movements, can be applied either remotely or via psycho-physiological methods such as Electromyography (EMG). Physiological measures can be grouped according to their physiological basis. These include the central nervous system via so-called brain-computer interfaces (e.g., Electroencephalography (EEG), functional near-infrared spectroscopy, and functional magnetic resonance imaging), as well as the peripheral nervous system (e.g., Electrocardiography (ECG), Electrodermal activity (EDA), electromyogram, respiratory measurements, facial tracking, skin temperature), and lastly also audio/visual activity (e.g., eye tracking,

blinking, and pupil dilation). Recent innovations in these domains led to works investigating the use of physiological measures to assess stress [132], concentration [133], engagement [134], emotions [135], immersion [136], [74], and user experience [137]. Drachen et al. [138] investigated the correlation between Heart rate (HR) derived from ECG, EDA values, and PX while playing three different first-person shooter games for 20 minutes each. Every five minutes, the extracted features are compared to subjective ratings. Results showed a significant correlation between psycho-physiological arousal (HR, EDA) and self-reported measures for capturing the PX in the game. Du et al. [139] recently researched facial expressions and HR during a gaming activity by a non-contact method assessing emotion recognition. Video-captured data was used to detect the player's emotions during 30 seconds stimuli, whereas the intensity of emotion was measured using HR values. Bevilacqua et al. [140] found that variations of HR and facial actions are different comparing boring and stressful parts of the gaming sessions on a group and individual level.

Subjective Assessments

Subjective assessment methods determine users' opinions on their experience during or after a presented stimulus. The evoked gaming QoE can be measured either qualitatively or quantitatively. This includes an explicit acquisition of their feedback with regard to specific QoE aspects of interest.

Qualitative measures collect players' emotional and cognitive responses to games through direct observation with methods such as thinking aloud, focus groups, or structured interviews. These methods are typically applied to gain conceptual knowledge leading to a deeper and more detailed understanding of gaming QoE concepts. However, the collected data is usually low in reliability and external validity. Thus, achieving generalizable findings of a population of interest and comparisons of findings across multiple studies is very challenging. Due to their lack of standardization and comparability [141], qualitative methods are best suited for exploratory research, or in conjunction with other measures, such as questionnaires [142].

Questionnaires are the most widely used quantitative assessment method which ideally allow quantitative statements to be made about the relationship between physical quantities, perceptual quantities, and assessment quantities. To this end, one uses the method of scaling, i.e., the assignment of (physical or perceptual) quantities to numbers according to previously defined rules [143],[95]. Questionnaires offer a relatively easy means to assess the subjective experience under consideration while ensuring consistency and uniformity of collected data [141], as they are usually the result of a psychometrical validation process.

Comparison of Classified Assessment Methods

As indicated above, all methods have different strengths and weaknesses. Some behavioral methods require specific behavioral responses (e.g., pushing response buttons), which might negatively influence a natural interaction with a gaming system. Also, psycho-physiological methods such as the oddball paradigm require a very specific test setup, which is difficult to use during a typical game session. While a deeper analysis of the player inputs to a game, e.g, by analysis of touch pressure or in-game performance metrics, already showed some promising findings, such methods lack generalizability with respect to the player as well as game characteristics. Additionally, drawing conclusions from behavioral measures about the underlying subjective experience is often ambiguous, as a physical reaction might be caused by an independent variable to investigate, such as a network delay but possibly

also by an event in the game, e.g., a scary monster appearing. The same applies to most physiological measurements. An increased heart rate might be a sign of an ideal level of challenge, which could lead to a flow experience, but it could also be caused by an unexpected event or due to frustration about a bad network condition or simply losing a game. While psycho-physiological assessments are potentially less dependent on specific behavioral tasks, the installation of measurement devices such as EEG electrodes on the scalp inevitably causes higher intrusiveness and potentially discomfort, which could negatively influence the gaming experience. Additionally, due to the movements when interacting with a game or different context factors such as room temperature or fatigue, physiological data is often affected by noise, which reduces the analytical power of one-time measurements severely. Despite these disadvantages, behavioral and psycho-physiological assessments also have apparent advantages over subjective assessments, as they offer means for real-time monitoring of the state of participants over longer periods of time, without the necessity to interrupt the experience of a player. Additionally, in case them being applied in a non-intrusive way and in the observer's normal environment, a higher ecological validity could be achieved [144] compared to traditional lab tests using questionnaires. While such questionnaires provide a conscious user response on a given scale, and with that also insights into internal perceptual and cognitive processes, they suffer from potential memory effects for long stimuli [144]. Also, the interpretation of scale labels and usage of the range of a scale often vary strongly between participants, which also must be motivated to properly focus on the rating task. However, the fact that responses of questionnaires can be directly linked to specific aspects of gaming QoE, and that there are validated tools available, which are cheap and easy to apply to research, make questionnaires the most common method in current gaming research. Most scholars agree that physiological and behavioral indicators are most useful as adjuncts to subjective methods [142]. Physiological and behavioral measures may only be useful "when specific independent variables are being manipulated (e.g., display characteristics)" [142]. In recent attempts to define the complex concept of immersive media experience, authors of [120] argue that a multi-method approach combining the three assessment types, with each method compensating the disadvantages of the others, would be an ideal way to assess immersive experiences in all their facets. The same may apply to gaming QoE assessment in general. Subjective measures could be used to make sure that behavioral and psycho-physiological assessments are not influenced by unexpected events, but the advantage of real-time monitoring might be ensured. However, more research is required to understand how the different methods correlate with one another. Biases inherent to the method as well as inter-individual differences (e.g., between less and more empathetic participants) must be investigated and collectively reduced.

3.1.2 Questionnaire-based Assessment of Gaming QoE

In the following section, an overview of available questionnaires used for the assessment of gaming QoE will be given. This knowledge is important for the selection of a measurement tool that fulfills the demands of the targeted research of the presented work. Pagulayan highlights that the overall quality of a game, mistakenly, is commonly denoted as fun [145]. While gaming as an entertainment service surely aims for a joyful experience, it was already pointed out in Chapter 2 that the concept of gaming QoE is highly multidimensional. Consequently, a great number of questionnaires assessing gaming QoE are available. However, Bernhaupt et al. argue that there is no common agreement about a general framework that should be used to evaluating the gaming experience [10]. Also, there are no guidelines available on how to identify and categorize the exact experience due to the multitude of paradigms

and methodologies available [146]. Even though there is still some disagreement regarding their definitions, engagement, involvement, immersion, presence, and flow are often considered aspects for game satisfaction and user experience [10], [147]. Additionally, also concepts related to playability, fun, aesthetics, ease of use (controls, interface), challenge, pace (the rate of new challenges), and motivation are taken into account [10], [145], [148]. For an overview, a detailed, but by no means complete, list of questionnaires for assessing gaming QoE aspects is compiled in Table 3.1. For an even broader overview, Abeele et al. provided a list of 124 scales[1] used in player-computer interaction research [149]. Some extended reviews of available questionnaires can also be found in [10], [141], [146], [147], [150], [151]. In the following, some of the most widely used and comprehensive questionnaires will be introduced in more detail.

Game Experience Questionnaire (GEQ)

The GEQ is a "self-report measure that aims to comprehensively and reliably characterize the multifaceted experience of playing digital games" [152]. The GEQ was first published in 2007 [152] and updated in 2013 [153]. The GEQ has a modular structure consisting of a core module (concerning the actual PX during a scenario), a social presence module (concerning involvement with other social entities such as empathy) as well as a post-game module (concerning experiences once stopped playing such as returning to reality) [153]. Each module uses a 5-point ACR scale using the labels "not at all", "slightly", "moderately", "fairly", and "extremely". The GEQ core module assesses seven factors of PX: sensory and imaginative immersion, tension, competence, flow, negative affect, positive affect, as well as challenge. Additionally, a shorter version of the core module, the so-called In-game GEQ (iGEQ), using an identical component structure consisting of 2 items per factor is available. While the authors demonstrated good reliability of the scales in [152], discriminant validity was only demonstrated by response patterns to variables such as gender and game type. In a systemic review in 2018, Law et al. found 73 publications which used the GEQ [154]. However, some concerns have recently been raised regarding the reliability and validity of the GEQ core module [154]–[157]. Norman states that the GEQ seems reasonable and applicable in studying PX with video games, but it might not be suitable for games that do not involve a narrative or for which the story is intended to put the player in a bad mood (e.g., survival horror) and non-competitive games (e.g., simulations) [158],[96]. This might be especially true for the iGEQ which uses an item to assess immersion that is asking about the story of the game. As not every game scenario offers a story *per se*, in particular not in a short duration of only a few minutes, an adaptation of this construct seems reasonable. Law et al. present the results of a validation study (N = 633), which detected several issues of the GEQ. The reliability of the sub-components was not satisfactory for negative affect and barely satisfactory for challenge. In addition, a factor analysis revealed problems with the originally postulated factor structure based on calculated model fit indices as well as discriminant validity issues [154]. Similar findings were found by Brühlmann et al. during an analysis of the GEQ factor structure as they conclude that a single negativity factor might be more appropriate than the constructs of negative affect, tension/annoyance and challenge [157]. The authors further argue that although evidence of a different and less stable structure of the GEQ was shown, this does not mean it can not be a valid predictor of a gaming experience. While these findings harm the inconceivable use of the questionnaire, a factor analysis of the iGEQ, as it only consists of two items per factor, is not available. However, for the items of the iGEQ, the reported rotated pattern

[1] The list can be assessed using the following URL: https://goo.gl/jxPttB

matrix (cf. Table A.1 in Appendix A) based on Law et al. [154] does not show any high cross-loadings nor issues with overlapping factors apart from negative affect and tension. The latter might be the case as participants solely played their game of choice, which might cause bad experiences related to frustration or boredom.

Player Experience and Need Satisfaction (PENS) Questionnaire

The PENS questionnaire [159] is designed to explain the gameplay factors that lead to enjoyable and meaningful PX. Its authors Rigby and Ryan argue that engagement is motivated by a players' abilities to satisfy the three fundamental psychological needs competence, autonomy (freedom of choice), and relatedness (interaction with other players). These aspects are part of the well-known Self-Determination theory. The PENS questionnaire contains 21 items assessing the following five components: competence, autonomy, presence, relatedness, and intuitive controls [96]. The questionnaire has been statistically validated to some extend, and extensively used in a number of studies [160], [161],[162]. Johnson et al. further investigated the validity of the questionnaire and argue that claimed structure is partially supported but that competence and intuitive controls appeared to be a single construct, rather than two separate constructs as hypothesized in the PENS [156]. Brühlmann et al. reported that the factor structure of the PENS appears to be consistent and invariant [157].

Core Elements of the Gaming Experience (CEGE)

Gamez developed the CEGE questionnaire using an iterative process and structural equation modeling for its validation (N = 598) [163]. The questionnaire, targeting to measure enjoyment, consists of two guiding elements called puppetry and video-game, followed by control, ownership, and facilitators. Studies revealed an adequate fit, even though not optimal, suggesting that the CEGE model is an accurate abstraction of the process of the gaming experience. The questionnaire was also successfully used to investigate different input devices and game storylines [163].

Playful Experiences Questionnaire (PLEXQ)

The PLEXQ [164] assesses four factors of playful experiences: stimulation (discovery, exploration, challenge, expression, fellowship, and nurture), pragmatic (completion, control, competition, sensation, and thrill), momentary (relaxation, humor, and captivation) as well as negative experiences (suffering, cruelty, and subversion). While the PLEXQ was developed to measure "playfulness inherent in and experienced by users in their interactions with interactive technologies" it was also used for gaming services and showed good reliability [164]. However, more research about its validity is required.

Player Experience Inventory (PXI)

One of the most comprehensive and recently published frameworks, which is also strongly in line with the taxonomy presented in Chapter 2, is the PXI. Based on feedback from 64 experts in the field of player computer interaction, Abeele et al. designed and refined a scale using two major concepts: quality aspects at the functional level and at the psychosocial level. Therefore, the 12 constructs enjoyment, competence, autonomy, ease-of-control, cognitive immersion, meaning, effort, aesthetic appeal, progress feedback, clarity of goals, challenge, were used as possible candidates for the scale development. The initial results of a principal factor analysis published in 2016, suggest the scale can be used accurately to evaluate PX [149]. In 2020, the scale was validated and evaluated over five studies (N=529) providing evidence for both discriminant and convergent validity. The final model consists of 30 items (3 items per factor) assessing meaning, mastery, immersion, autonomy, curiosity, ease of control, audiovisual appeal, challenge, goals, and rules, as well as progress feedback [97].

Table 3.1: Overview of questionnaires for assessing gaming QoE aspects.

Name of Questionnaire	First Author	Year	Quality Aspects	Items	Scale
Cognitive Absorption Scale (CAS)	Agarwal [115]	2000	cognitive absorption, ease of use, usefulness, personal innovativeness, playfulness, intention to use, self efficacy	52	7-point
Core Elements of the Gaming Experience (CEGE)	Gamez [163]	2009	enjoyment, frustration, control, puppetry, facilitators, ownership, game-play, environment	38	7-point
EGameFlow	Fu [165]	2009	immersion, social interaction, challenge, goal clarity, feedback, concentration, control, knowledge improvement	42	7-point
Engagement Questionnaire (EQ)	Mayes [166]	2001	interest, authenticity, curiosity, involvement, and fidelity	46	7-point
EVEQ-GP Questionnaire	Takatalo [167]	2011	presence, involvement, flow	139	7-point
Flow State Scale (FSS)	Jackson [168]	1996	challenge-skill balance, action-awareness merging, clear goals, unambiguous feedback, concentration on task, paradox of control, loss of self-consciousness, transformation of time, autotelic experience	36	5-point
Flow-Short-Scale (FKS)	Rheinberg [169]	2003	fluency of performance, absorption by activity	13	7-point
Game Engagement Questionnaire (GEngQ)	Brockmyer [170]	2009	immersion, presence, flow and psychological absorption	19	3-point
Game Experience Questionnaire (GEQ)	Poels [152],[153]	2007	immersion, flow, competence, positive and negative effect, tension, challenge	33	5-point
Gameplay Scale	Parnell [171]	2009	affective experience, focus, playability barriers, and usability barriers	26	5-point
Gaming Engagement Questionnaire (GEQ_c)	Chen [172],[173]	2005	unknown	25	7-point
Igroup Presence Questionnaire (IPQ)	Schubert [174]	2001	spatial presence, involvement, realness	14	7-point
Immersive Experience Questionnaire (IEQ)	Jennett [74]	2008	person factors: cognitive involvement, real world dissociation and emotional involvement; game factors: challenge and control	33	7-point
ITC Sense of Presence Inventory (ITC-SOPI)	Lessiter [175]	2001	spatial presence, engagement, naturalness, negative effects	17	5-point
Locus of Control Scale (LOC)	Keller [176]	2008	focused concentration, enjoyment, involvement, motivation	13	7-point
Play Experience Scale (PES)	Pavlas [177]	2012	autotelic experience, freedom, focus, non-extrinsic, direct play	16	6-point
Player Experience Inventory (PXI)	Abeele [149]	2016	meaning, mastery, immersion, autonomy, curiosity, ease of control, audiovisual appeal, challenge, goals and rules, progress feedback	30	7-point
Player Experience of Need Satisfaction (PENS)	Ryan [159]	2006	competence, autonomy, relatedness, and intuitive controls; presence dimensions: narrative, emotional, physical	21	7-point
Playful Experiences Questionnaire (PLEXQ)	Boberg [164]	2015	stimulation, pragmatic, momentary, negative experiences	51	5-point
Presence Questionnaire (PQ)	Witmer [178]	1998	main: control, sensory, distraction, realism; sub: involvement, natural, auditory, haptic, resolution, interface quality	29	7-point
Temple Presence Inventory (TPI)	Lombard [142]	2011	spatial presence, social presence, presence as engagement, presence as realism, and others	42	7-point
User Engagement Scale (UESz)	Wiebe [179]	2013	focused attention, perceived usability, aesthetics, and satisfaction	28	5-point
User Engagement Scale revised (UES)	OBrien [128]	2018	aesthetic appeal, focused attention, perceived usability, reward	12	5-point

Gameplay Scale

Parnell developed the Gameplay Scale to assesses players' attitudes towards a game's appeal and quality. The Gameplay Scale is validated across two studies with a rather low sample size (N = 98 and N = 17) in which players had to answer a web survey after playing downloadable games. The questionnaire is composed of four sub-scales measuring different gameplay constructs: affective experience, focus, playability barriers, and usability barriers. While it was shown that the Gameplay Scale accounted for 73% of the variance in a game's initial appeal, further research is required for its validation [171].

User Engagement Scale (UESz)

The UESz, a tool to measure engagement during video game-play, was developed by Wiebe et al. The authors applied and adapted the User Engagement Scale by O'Brien and Toms to gaming experiences. The revised UESz is comprised of 28 items and four sub-scales: focused attention, perceived usability, aesthetics, and satisfaction. The factors were revealed by an exploratory factor analysis based on ratings of 413 participants. The scale showed an acceptable (Cronbach's $\alpha = .91$) reliability. Additionally, a satisfying validity analysis was performed by a comparison with the Flow State Scale [179].

Immersive Experience Questionnaire (IEQ)

Jennett developed the IEQ based on Brown and Cairns' grounded theory of immersion, as well as previous studies in the related areas of flow, cognitive absorption, and presence as a basis. The IEQ aims to assess the levels of immersion experienced by players by addressing varying degrees of attention during the task (basic attention, temporal dissociation, and transportation) as well as factors that could influence a person's motivation during the task (challenge, emotional involvement, enjoyment) [74]. A factor analysis demonstrated that there might be five underlying concepts of immersion: person factors (cognitive involvement, real-world dissociation, and emotional involvement) and game factors (challenge and control). The IEQ was statistically validated and has been used extensively across a diverse array of use cases and game genres [141], for example to investigate the effect of surroundings, controllers, challenge, and screen sizes [146]. Despite representing a suitable tool to measure gaming immersion, the IEQ is with its 31 items a rather lengthy questionnaire.

Flow State Scale (FSS)

The FSS developed by Jackson and Marsh assesses participants' level of flow experience [168]. The questionnaire consists of 36-items on 9 sub-scales. The concidered concept of flow include challenge-skill balance, action-awareness merging, clear goals, unambiguous feedback, concentration on task, paradox of control, loss of self-consciousness, transformation of time, and autotelic experience [180]. Jackson and Eklund improved this method and created the Flow State Scale-2, which is assessing flow in a physical activity at two levels based on the frequency of flow experience in particular domains, and the extent of flow experienced in a particular activity [181].

EGameFlow Questionnaire

The EGameFlow questionnaire created by Fu et al. is an adapted version of Sweetser's & Wyeth's GameFlow framework to assess the enjoyment offered by e-learning games. The scale consists of eight components: immersion, social interaction, challenge, goal clarity, feedback, concentration, control, and knowledge improvement. It contains 42 items and was developed by scale verification studies (N = 166) including four games. A satisfying reliability and validity based on a proper factor analysis was proven [165].

The Dilemma of Questionnaires for Assessing Gaming QoE

As shown, there is a great number of available questionnaires that were developed to assess gaming experience related aspects which were used in a variety of studies. This makes the selection decision for researchers very difficult which is even more complicated by the fact that a variety of questionnaires aim to measure the same construct and no guidelines about ideal usage scenarios are currently known. On the other hand, most questionnaires focus on a specific aspect of gaming QoE and do not consider other important factors [182]. This inevitably leads to the necessity to combine various questionnaires in case someone aims to assess the full spectrum of a gaming experience. Not only could this cause overlapping constructs to be measured, but it would also increase the total number of items a participant has to rate significantly. In fact, most of the questionnaires are by themselves already very extensive (cf. Table 3.1). In case that there is no adequate balance between experience time and questionnaire-answering time, the target of the measurement (i.e., the experience) might get blurred or even vanish. In turn, the measurement outcomes might become questionable [11]. Abeele et al. argue based on their expert interviews that concepts at one level can be causal to the higher psychological constructs. They also revealed a different focus of 'academic' game researchers and 'designer' game researchers which complicates finding a consensus on how to measure PX empirically [149]. Phan further criticizes that the majority of the existing questionnaires are based on a limited number of video game titles and have not been published and validated properly as researchers often do not follow guidelines for scale development such as an exploratory factor analysis followed by confirmatory factor analysis [182]. Lastly, Möller et al. argue that it may be doubted whether questionnaires are an optimum tool for measuring user states such as immersion or flow, as being confronted to the assessment task (i.e., questionnaire answering) may destroy the corresponding state, and it might be hard to remember for an in-retrospect judgment. In turn, as long as no alternative validated methods are available for measuring these states (such as a multi-method approach proposed in [120]), questionnaires may be the only way to approach such constructs [11].

3.1.3 Methods Considered for Research Objectives

Even though a systematic evaluation of all the former methods and their possible combinations would be a highly valuable contribution to the research community, such a strategy would be too ambiguous for the frame of the present work. Thus, at this point, it should be discussed which assessment methods are suitable within the scope of this thesis.

First of all, it must be emphasized that neither detailed research about game design leading to a good gaming QoE, nor virtual reality gaming targeting the concept of physical presence, nor the improvement of user motivation is in the focus of the presented work. Thus, traditional and game usability assessment methods are not promising for the evaluation of cloud gaming services, since the major focus is about investigating the influence of network and encoding parameters, or important quality factors which can be considered by service and network providers. Consequently, within this work, methods such as focus groups or structured interviews will only be used for exploratory research about the completeness of the taxonomy presented, about characteristics of games which might moderate the impact of network and encoding parameter, to gain insights about the importance of quality aspects on the overall experience, and lastly to receive feedback about flaws in the design of subjective experiments, e.g., with respect to the duration of a stimulus. Concerning behavioral and psycho-physiological assessments, even though some interesting findings and methods are available at current, the research is not advanced

enough to measure specific quality features reliably as changes in these measurements often correlate to multiple aspects. Since one important aim of the present work is to empirically validate the cloud gaming taxonomy, quantitative data about the most important quality aspects presented in Chapter 2 are required. Thus, the preferred method for the assessment of gaming QoE throughout the majority of studies will be the use of questionnaires.

However, it must be noted that even though there is a multitude of validated questionnaires assessing a large variety of concepts available, there is no one-size-fits-all approach available to cover all the quality features covered by the taxonomy presented earlier. To the best of the author's knowledge, the only framework which considers psychosocial and functional consequences at the construct level is the PXI based on the Mechanics-Dynamics-Aesthetics (MDA) framework and Means-End theory [97]. However, this method is not designed for cloud gaming and is missing fundamental components such as the input quality and output quality. Furthermore, it targets enjoyment instead of overall gaming QoE. Therefore, questionnaires have to be combined for assessing the full range of required features or used in separated studies. The selected questionnaires need to fulfill several requirements. As many quality features must be assessed in parallel, the number of items of the full measurement method should be limited as much as possible to avoid fatigue of participants which might negatively impact the quality of the assessed data in respect to its reliability and validity. The used scales should be as consistent as possible to avoid confusion of participants. Lastly, general guidelines for good questionnaire development (cf. Chapter 5) such as avoiding negative items, long and multifaceted labels should be followed. The concrete combination of selected questionnaires will be described during an exemplary test design presented in Section 3.3. In preparation for this design, in the following a study about different test paradigms will be presented.

3.2 Comparison of Interactive and Passive Test Paradigm (Study 3.2)

As a game is an interactive endeavor and differs in many aspects from task-directed interactions, typically interactive tests with longer test sessions are conducted to assess the gaming QoE including aspects such as the interaction quality, fun, immersion, and flow. However, since the duration of such tests must be limited due to fatigue, the number of stimuli under test is also limited. Furthermore, the cognitive load while playing a game might negatively reduce the resources of participants to concentrate on the evaluation task. A passive viewing-and-listening test, referred to as a passive test in the following, where the quality of pre-recorded or generated gaming content without playing actively will be rated, possibly allowing the use of shorter stimuli, would be very useful to overcome the aforementioned disadvantages of interactive tests. Additionally, every participant would rate the same audio-visual stimuli independent of their playing abilities. While also such passive tests have their drawbacks, e.g., the incapacity to evaluate interaction quality or the impact of delay, they might be valuable for other, more appropriate use cases and result in valid and comparable studies. These thoughts motivated the continuation of research about two different test paradigms, passive and interactive tests, which will be explained in more detail in the following section. The work is largely based on a publication of Schmidt et al. [15] in 2018.

In an ITU-T Contribution by Orange SA [183], the authors proposed a passive evaluation method for gaming QoE research by conducting an experiment with video quality assessment followed by an interview on QoE. The following four quality aspects are considered: visual fatigue, fluidity, visual

discomfort, and gameplay. More concretely, the SAMVIQ methodology using a multi-stimuli with random access approach with a sequence viewing duration of 10 or 15 seconds is suggested for stabilized and reliable quality scores (cf. [184]). With the aim to prove the suitability of the proposed method, Beyer et al. [185] carried out a subjective experiment comparing the interactive (stimuli duration of 2 minutes) and passive test paradigm (stimuli duration of 10 seconds). The results showed that participants rated the overall and video quality significantly higher during the passive test for a bit rate of 100 Mbps (no degradation), but significantly lower for a bit rate of 3 Mbps (blockiness). Furthermore, the authors reported that participants used a greater range of the scale during the passive test. A similar experiment, leading to slightly different results, was carried out by Sackl et al. [80]. A comparison of video quality ratings between interactive and passive tests of the same game under different levels of bit rate revealed that participants rated the video quality slightly higher when playing compared to watching the video games for 3, 5, and 10 Mbps, but the differences were not statistically significant. Unfortunately, the duration of the stimuli was not reported. Using a duration of 10 seconds is a standard in the video quality community as it is proposed in ITU-T Rec. P.910. However, the knowledge of the influence of the stimulus duration for gaming research so far is very limited. While in [183] it was stated that the duration must be limited to 15 seconds due to memory capabilities of participants (a justification or reference was not given), Mullin criticized a stimulus duration of 10 seconds for listening tests. He argues that "this length of time does not afford the opportunity to experience the unpredictability of some networks or, if loss rates are low, the full potential of the resulting impairment" [186]. For the use case of passive tests, it must be noted that some types of degradations, e.g., network delay, cannot be assessed. Thus, when comparing interactive and passive tests, suitable degradations must be chosen based on the state-of-the-art presented in Section 2.2.2.

Applied Methodology to Investigate the Test Paradigms

With the aim to compare the two test paradigms, interactive and passive test, a subjective experiment as outlined in Fig. 3.1 was designed. In contrast to the work in [183],[185],[80], it was decided to use two different stimuli durations, namely 10 and 90 seconds, for the passive tests. Additionally, the PX during the passive test was assessed, and the influence of the player performance for the selection of the video material was investigated.

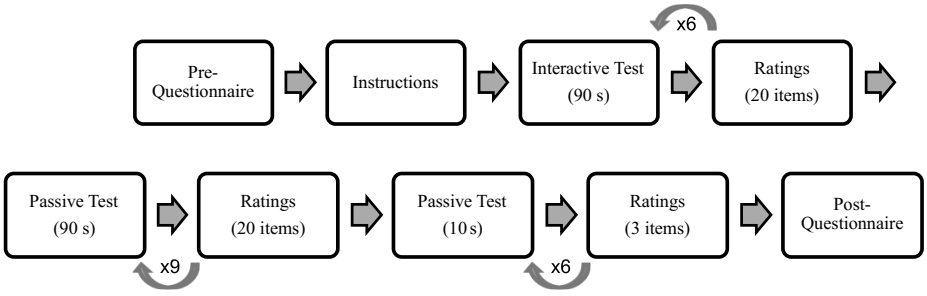


Figure 3.1: Schema of study design (cf. [15]).

Since the game content was identified as an important influencing factor [86],[72], two games for both test paradigms were selected: the action game Grand Theft Auto (GTA) 5 and the racing game Project Cars. More information about the games can be found in Appendix D, in which also information about all other games used throughout the thesis is summarized. A within-subject design to reduce the impact of user factors was chosen, and the order of the interactive and passive tests as

well as of the stimuli (games or gameplay videos under certain encoding conditions) was randomized. The study started with a pre-questionnaire to assess user factors and ended with a post-questionnaire asking participants about their judgment process and peculiarities during the tests.

As both test paradigms should be able to assess gaming QoE under various degradations, three different video encoding conditions were selected: a reference condition without visually perceivable degradations using a bit rate of 30,000 kbps at 60 fps, a low bit rate condition resulting in blockiness in the video using a bit rate of 1,500 kbps at 60 fps, and a low frame rate condition causing jerkiness in the video using a bit rate of 30,000 kbps at 5 fps. The conditions were chosen to represent a large range of different qualities. While such a low frame rate is very uncommon for pure video streaming, it is indeed possible for online gaming, e.g., caused by excessive server loads.

Participants were instructed regarding the test procedure including information about the duration of the stimuli, how and when to rate their gaming QoE, as well as about the rules and goals of the games. Furthermore, the difference between video quality and graphical quality was explained and the questionnaire used after each stimulus was examined. For the interactive test, a duration of 90 seconds for each stimulus was used. In pre-tests and during post-test interviews, participants majoritarian stated that this duration is sufficiently long enough to receive an impression of the condition and to rate their experience. As a cloud gaming service, the Steam In-Home streaming application was used. The settings for the bit rate and frame rate used for each stimulus can be changed via console commands. The average bit rate received by the client was in line with the parameters set on the server, even though there were some small changes during low-motion scenes.

For the passive test, first, a game scenario, i.e., a timely limited section of a game, referred to a scene in the video community, was captured lossless for each game using FRAPS[2]. The degraded stimuli were generated using FFmpeg with the same encoding parameters as for the interactive test. The video material was provided to the participants with two durations, 90 seconds as in the interactive paradigm and 10 seconds as suggested in [96]. Representative game scenarios similar to the interactive scenario were chosen. To investigate the impact of the player performance on quality ratings in a passive test, two different game scenarios of Project Cars were created for the long passive test; one showing a very good and one a very bad player performance. Regarding the instructions for the passive test, the main difference was that participants were told to imagine they would have played the game under the conditions they see in the video and rate their gaming QoE accordingly. This enables them to judge also PX aspects such as fun (positive affect) or frustration (tension).

To assess the gaming QoE, a 20-item questionnaire was used for the 90-second tests, and a 3-item questionnaire was used for the 10-second test, as for the later a judgement of PX is not feasible. An electronic questionnaire provided the participants with 7-point continuous scales as proposed in ITU-T Rec. P.851 [187]. In addition to the overall quality (“How do you rate the overall quality?”), the video quality (“How do you rate the Video Quality?”), reactiveness (“How frequently did you notice delayed reactions of the game?”), performance (“How do you rate your own performance?”), control (“I had control over the game.”), willingness to play again (“Would you play the game under these conditions again?”), and the 14 items of the in-game Game Experience Questionnaire (iGEQ) [153] were assessed. It must be noted that the reactiveness is negatively coded, i.e., a high value corresponds to a low reactiveness of the game. All games were played in a resolution of 1080p on a 24’ standard Monitor with a refresh rate of 60 fps. Twenty-one participants (7 females, 14 males), aged between 20

[2]https://fraps.com/

and 32 years (Mnd = 25 years) provided valid data for the analysis. The majority were students and non-expert gamers, but familiar with controlling similar games.

Impact of Test Paradigm on Game-related Quality Aspects

To compare the different test paradigms with each other, a three-way repeated measure Analysis of Variance (ANOVA) was calculated. The 7-point ratings are transformed to 5-point ratings as shown in [188] to match to commonly used MOS scales. The *game, encoding condition*, and *test paradigm* were used as independent variables. The *overall quality, reactiveness*, and *video quality* were used as dependent variables.

While there was no main effect of the game on any dependent variable, there were very strong main effects of the encoding condition on the overall quality ratings, $F(1.5,29.0) = 464.9$, $p < .001$, $\eta_p^2 = .96$ (due to violation of sphericity, a Greenhouse-Geisser correction was performed), on the reactiveness, $F(2,40) = 1402.4$, $p < .001$, $\eta_p^2 = .98$, and on the video quality, $F(2,40) = 144.5$, $p < .001$, $\eta_p^2 = .88$. This effect was expected since the conditions were chosen with the intention to have strong variations in the ratings. The means, standard deviations, and results of the ANOVA, including the effect size η_p^2, for the main effect of the test paradigm are summarized in Table 3.2.

Table 3.2: Descriptive statistics (averaged over both games) and ANOVA statistics of each game for the main effect of test paradigm.

Quality Aspect	Interactive Test (90 s)		Passive Test (90 s)		Passive Test (10 s)		ANOVA statistics GTA 5			ANOVA statistics Project Cars		
	M	SD	M	SD	M	SD	F	p	η_p^2	F	p	η_p^2
Overall Quality	2.90	1.16	3.10	1.12	3.35	1.10	7.27	< .01	.27	13.00	< .001	.39
Video Quality	2.90	1.21	3.07	1.16	3.58	0.96	10.53	< .001	.35	28.07	< .001	.58
Reactiveness	2.70	1.55	2.63	1.57	2.38	1.50	11.00	< .001	.36	17.82	< .001	.47

It can be observed based on the mean values that the ratings of the overall quality, video quality, and reactiveness during the interactive test and the long passive test (90 s) are very similar. Post-hoc tests showed no significant differences, $p > .05$. However, the ANOVA revealed a main effect of the test paradigm for all three dependent variables. This is due to the ratings during the short passive test (10 s), which are in general more positive. The highest difference of 0.68 between the interactive and short passive test can be seen for the video quality. On the first note, the similarities of the ratings during the interactive and passive test with a duration of 90 seconds are surprising and promising. It appears that such a passive test indeed could replace an interactive test. However, as shown, the duration and possibly other aspects need to be considered carefully. In the following, a critical review of the results for both games separately and video encoding conditions will be presented. To get a visual impression of the results, Fig. 3.2 shows the bar plots (including p-values for significant differences between the test paradigms) of the overall quality, video quality, and reactiveness for both games used in this study. For both games, a two-way ANOVA yielded a main effect of the test paradigm on the dependent variables, as shown in Table 3.2.

For GTA 5, pairwise comparisons (adjusted significance level using the Bonferroni correction) between the test types revealed significant differences of video quality ratings during the short passive test (10 s) compared to the long passive test (90 s) as well as to the interactive test, for each video encoding settings. This also is the case for Project Cars during the reduced bit rate (1,500 kbps and 60

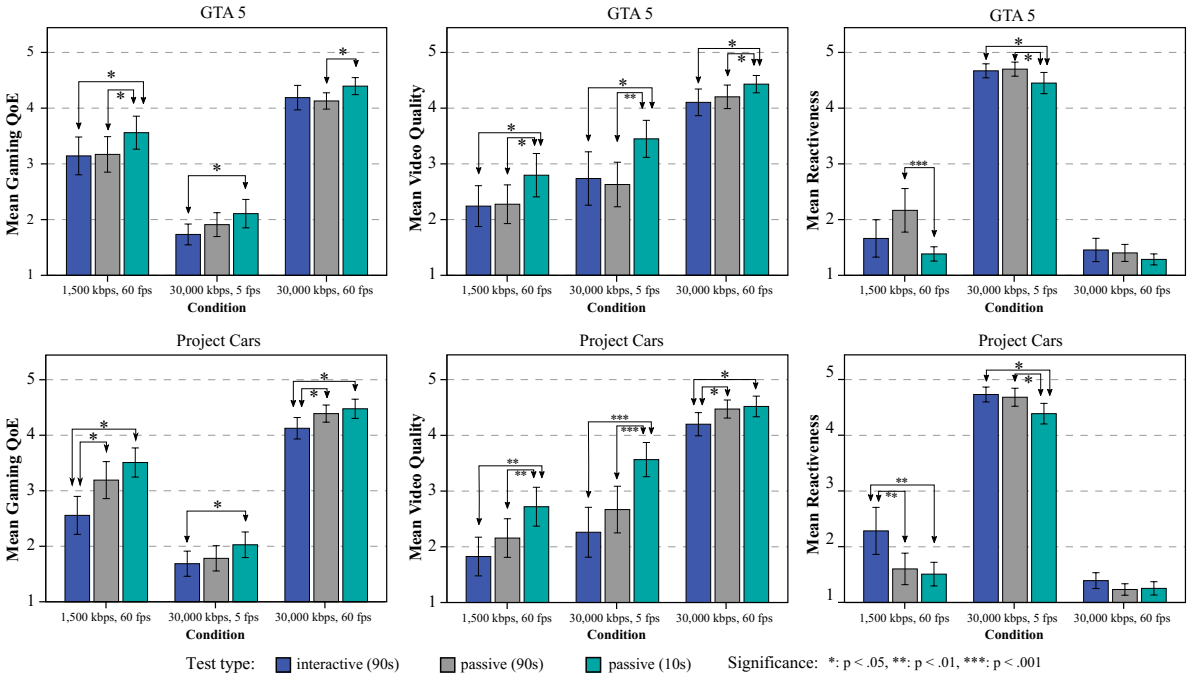


Figure 3.2: Bar plots of means and 95 % confidence interval for overall quality, video quality and reactiveness (labels for gaming QoE and video quality (1-5): bad, poor, fair, good, excellent; labels for reactiveness (1-5): very rarely, rarely, sometimes, often, very often) (cf. [15]).

fps) as well as with the reduced frame rate (30,000 kbps and 5 fps) encoding condition. However, for the reference condition of Project Cars, the ratings of the interactive test differ significantly from both passive tests. With respect to the reactiveness ratings, the behavior of both games is slightly different. While for the low frame rate condition, again the short passive test differs from the other two test types (no effects are found for the reference condition), the reactiveness during the interactive test was rated significantly higher for Project Cars and during the long passive test significantly higher for GTA 5. The higher ratings during the interactive test for Project Cars might be caused by the inertia of the car. Since most participants were non-expert gamers, they might not have been able to differentiate between an intended and non-intended delay in the game. Regarding the overall quality, the findings are generally in line with the relations reported before. However, due to the bad reactiveness during the interactive test for Project Cars, strong differences compared to the other two test paradigms are reported. Similar to the video quality, also for the reference condition a difference between the interactive test and long passive test can be seen for Project Cars. In contrast to the video quality, for the low frame rate condition, there is no difference between the long and short passive test for the overall quality.

Impact of Test Paradigm on Player Experience Aspects

In the following, it will be investigated whether there are differences for the PX assessed with the iGEQ during the interactive and long passive test (90 s). Since an in-depth analysis for every quality aspect would be too extensive, only the mean values of each condition and ANOVA statistics for the main effects of encoding condition and test paradigm are reported in Table 3.3. When comparing both used games with each other regarding the PX aspects, all ratings are remarkably equal for each encoding condition. For this reason, the mean values are not separated by the game. As shown in the previous subsection, there is no statistically significant difference between the overall quality rating for both test paradigms. This finding can be explained by comparing the remaining PX aspects. Even though the

aspects competence, control and judgement of own playing performance differ significantly, there is no main effect of the test paradigm for any other PX aspect with exception of positive affect and tension. It is plausible that the ratings for the three first mentioned aspects are different since the provided video material during the passive test only contained game scenes with a high performance. It is an important finding that despite differences between the test paradigms with respect to the feeling of competence and control, as well as the performance, no differences in the overall quality rating were observed.

Table 3.3: Mean values of player experience aspects during interactive and passive test (90 s). Significant differences are highlighted in bold. The abbreviations IA and PA refer to interactive and passive tests using a duration of 90 seconds.

	1.5 Mbps, 60 fps		30 Mbps, 5 fps		30 Mbps, 60 fps		ANOVA for condition			ANOVA for paradigm		
Quality Aspect	IA	PA	IA	PA	IA	PA	F(2,80)	p	η_p^2	F(1,40)	p	η_p^2
Positive affect	3.17	3.50	2.09	2.44	3.76	3.90	208.47	**<.001**	.84	8.02	**.01**	.17
Negative affect	2.53	2.46	3.32	3.34	2.27	2.15	94.87	**<.001**	.70	0.80	.38	.02
Competence	3.17	3.77	2.22	3.24	3.37	3.99	58.14	**<.001**	.59	44.11	**<.001**	.52
Challenge	3.26	3.23	2.74	2.77	3.44	3.46	43.34	**<.001**	.52	0.02	.89	.01
Flow	3.05	3.13	2.42	2.43	3.35	3.43	64.52	**<.001**	.62	0.52	.47	.01
Tension	2.48	2.16	3.36	3.28	2.09	1.75	120.74	**<.001**	.75	4.61	**.04**	.10
Immersion	3.03	3.18	2.49	2.64	3.45	3.67	89.12	**<.001**	.69	2.39	.13	.06
Performance	3.03	3.88	2.33	3.70	3.24	4.08	26.30	**<.001**	.40	65.10	**<.001**	.62
Control	3.47	4.00	2.00	2.95	3.99	4.25	155.27	**<.001**	.80	22.70	**<.001**	.36

Impact of Player Performance during the Passive Test

Next to the duration and complexity of a video scene, the question arises whether the behavior of a player shown in the video material, i.e., the player performance, does have an impact on user ratings. To investigate this matter, subjective ratings in a long passive test (90 s) using video material with a good and bad player performance were collected. Participants rated the performance (for all three conditions together) of the good performance condition with a mean of 4.00 (SD = 0.58) and of the bad performance with a mean of 1.61 (SD = 0.48). These differences are much higher compared to the performance values presented in the previous subsection. The same video encoding conditions are used as in the previous analysis. For statistical analysis, a repeated measure ANOVA using the *player performance* and *encoding condition* as independent variables was calculated. Fig. 3.3 shows the bar plots for the overall quality and video quality.

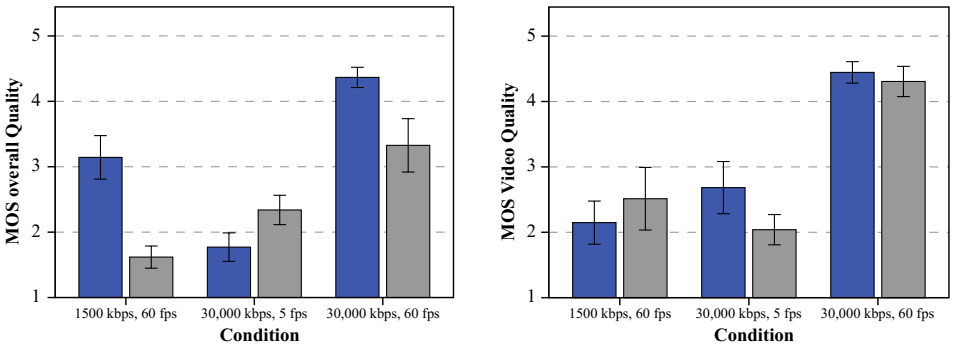


Figure 3.3: Bar plots of means and 95 % confidence interval showing the impact of player performance (blue: good, gray: bad) on overall quality and video quality during the long passive test (90s) (cf. [15]).

An important finding is that the ANOVA did not yield a main effect of the performance on the video quality. However, significant differences can be observed for the overall quality for all three encoding conditions, $F(1,21) = 36.85$, $p < .001$, $\eta^2 = .64$. This might be due to differences in the PX aspects since participants were asked to rate their overall experience by imagining that they would have played the game as shown to them. For the PX aspects, differences for the encoding conditions can be observed. For positive affect, flow, tension, and immersion, the ratings during the good performance are significantly higher and for negative affect, challenge, and tension significantly lower for the reference and low bit rate condition, while there is no difference between these quality aspects for the low frame rate condition.

Discussion about Findings

Regarding the test paradigm comparison, a strong similarity between the ratings of the interactive and long passive test was revealed but the short passive test overestimates the overall quality and video quality. Someone could argue that the differences in the video quality ratings between the long and short passive tests are a result of changes in the video complexity. For this reason, it is worth considering the Spatial information index (SI) and Temporal information index (TI) of a video scene as described in ITU-T Rec. P.910. Thus, the mean values of SI and TI (instead of maximum as suggested in [184] due to scene changes in the long video) for the reference videos used in the short and long passive test was calculated, and the difference (ΔSI and ΔTI) between each test paradigms for both games was derived. It can be summarized that there were only minor differences for GTA 5, $\Delta SI = 2.90$ and $\Delta TI = 1.10$, and Project Cars, $\Delta SI = 1.95$ and $\Delta TI = 0.03$. Consequently, the differences in the video quality ratings are arguably not caused by video complexity differences between the stimuli. Furthermore, results show that for the reference condition there is also a significant difference between the interactive test and long passive test for the overall quality. Yet, the differences are very small, thus only a very weak effect is observed. The differences in the quality ratings for Project Cars between interactive and passive test are higher than for GTA 5. A reason for this might be that participants during the interactive test perceived the movements of the car always as delayed (inertia), even though this is intended by the game. This illustrates that for the passive test to work also for overall quality judgements, domain knowledge is necessary, i.e., a proper selection or training of participants is a requirement.

On a general remark, there was a very high similarity of PX ratings between the interactive and passive test (90 s) for different games and conditions. It is worth underlining again that participants were solely imagining they would play the game under the given conditions. When considering that even repeated measures of the same interactive test showed slightly different results [72], [86], this is an astonishing finding, which might indicate that participants judge their experience strongly related to technical degradations. However, it was shown that the differences between the long passive and short passive tests are smaller for the overall quality compared to the ratings for the video quality. A reason might be that participants were not able to judge if a game would be fun to play in such a short time (positive affect has a strong impact on overall quality). However, due to the significant differences shown for positive affect and tension, when comparing the interactive and passive test results, one must conclude that a passive test cannot replace an interactive test in the case that the PX is the targeted measure. Lastly, the results of the passive tests with respect to PX indicate that participants expect jerkiness in the game to affect their positive PX aspects stronger than blockiness in the video, which was also reported by the participants in the post-questionnaire.

It must be noted that this study also has limitations with respect to the targeted investigation of the two test paradigms. The limited number of games tested makes it difficult to generalize the results. Even though it was shown that at least the long passive test resulted in valid outcomes for the video quality and (in most cases) for the overall quality, this does not prove that a passive test is also valid for every game. A second limitation of the study is that the encoding conditions were extreme cases and it is unclear if small quality degradations can be accurately assessed with passive tests compared to interactive ones, even though there is no indication or logical reason for different behavior. Furthermore, the duration of the stimuli, even for the long conditions, was not very high. Especially for the assessment of the PX with the iGEQ a longer duration would be desirable but the duration had to be limited to prevent fatigue of the participants. Finally, there is no concrete knowledge at which duration the passive test starts to differ from the interactive test. The study only revealed that the threshold is somewhere between 10 and 90 seconds.

3.3 Designing Subjective Tests Measuring Gaming QoE

A suitable measurement instrument to measure gaming QoE is not sufficient on its own to enable the assessment of reliable and valid results from subjective tests. Thus, before describing the final selection of the measurement instruments, in the following section, the design of subjective tests for gaming QoE evaluation will be discussed.

3.3.1 Standardization Activities

The development of standardized methods for the evaluation of gaming QoE would be of great use for the research community and also industry stakeholders as they allow a comparison of study results and will hopefully lead to valid and reliable findings. In 2013, Beyer and Möller introduced the cloud gaming taxonomy to the ITU-T SG12 [20]. As the work rouse interest of the standardization body, they proposed a first structure of the targeted recommendation in September 2014 including information about the experimental setup to be used in the subjective evaluation, thoughts on questionnaires for quantifying user perception throughout the evaluation, and about the usefulness of performance measurements as well as physiological response measurements [189]. The draft was not only discussed at the ITU-T meeting but also among experts of the COST action IC-1003 Qualinet (“European Network on Quality of Experience in Multimedia Systems and Services”) which resulted in a publication in 2015 [190]. The valuable comments were also presented to the ITU-T SG12 by Möller et al. in [191]. Experts stated that the selection of the measurement tool should be decided based on the aim of the research, in particular by considering the effort caused by the rating task of participants. When evaluating the impact of network conditions on the user’s willingness to continue playing, interruptions resulting from the rating task should be kept to an absolute minimum. In turn, if the purpose is to identify and understand perceptual features underlying the gaming activity, then more detailed questionnaires must be used. Additionally, they highlighted the importance of the game scene selection which should be comparable between participants by taking into account a certain level of difficulty, its interactivity, spatial and temporal information as well as other factors that might potentially influence the interaction behavior of participants. Interestingly, they also proposed research about the use of crowdsourcing for gaming quality assessment (cf. Chapter 4) as well as about meaningful categorizations of games (cf. Section 6.2). Next, more knowledge about the requirements of participants should be collected and a

screening questionnaire that quantifies participant characteristics as far as possible (cf. Section 3.3.2 and 7.1.2) should be used for future studies. Lastly, it was recommended that the recommendation should include both passive and interactive procedures (cf. Section 3.2).

From experiences gained by conducting studies assessing gaming QoE for an investigation of the impact of delay moderated by the game scenarios [14], a comparison of different gaming platforms [192], the study presented in the previous section, as well as a great number of insightful feedback from test participants, the author of this thesis decisively contributed to the finalization of the ITU-T work item P.GAME about subjective evaluation methods for gaming quality in 2018.

In 2017, Schmidt et al. submitted an update on the recommendation draft with a special focus on the definition of engagement-related concepts including involvement, immersion, flow and presence, Relationships between engagement concepts, the study design with respect to the test environment, display specifications, social aspects, duration and presentation of stimuli as well as participant instructions and user factor assessments, selection of test material and available questionnaires [21]. The research about the two test paradigms presented in the previous section was discussed in the next SG12 meeting [22] and finally lead to the inclusion to the ITU-T Rec. P.809 which recommends the usage of passive tests with audio-visual stimuli with a duration of 30 seconds for experiments that are not focused on interactivity or the full spectrum of the PX [96]. Creating a dataset that covers all possible encoding parameters such as framerates, bitrates, and resolution as well as a high number of gaming contents while using a stimulus duration of 90 to 120 seconds would result in a not feasible effort. Thus, even though the recommendation is to some extend a pragmatic decision as this concrete duration was not explicitly investigated, it appears to be a reasonable compromise to enable work related to the broad range of possible encoding settings of a cloud gaming service. The duration of 30 seconds was also used in the work of Claypool [193], where the author investigated the motion and scene complexity for a wide variety of video games. However, to date, there is no study substantiating the validity of using 30 seconds long stimuli. In 2018, based on a final contribution by Schmidt et al. [23], the ITU-T Rec. P.809 was finalized. Apart from an overview about gaming QoE aspects covering hedonic and pragmatic quality as well as PX (cf. Chapter 2) as well as a summary of available questionnaires for the assessment of gaming QoE in order to help choosing suitable methods to conduct subjective experiments (cf. Section 3.1.2), the recommendation provides information about the test environment and test setup, participant instructions, and selection of game materials. For more details, the reader is referred to [96].

To allow a comprehensive understanding of all important parts of a study design to assess gaming QoE, in the following, work related to the ITU-T activities will be introduced.

At first, there is a variety of subjective scaling methods available which are used in the domain of speech quality assessment (cf. ITU-T Rec. P.80, P.800, P.851) as well as for video quality assessment (cf. ITU-T Rec. P.910 and P.911). A method suited for qualification tests and to investigate QoE in the common use of the system under test is the Absolute Category Rating (ACR) method [184]. Traditionally, a discrete 5-point quality category scale using the five attributes "excellent", "good", "fair", "poor", and "bad" is used to assess conversation quality [44] or video quality [184]. This scale is commonly known as the "MOS-scale". However, it must be noted that also for various other scale formats, a Mean Opinion Score (MOS) can be calculated and is frequently used in the QoE domain.

Alternatively, a slightly modified version of the ACR is often used: the ACR with hidden reference (ACR-HR) which demands the inclusion of a reference version of each test material [184]. The scale promises that the perceptual impact of the reference condition can be removed from the subjective scores including the impact of material bias (e.g., participants liking or disliking a content), the quality of the source material (e.g., small issue due to the recording process), as well as equipment factors (e.g., professional equipment versus consumer-grade) upon the final scores.

Another important topic is the dimension-based subjective quality evaluation for video content published in the ITU-T Rec. P.918. Instead of solely assessing the overall video quality, the methodology yields scores for five perceptual video quality dimensions which provide diagnostic information on what may cause a degradation.

To identify the video quality relevant perceptual dimensions, Schiffner and Möller [194] performed a pairwise similarity experiment with a subsequent multidimensional scaling and a semantic differential experiment with a subsequent principal component analysis [195]. Applying both test paradigms in separate experiments resulted in the five perceptual dimensions for transmitted video which are summarized in Table 3.4. Each dimension is explained to participants during the introduction in written

Table 3.4: Overview of the five identified and proposed perceptual video quality dimensions [195].

Name	Description	Example Impairment
Fragmentation	Fallen apart, torn and disjointed	Packet loss
Unclearness	Unclear and smeared image	Low coding bitrate
Discontinuity	Interruptions in the flow of the video	Buffer delay and limitations
Noisiness	Random change in brightness and colour	Quantization, circuit noise
Suboptimal Luminosity	Too high or low brightness	Over- and under-exposure

form using describing adjectives and in form of example videos. For the assessment of the dimension, the Direct Scaling method [196] using 7-point continuous bipolar scales (labels consist of antonym pairs such as discontinuous and continuous) is applied. To use this method for research in the gaming domain, typical gaming content is used for the introduction videos. Furthermore, in the specific context of a cloud gaming service, the discontinuity will be in the following considered as Temporal video quality (TVQ), whereas the remaining dimensions form the Spatial video quality (SVQ).

3.3.2 Test Design for ITU-T Rec. G.1072

In the following section, it will be described how a subjective experiment can be designed based on the knowledge provided in the ITU-T Rec. P.809 and some fundamentals of test design. Thus, the section targets the RQ3. The test design was particularly used to create a database of subjective ratings for the development of an opinion model predicting the QoE of cloud gaming services. Thereby, the passive as well as the interactive paradigm are considered. Whereas the dataset, which is of high importance also for the upcoming chapters and led to the ITU-T Rec. G.1072 in 2020, will be presented in more detail in Chapter 7, in the following the focus will be on its design and requirement specifications as those can and should be applied to the design of subjective test assessing gaming QoE in general. Many of the following information was also presented in various SG12 meetings [24],[25],[26].

Scaling Method and Rating Scale

In order to collect meaningful user ratings, a subjective scaling method has to be selected. As stated by the ITU-T Rec. P.910, this selection "for a particular application depends on several factors, such as the context, the purpose and where in the development process the test is to be performed". While the PC and DCR methods would be suitable to investigate the fidelity of the transmitted video during the use of a cloud gaming service, especially for stimuli that are very similar to each other, they would strongly reduce the number of different test conditions per study, thus causing a significant amount of additional effort, and they do not represent a common scenario in which a player would typically evaluate a cloud gaming system based on “world knowledge” and experiences in the home environment. Therefore, the ACR-HR method, which offers some additional benefits over the normal ACR method such as a possible reduction of the influence of the stimulus material, was selected a subjective scaling method.

Regarding the rating scales, a great number of research work is available investigating a broad range of scale aspects, e.g., optimal number of categories, usage of middle point, labeling of all point, or just endpoints. According to Krosnick and Fabrigar, the optimal number of scale categories is 5 to 7 [197]. In a systematic comparison, Weijters et al. [198] investigated different scales using three key response bias states. The authors state that a neutral scale point, i.e., using an uneven number of categories, typically leads to an increase in user tendency to agree. While balanced scales, i.e., same number of positive and negative categories, could counter this effect, an increase in misresponse to reversed items (MR) was shown. Lastly, the authors provide advantages of bipolar scales, i.e., only labeled endpoints with opposite meanings, by arguing that these scales are easy to construct as only two labels have to be formulated and that they are intuitively more in line with an interval scale assumption. This type of scale was used for the ITU-T Rec. P.918 to assess the TVQ and SVQ. On the other hand, Wildt and Mazis point out that a fully labeled scale facilitates the interpretation both by respondents and researchers [199] and also causes less MR, as it reduces the cognitive load by clarifying the meaning of response categories.

Therefore, to enable a higher discriminative power, a 7-point extended continuous ACR (EC-ACR) scale proposed in ITU-T Rec. P.851 and P.908 was used to assess the overall gaming QoE. The same scale was also used for the other items of the final questionnaire (labels were adapted) to be more consistent and to avoid confusing the test participants. While a comparison of the ACR and EC-ACR scale was performed by Köster et al. [188], who also provided a transformation of the 7-point ratings to typically used 5-point ratings, such research is still a topic for future research in the domain of gaming QoE. However, the possible higher sensitivity compared to the ACR scale and the assumption that naïve participants more frequently use the extreme categories of the scale due to its overflow region (while no signs of negative consequences are known yet) motivated the decision for this scale.

Test Structure

As it is the case with all evaluation methods, the chosen method should reflect the later use scenario as closely as possible to reach ecological validity. For gaming, this requirement would make interactive tests necessary to reflect the interactive usage situation (playing a game) the player will be in. Further, a realistic interaction experience will only be reached if the experience lasts for a certain period of time. Thus, ideally, gaming experience evaluation would require test users to play games and rate their resulting experience after having played one or several games for the duration of a typical game. Unfortunately, such interactive tests come with several disadvantages. First, test participants are subject

to fatigue, and thus lengthy gaming interactions will strongly limit the number of test conditions that can be evaluated in one test session. Extending a test over a number of test sessions would help to fight fatigue, but would render a direct comparison of test conditions difficult, as test participants can be expected to dynamically change their playing and rating behavior. Second, the concentration required to play a game will make it difficult for test participants to concentrate on certain aspects of a game, which might be the focus of the evaluation. For example, flow and immersion will require an interactive experience of a certain duration to evolve during the test, but this may make it difficult for test participants to concentrate on the video and audio quality of the game scenes. If the latter quality aspects are of interest (e.g., to compare the effect of frame rate or other video coding parameters), it may be advantageous to evaluate short sequences of audio-visual material which is typical for a gaming session in a passive paradigm. Thus, the ITU-T Rec. P.809, based on the research presented in Section 3.2, recommends the use of the two test paradigms, passive tests with audio-visual stimuli as well as interactive tests with game scenarios.

In Fig. 3.4 the general test structure of interactive tests is shown which will be described in more detail in the following. It must be mentioned that, unless the target of the research requires it, multi-user tests should be avoided to prevent the impact of social influencing factors. Thus, the structure only considers the evaluation of a single user using a cloud gaming service.

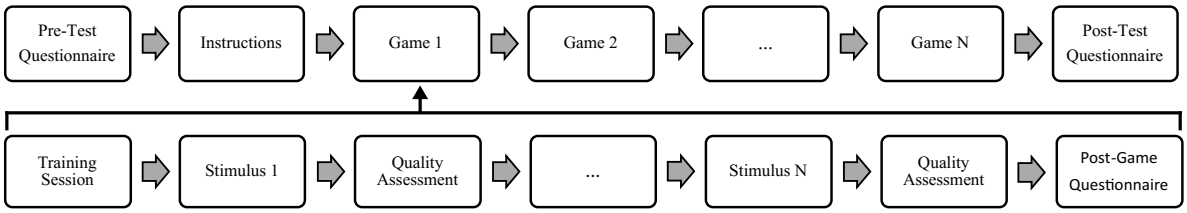


Figure 3.4: Schema of study design for interactive tests [24].

At the start of a subjective test, participants are asked to fill out a pre-test questionnaire in which user factors are assessed. This is followed by an introduction of the test participants regarding the procedure of the test, measurement instruments, and rating task. Also, prior to a test, participants should be screened for normal visual acuity or corrected-to-normal acuity and for normal color vision. In case audible stimuli will be presented, appropriate pre-screening procedures like audiometric tests should be selected. Due to practical reasons of setting up the interactive test stimulus, i.e., navigating to the game scenario under test, and also to reduce the load on participants, it is advisable to test all conditions of interest for one game at a time. While it is required to use multiple games if the research targets the investigation of the content, in the scope of the ITU-T G.1072 database only one game was tested in each subjective test as 17 different conditions were tested. In each game block, a training scenario under the best possible conditions should be performed to make sure the participants understand the rules and mechanics of a game and are able to control it properly as this is a requirement for the judgement process. A duration of 5 minutes should be considered sufficient unless a participant is already very familiar with the selected game. Next, a stimulus is presented to the participant who rates the gaming experience using a digitally presented questionnaire. This process is repeated till all conditions are tested. Within a game block, the conditions should be randomized to reduce learning and order effects. Therefore, a Latin square can be applied. At the end of each game block, a post-game questionnaire can be used to ask about quality aspects linked to the game *per se* such as aesthetics or learnability. Once the same procedure is used for the remaining games, a post-test questionnaire is used to ask

about the possible issues during the test as well as about which quality aspects were considered for the overall gaming QoE judgement. At the end of a subjective test, a monetary compensation is paid to participants.

For interactive tests with game scenarios, the ITU-T Rec. P.809 distinguishes between short interactive tests, in which a typical stimulus (interactive gameplay) length is between 90-120 seconds, and in which it is possible to assess the interaction quality (e.g., the impact of delay on the control), and long interactive tests with a duration of 10-15 minutes to ensure that players get emotionally attached to a game scenario, e.g., for measuring affect or flow. Especially with long test sequences care should be taken that participants do not get tired, and thus their reduced cognitive state may influence test results. In the scope of the development of the dataset used for the ITU-T Rec. G.1072, short interactive tests were carried out as the focus was not on user engagement. In contrast to the ITU-T Rec. P.910, a stimulus duration of 30 seconds for the passive-and-viewing tests was used as recommended in the ITU-T Rec. P.809. While for the latter test paradigm, the total rating duration, i.e., watching and rating game video scenes, should not exceed 60 minutes, a slightly longer time might be used for an interactive test as they usually engage participants more and are also more enjoyable. However, for both paradigms a break should be included in the middle of the test and whenever a participant requires it. In their research presented in [200], Schatz et al. argue that for comparable QoE lab user experiments, a limit of 90 minutes should not be exceeded in order to achieve a good balance between quantity and quality of results. Finally, the structure of a passive test is very similar to the interactive tests. However, the block design per game as well as training scenarios of the games do not apply there.

Test Setup

For the interactive tests, the following test setup is used. The structure of the setup, which requires the following elements, is shown in Fig. 3.5 and contains: a) a server PC fulfilling the hardware recommendations as listed by the developer of the used game, b) a client PC capable of playing the video stream smoothly, c) a local network for controlled streaming conditions, and d) peripheral components such as monitors, headset, and input devices.

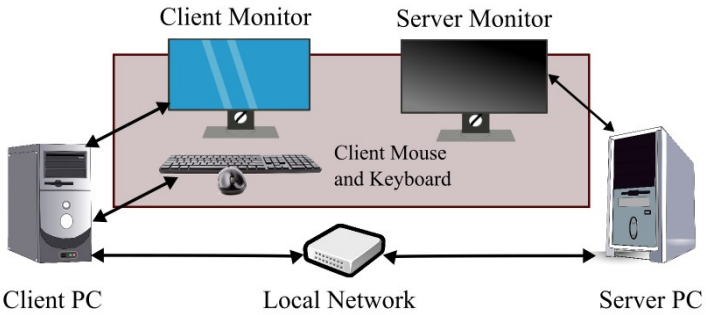


Figure 3.5: Setup for interactive subjective tests [26].

For the subjective tests, the light conditions in the test rooms, its acoustical properties, as well as the viewing distance (which is ideally three-times as large as the video height), and position of participants should be as consistent as possible for all experiments. While in general, the ITU-T Recommendations P.910 and P.911 should be followed also for gaming QoE assessments, some specifics to gaming research will be highlighted in the following sections. For both test paradigms described in the previous section, ideally the same client monitor should be used. However, unless it is part of the parameter

under investigation, for the video presentation a 24" LCD monitor with HD1080 resolution, a refresh rate of highest targeted encoding framerate, and no G-Sync/Free-Sync should be used.

For the subjective tests specifically carried out for the development of the ITU-T Rec. G.1072, as a server, a Windows PC running on an Intel®Core™ i7-7700K (4.2 GHz) processor, 16 GB RAM and an Nvidia GeForce GTX 1070 graphics card was used. As a client, a Windows PC using an Intel®Core™ i5-4460 (3.4 GHz) processor, 32GB RAM and an Nvidia GeForce GTX 960 graphics card was selected. Lastly, an ASUS VG248QE (24 inches) gaming monitor was connected to the client in addition to a standard keyboard and HoLife gaming mouse (up to 7200 DPI). As a cloud gaming application, the Steam Remote Play service, the former Steam In-Home Streaming, was used in a local network environment. The console of the software was used to change the encoding settings for the video streaming. To change the encoding or network conditions, a script controlling all settings on the client PC was implemented using AutoHotkey[3], a free, open-source custom scripting language for Microsoft Windows. As test environment, lab rooms offering high control about the setup and events during a test were selected, adhering to ITU-T Rec. P.910. In addition to the components listed above, the local network was controlled using Linux's *NetEm* network emulator kernel module [201] running on an Ubuntu system. NetEm enables the simulation of networks and packet losses. For the created dataset only a constant delay and uniform packet loss were used, which does not represent real networks sufficiently, but strongly reduced the parameter space to be tested. To allow communication between the client PC and the Ubuntu system, the AutoHotkey script executed a Python code using simple post and send messages. Using such an automated way of setting up the conditions of the test stimulus avoids disturbing the participants and avoids human errors during the setup. The AutoHotkey script also provided a clear indicator, when the stimulus starts as participants had to press a hotkey to disable a black screen, and also when the stimulus should be rated due to a message appearing on the screen. The screen was faded out to a gray background in parallel to muting the sound to avoid an abrupt interruption of the experience which could possibly also influence ratings related to concepts such as flow. Lastly, the script was used to record the input commands such as mouse clicks and keystrokes to derive statistics such as the APM for possible future analysis.

For passive tests, the videos were presented directly within a digital questionnaire tool used for the rating process. The questionnaire is a self-designed JavaScript website that used some components of a tool called TheFragebogen, a web browser-based questionnaire framework for scientific research [202].

Participant Requirements

As illustrated in Chapter 2, gaming QoE is subject to a variety of possible user factors. While it is not the aim of the presented research to accurately predict which user types prefer certain games or which game elements will result in a higher QoE, there are certain characteristics of participants that should be controlled and instructions for the participants should aim to maximize reliability, validity, and objectivity of results. For each subjective test, about 25 participants should be invited to allow a statistical data analysis. Ideally, a gender balance or a maximum of 60-40 split should be targeted to represent the target group sufficiently well. Furthermore, participants should fulfill the following criteria:

- Participants should have played video games using the input device used in the test before and judge themselves to be capable of controlling a video game properly.

[3] https://www.autohotkey.com/

- Participants should have normal (20/30) visual acuity with or without corrective glasses (per Snellen test) and normal color vision (per Ishihara test).
- Participants should have a fair level of language skills (at or above B2 level by Common European Framework of Reference for Languages) to answer the questionnaires.
- Participants should not have relevant neurological diseases (e.g., epilepsy) or sensory-motor dysfunctions (e.g., movement disorder).

Participant Instructions

A summary of instructions for participants specific to gaming research will be given below, whereas more detailed information can be found in the ITU-T Rec. P.809.

- To ensure a similar user behavior among participants leading to more comparable results, specific tasks might be given to the test participants and they should be asked to sufficiently interact with the game.
- The rules and goals of the games as well as the questionnaires used should be explained before the first condition is tested.
- The difference between video quality and graphical quality must be explained (e.g., a very abstract game consisting of only a few blocks *per se* does not have a bad video quality but often a low graphical quality).
- Participants should be made familiar with games even in passive tests to avoid confusing game events with video artifacts.
- Apart from paying participants for their participation, an additional bonus payment can be given to the best performing players to keep them motivated and give more realistic meanings to perceived degradations.

Additionally, samples of the used conditions during the test should be shown to the participants before the test, but it must be avoided to indicate whether this is a good (or best) or bad (or worst) condition, as the ratings of participants should be based on their expectations and not on finding previously quantified degradations. Ratings of these sample conditions can be used as anchor conditions.

Game Material Selection

For the stimulus generation for passive tests, in general, there are two methods available to create audiovisual stimuli: encoding a reference video using different encoding parameters, or recording actual gameplay using different streaming parameters. If the encoding process of the evaluated cloud gaming service is known, the former represented the method with least effort but highest consistency. With respect to the interactive tests, the chosen scenario from the game must be representative of the game, i.e., avoid dull moments such as showing a menu for a longer period of time and cover a normal interactivity, and different test subjects will need to be able to repeatedly experience them in similar ways. This implies that the scenarios must not be too difficult for one player, but also not too easy for another player to cause a similar challenge to the players' abilities. Depending on the range of the players' skills, this might lead to conflicting requirements, which need to be addressed in an appropriate way. As an example, the difficulty of the game can be adjusted to meet a particular player's skill. The use of horror games or overly violent games should be avoided if possible, not only for ethical reasons but especially when physiological measurements are used to capture the user's state. If the

game content is not in the focus of research, participants can be allowed, which was also the case for the development of the dataset, to select their favorite game out of a list of available games. This often ensures that participants are enjoying the game, are more familiar with its rules and mechanics, and thus, represents also a more realistic use case.

The selection of game content is a complex challenge as there is still a lack of a consent on a suitable game classification which also takes the influence of network impairments or encoding settings into account. As the video quality and interaction quality have turned out to be very important quality aspects for gaming QoE, but are influenced by different influencing factors, it was decided to choose a set of games that covers a broad range of degradation possibilities regarding these three dimensions: encoding complexity, sensitivity towards delay, and sensitivity towards frame losses. Research about this content classification, which is also described in [17], [18], [28], [203] will be presented in Section 6.2 in more detail. The final selection of the used games will be presented in Section 7.1.

Parameters Under Investigation - Independent Variables

While the selection of the parameters under investigation is different for each research purpose, it is fundamental to keep all factors, which could potentially put an unintended influence on the user ratings, as constant as possible to allow reliable and valid results. This especially includes dynamic user factors and context factors as described before. To this end, the quality factors summarized in the ITU-T Rec. G.1032 [60] should be considered and reported to the best knowledge of researchers. Furthermore, it is advisable that the selected parameters, e.g., a network delay, should cover the full range of the rating scales. This was especially important for the created dataset for the ITU-T Rec. G.1072. Therefore, either expert judgements or pre-tests should be used to derive suitable parameter values. Additionally, if the test design allows, parameters should be used in combination, i.e., not only delay but mixed with encoding issues and frame losses for example, since otherwise, participants might only focus strongly on one aspect to derive their quality judgement. As cloud gaming and gaming QoE are highly complex, there is an almost infinite number of possible parameters to consider. Thus, the scope of the research (cf. Section 1.2) must be clearly described and motivated (e.g., excluding Head-mounted displays or the highest framerate considered) as this also explains the limitation of the work as well as its use case.

Measured Quality Aspects - Dependent Variables

At the time of the start of the presented research in this thesis, a comprehensive understanding of the importance of individual quality aspects such as video quality, flow, and immersion for the overall gaming QoE as well as the interplay of various aspects was not available. Even though some promising and scientifically sounds models of gaming satisfaction and enjoyment such as the GUESS, CEGE, and PXI were developed in the meantime, these models do not target cloud gaming services and, thus, are missing fundamental quality aspects such as input quality and audio-visual quality from a technical perspective. Instead, they rather focus on UX and are strongly game design-focused. Based on the overview of questionnaires presented in Section 3.1.2, a summary of various quality aspects, for which a measurement instrument is existing, is presented in Table 3.5. In summary, it can be seen that many tools cover game and motivation-related aspects such as escapism, competition, goals, and narrative elements. Furthermore, a strong focus of engagement concepts such as involvement, immersion, presence, and flow as well as social aspects such as relatedness and social presence can be observed. Concerning the PX aspects, a few concepts such as fairness and aggression are used, but

in general, a strong overlap exists with the aspects covered by the GEQ, which forms the basis of the gaming taxonomy presented in Chapter 2. Thus, no severe gaps in taxonomy were identified which is highly relevant for the RQ1.

Table 3.5: Overview of quality aspects based on available questionnaires.

User-related	Game-related	Engagement	Social Aspects	Player Experience
intrinsic motivation	**system**	**engagement**	empathy	positive affect
interest	interaction quality	cognitive involvement	cooperation	negative affect
dissociation	immediate feedback	emotional involvement	socialisation	tension
distraction	technical quality	immersion	group identification	challenge
diversion	service mechanisms	absorption	team identification	competence
escapism	task completion	attention	companionship	enjoyment
exploration	image characteristics	focus	relatedness	satisfaction
fantasy satisfaction	resolution	**flow**	ownership	fun
curiosity	auditory	challenge-skill balance	co-presence	autonomy
discovery	haptics	action-awareness merging	social presence	control
excitement	**game**	clear goals	embodied presence	fairness
stress relief	game control	unambiguous feedback	identification	intensity
competition	game help	concentration		valence
development	game identification	sense of control		arousal
reward	game play	transformation of time		tiredness
achievement	game goals	autotelic experience		aggression
skill development	playfulness	loss of self-consciousness		annoyance
strategy	interactivity	**presence**		boredom
extrinsic motivation	story-telling	awareness of surroundings		frustration
incentives	realism	physical presence		
purchase intentions	learnability	returning to reality		
intention to (future) play	interface quality	spatial presence		
willingness to recommend	ease of use	transportation		

Consequently, it was decided to find a concise way of measuring the majority of the quality aspects covered in the taxonomy. However, it can be assumed that some quality aspects such as aesthetics are rather game dependent than influenced by transmission and encoding parameters. Thus, such aspects can be assessed for each game by the post-game questionnaire instead of after each stimulus during a subjective test. Lastly, a modular approach is favorable as some parts of the full measurement method could be replaced by improved methods in the future.

In the following, the final selection of the measurement tools will be presented.

Pre-test Questionnaire

The pre-test questionnaire assessed the following user factors: year of birth, gender, time spend playing per week, frequency of playing per week, self-judgement of gaming expertise, desire to play, typical game devices and monitor size used, and experience with the game under test.

Post-game Questionnaire

The post-game questionnaire covers the following aspects: Performance indication (PI), Learnability (LE), Appeal (AP), and Intuitive controls (IC). Each factor is assessed by a 7-point continuous scale and 3 items per factor, which were partially adapted to the gaming domain. The development of the PI items will be explained in detail in Chapter 5. The items for learnability were derived from the Perceived Ease of Use factor of the Cognitive Absorption Scale (CAS) by Agarwal and Karahanna

[115]. The User Engagement Scale by O'Brien et al. [128] was used to cover appeal. Lastly, intuitive controls was measured using the items of the PENS questionnaire [204].

Post-condition Questionnaire

The questionnaire to assess the gaming QoE after each stimulus first started with the overall gaming QoE. Here, the 7-point EC-ACR scale proposed in ITU-T Rec. P.851 and P.809 was used. Next, participants were asked to indicate how they felt while playing the game scenario for each of the following quality aspects measured again on the 7-point EC-ACR scale:

- Input Quality: Controllability (CN), Responsiveness (RE), Immediate Feedback (IF)
- Output Quality: Audio Quality (AQ), Video Quality (VQ), Video Fragmentation (VF), Video Unclearness (VU), Video Discontinuity (VD), Suboptimal Video Luminosity (VL)
- Player Experience: Immersion (IM), Competency (CO), Negative Affect (NA), Flow (FL), Tension (TE), Positive Affect (PA), Challenge (CH)
- Self-judgement of Playing Performance (PR), and Service Acceptability (AC)

For the video quality features VF, VU, VD, and VL, the ITU-T Rec. P.918 was followed. However, it must be noted that the proposed dimension Noisiness was excluded as this type of artifact is not common for cloud gaming services. For the PX, the iGEQ was used. Despite the criticism regarding the long GEQ, the iGEQ was considered to be an appropriate tool as pointed out earlier. Lastly, as no validated questionnaire to assess the input quality was identified, the concepts controllability, responsiveness, and immediate feedback were used for its measurement. The questionnaire was developed within the frame of the presented research. More details can be found in Chapter 5. A full list of the items and scales used in each questionnaire is accessible in Appendix B.

3.4 Summary

This chapter provided an introduction to assessment methods categorized into behavioral, psycho-physiological, as well as qualitative and quantitative subjective assessments. While behavioral and psycho-physiological methods promise real-time monitoring of the state of participants and higher ecological validity, they often require specific behavioral responses that might negatively influence a natural interaction with the cloud gaming services or lack generalizability. Subjective methods, on the other hand, rely on post-memory judgements and interrupt the interaction with the service. While a multi-method approach appears to be of high value, more extensive research is required to understand biases inherent to each method as well as to reduce inter-individual differences of players. Thus, within the scope of this dissertation, the focus will be quantitative subjective methods, which will allow an empirical investigation of the cloud gaming taxonomy.

Even though many methods for traditional services such as telephony or web browsing are available, knowledge about methods dedicated to the evaluation of cloud gaming services is rare. While there is a wide range of methods provided by the UX community, many of these approaches such as heuristics or game usability methods focus on the design of games and leave out the control and video stream of cloud gaming services. Chapter 3 contributed an extensive list of available questionnaires to assess player experience aspects such as immersion. However, there appears to be no solution to cover the full

spectrum of quality aspects covered by the taxonomy in a concise way. This is caused by the fact that many questionnaires are too long and in case of their combination, cover many overlapping constructs. Furthermore, the lack of validated tools specifically for cloud gaming leads to the circumstance that many researchers are using either self-constructed questionnaires or adapt standardized questionnaires from other domains. This is particularly true for one of the important aspects, the input quality, for which no validated methods for its assessment could be identified. Thus, Chapter 5 will present the development of a new scale to measure the input quality of cloud gaming services.

Furthermore, Chapter 3 provided insights to standardization activities which allow researchers to carry out more reliable, valid, and comparable empirical studies. The lessons learned from the literature, gaming experts, and conducted subjective tests led to the ITU-T Rec. P.809 about gaming QoE assessment methods. For the assessment of the full range of quality aspects covered by the cloud gaming taxonomy, a concise modular approach was provided. The method combines the ITU-T Rec. P.809 with respect to the design of subjective gaming QoE tests with the ITU-T Rec. P.918 for the assessment of video quality dimensions, as well as the questionnaire measuring input quality (cf. Chapter 5). Whereas the aspects appeal and playing quality will be on assessed during a post-game questionnaire, the iGEQ will be used as a tool to assess the PX after each test condition.

Regarding the planning of subjective tests, on the example of a test design for the ITU-T Rec. G.1072, all gaming-specific considerations were summarized. Here, knowledge about important influencing factors summarized in the ITU-T Rec. G.1032 was considered, and a split into interactive and passive test was performed. In a study, it was shown that passive tests, if they follow appropriate participant instructions and stimulus duration, can deliver comparable results to interactive tests for the assessment of video quality. However, for the assessment of the PX, the results emphasize the need to carry out interactive tests. Thus, it is proposed to assess the large variety of encoding parameters in short passive tests (30 seconds), while the input quality and PX should be assessed in interactive tests (90 seconds) that focus especially on network parameters.

Chapter 4

Crowdsourcing for Gaming QoE Assessment

For many research purposes, there is an interest in gathering a large amount of data in a short time frame of a demographically diverse audience. To assess the QoE of multimedia services, traditionally lab studies are conducted. While this offers a controlled environment, these experiments are often time-consuming and expensive. Therefore, the method of Crowdsourcing (CS) has become very popular in recent years. Participants of such tests, referred to as (crowd) workers, will typically be recruited via platforms such as Amazon Mechanical Turk (MTurk)[1] or Microworkers[2], and will solve short Human Intelligence Tasks (HITs) compensated with monetary reward. CS can be used to debug applications, to gather data about network connections and localization data, and for labeling tasks, but it recently gained also attention for the quality assessment of diverse media contents such as speech, audio, and video quality [205]–[209]. As within the scope of the present work, the CS approach will be applied to the gaming domain, it should be referred to as Crowdgaming (CG) in the following.

Apart from the fact that the CG approach is in general highly interesting from a methodologically perspective due to the advantages mentioned above, the method will be used for the present research for two purposes: a) the development of a questionnaire measuring input quality described in Chapter 5, and b) the investigation of the mediation of game characteristics on the impact of network impairments on gaming QoE reported in [18], [210], [211] as well as Section 6.1 and 6.2.

However, obtaining valid and reliable results is very challenging as there is no direct contact with the crowd workers, thus, it strongly depends on the purpose and content of the experiment. While CS is a very promising method, it also faces several challenges regarding the test design, reliability of users, incentives and payment schemes to motivate users, hidden influencing factors in the uncontrolled environment, and statistical analysis of the results [212].

In recent research activities targeting media quality assessment, workers participate in subjective tests from their own working environment while using their own hardware which differs from the hardware used in controlled lab studies [208], [209]. This approach provides higher ecological validity as the situation is more realistic than the lab environment. However, internal validity is often endangered as the CS method is vulnerable to effects of uncontrolled influencing factors. To overcome the multitude of challenges to conduct CS tests offering reliable and valid results, a variety of influencing factors and methods for media quality assessment have been investigated, and different guidelines were provided

[1] https://www.mturk.com
[2] https://www.microworkers.com

in the last years [208], [212], [213]. Among other insightful research such as [206], [207], [214], the authors of [207] investigated the influence of trapping questions on the reliability of collected crowd data. They suggest to emphasize to workers that high-quality responses are very important for their research and asked them to select a specific item to show their concentration. The lessons learned from recent work led to the ITU-T Rec. P.808 on the use of CS for subjective evaluation of speech quality. The recommendation describes the creation of test materials, experimental designs, and the procedure for conducting listening tests in the crowd, as well as how to report the results. In addition to ITU-T Rec. P.808, the ITU-T offers a technical report [213] on subjective evaluation of media quality using micro-task CS. The document focuses more on general aspects of CS and lists influencing factors, describes the experimental design and test procedure, and gives a brief overview of required statistical analysis. With respect to crowdsourced quality assessment of gaming applications, there have been only a few researches carried out. In [215] and [216] a few suggestions on a CS approach for online gaming tests are given. The authors of [217] present a CS game platform that can be used to create and share simple games, and collect data for different purposes. Lastly, in [218] a crowdsourcing-based approach to objectively assess the impact of game impairments on the player performance was investigated. However, for the dissemination of the used game, DotA2, the well-known platform Steam was used. While this offers some benefits such as additional user statistics, the research was limited to an already existing and thus not open-source version of a game. Furthermore, a thorough search of the relevant literature did not yield any work on simulating network impairments which is a fundamental need for the development of the previously mentioned questionnaire in the next chapter.

Thus, in the following section a newly developed CG evaluation method based on the recently published ITU-T Recommendations P.808 [212] and P.809. [96] will be described. The presented work targets the RQ4 about an alternative to traditional lab studies for gaming QoE assessment. The method will be tested in terms of expected ratings of gaming QoE features in a total of six studies investigating the impact of network and encoding parameters, namely delay, packet loss, and framerate, as well as changes in the game design on gaming QoE using a CS approach. Steps to investigate appropriate participation of workers, to increase their motivation to focus on the rating task, and how to control typically considered system influencing factors for gaming research will be highlighted. Finally, the validity of CS data for gaming QoE assessment by comparing lab and CS test results will be examined. The work is currently a part of the ITU-T work item P.CrowdG which aims to give guidance about the subjective evaluation of gaming quality with a CS approach. Additionally, the framework and some study results are published in [16].

4.1 Development of Crowdgaming Framework

In this section, all components of the interactive CG test framework, as shown in Fig. 4.1, will be described. Also, recommendations on how to gather reliable and valid results despite the absence of an experimenter, controlled network, and visual observation of test participants will be given.

Game Implementation

The game is a central aspect of an interactive gaming study. However, it is not only the stimulus to investigate, it can also be the bridge between the user, server, and CS platform. As the game is an important influencing factor on gaming QoE, a total of six web-compatible JavaScript games were developed and modified to fulfill the needs of the integration to a CG test. The games are hosted on

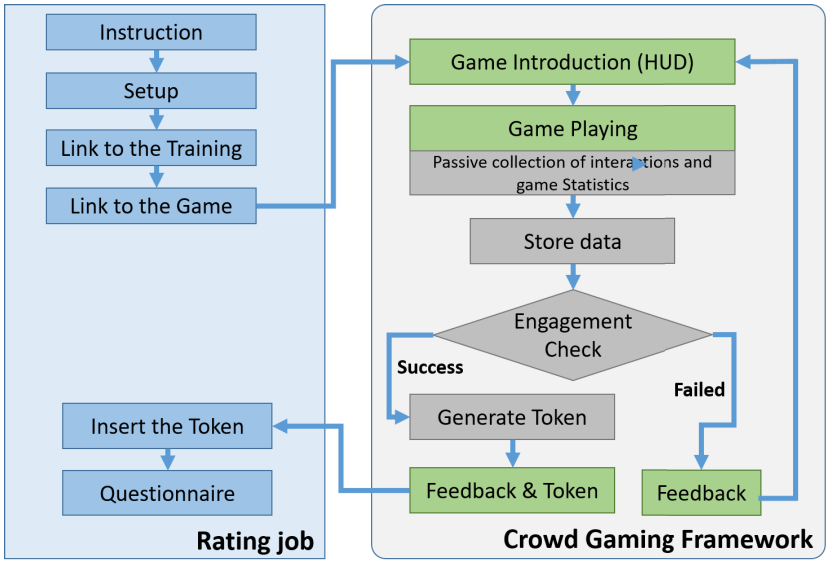


Figure 4.1: Components of CG Framework (cf. [16]).

a web server, which workers can access via a Uniform resource locator (URL) available on a crowd platform. For the development of the games, the p5.js library[3], which offers a set of drawing functions and add-ons for interaction with other HTML5 objects, was used. An alternative would be the use of the cross-platform game engine Unity to create a WebGL game. The open-source nature of this approach offers customization of the game to achieve several important advantages, which will be described in the following.

It must be noted that the developed games are of a rather abstract nature as for the purpose of this research, only fundamental concepts of games and typical interactions are required. However, this is no strong limitation of the method *per se*. As a basis for a majority of the games, the "A game a day" project by Kael Kirk[4] was used. The games had to be modified to fit the aspect ratio of the browser window, animated mouse cursors were added, reward systems were improved, a restart in case of losing the games was added, the network impairment simulations were added using buffers or random generators, and the communication with the web-server including the generation of tokens and logging information was implemented.

The final game dataset consists of the games *Dodge, GTA, Shooting Range, Flappy Bird, Rocket Escape,* and *T-Rex*. While in Dodge (dexterity game) and T-Rex (jump and run game) obstacles have to be avoided by well-timed keystrokes, Rocket Escape (racing game) and Flappy Bird (arcade game) require frequent player input to balance the position of the character. Finally, GTA and Shooting Range (both action/shooting games) require in addition spatially accurate mouse inputs. Two screenshots showing the games T-Rex and Shooting Range are shown in Fig. 4.2 and 4.3, respectively.

Game Introduction

The ITU Rec. P.809 suggests that players must learn the controls and rules of the game before rating the first condition. Therefore, workers had to pass a training session. Before each game scenario, a screenshot of the game with labeled Heads-up display (HUD) (e.g., timers, scores) and game elements (e.g., characters and targets), controls, and a description of the rules and goals of the game was shown.

[3]https://p5js.org/
[4]https://github.com/Kaelinator/AGAD

Figure 4.2: Screenshot of jump and run game called T-Rex.

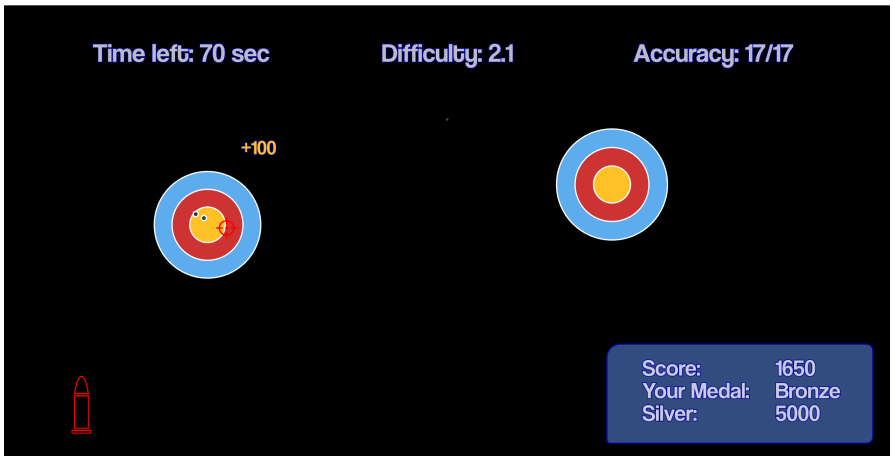


Figure 4.3: Screenshot of shooting game Shooting Range.

Token System

Per se, there is no information available to find out if someone who is playing a game on the webserver is also participating in the survey on the CS platform. For this reason, a 36 character long universally unique identifier (UUID) was generated after each gaming session and served as a token. The token was stored in server logs among other information, and workers were asked to copy this token and paste it back to the survey at the beginning of the rating process. If a valid token was used, the rating scale was shown. This method ensured that workers really played the game until the end. It also enables us to know which information stored on the server belongs to which worker ratings. While well-versed workers may figure out the method of the token creation, a mismatch of a potentially manipulated token and those stored on the server would lead to discarding the ratings of such workers.

Stimulus Generation

Essential for conducting an interactive gaming QoE assessment test is the generation of a stimulus. While from a technical point of view, the content of a game does not play an important role in conducting a CG test, there are a few aspects that should be considered. Firstly, the duration of a stimulus should be limited and clearly indicated to participants. A timer of the remaining playing time was added to the HUD of each game. Secondly, an automated restarting of the game scenario should be implemented in

case of a defeat of the player. Thirdly, an open-source game offers means to artificially add network impairments or encoding artifacts to a game. A network delay can easily be simulated by buffering input commands, input packet loss can be simulated by using a random number generator and discarding functions called for input events, and even frame rates can be changed by skipping the drawing function, which is called every frame. Using this approach, degradations can be artificially simulated without manipulating and controlling the network conditions of workers. However, it has to be noted that these degradations might have small differences to the real end-to-end delay or packet loss, they should be carefully designed similar to the real scenario.

Game Design

Especially for fundamental research, it is highly beneficial to be able to change the design of a game. Different methods of controlling a game, interface design, balancing, or characteristics such as the pace or predictability, which may influence the impact of a network delay on gaming QoE, can be investigated. Furthermore, information of the game state such as performance indicators, which can be added to the HUD, and logs of the player inputs can be generated.

Engagement Check

One challenge to overcome in a CG test is to find out whether a worker played a game scenario as intended. While in a lab study this can be observed visually by the experimenter, information generated by interacting with the game can be used in the crowd. Therefore, in the developed CG framework an engagement check at the end of each stimulus is implemented. During a pre-test, the number of inputs, i.e., mouse clicks or keystrokes, for each game during the most strongly impaired condition, e.g., a delay of 300 ms, were analyzed to derive an activity threshold. It is also possible to derive such a threshold by an expert judgement. Workers passed the engagement check if their number of inputs was higher than 20 percent of the typical number of inputs derived from the pre-test, scaled by the ratio of stimulus duration and duration of pre-testing. If workers failed this check, they were told that they did not put enough attention to the game and were asked to play the condition again. The 20 percent rule was established based on the feedback of workers who failed the engagement test even though they mentioned a strong focus on the task. Not only does this method prevent workers from cheating, but it also is of high value for the training session to make sure workers understood the rules and controls of the game in the short amount of time available. If knowledge about typically reached performance values such as points is available, also such information could be used in addition to the input information.

Crowdsourcing Workflow

The following steps are adapted from ITU-T Rec. P.808 to design the CG tests. For the six studies carried out in the scope of the method development, MTurk was used as the CS platform as it is most widely used and offers a pre-selection of workers with diverse backgrounds, English speaking workers, dynamic content creation, and easy payment of participants.

In the following, the procedure of the CG approach will be explained.

HIT Recruitment

Depending on the purpose of the study, it may be beneficial to select a specific target group for the study. Therefore, a screening HIT can be published before the actual CG test. Here, aspects such as age, gender, playing frequency, gaming skills, as well as game and device preferences can be assessed (cf., ITU-T Rec. P.809) to create a user profile. If a profile is suitable for the research, the worker can later be invited to participate in the test based on the profile, which also contains the worker ID. For the conducted tests, the most important criteria were that workers like to play video games and that they can control them sufficiently. Additionally, some platforms offer worker profiles based on a variety of characteristics. Only workers who fulfilled the following three criteria were recruited: their location is in the United States, their HIT approval rate is over 98 percent, and their number of approved HITs is greater than 500 (cf. ITU-T Rec. P.808).

HIT Requirements

Every HIT started with a summary of the requirements. Workers were asked to only participate in the test if they fulfill the following requirements: a) they should have played video games in the past year, b) they should be interested in playing video games, c) they are using a desktop (PC) or a laptop for the job, d) their device has a keyboard and mouse connected, e) their device is connected with power, f) their device must be able to play stereo sound.

HIT Instruction

The procedure of the test, what is expected from the workers, and how to use the rating scale should be explained to the participants using short and clear sentences for each step. In the instruction of the performed tests, it was explained that they will play different simple game scenarios and rate their experience after each scenario and it was recommended to use a modern web browser for the test. Next, it was clearly stated that responses will be used for scientific research and that especially the questionnaire should be treated very seriously. Afterward, the estimated total duration of the HIT, the duration of each scenario, and the structure of the HIT, which was split into several sections, were listed. The structure of the crowdsourcing survey was also shown visually, as it can be seen in Fig. 4.4. In case of unforeseen issues, workers were offered to contact the experimenter via a given Email address. As an alternative, a forum post could be used. Additionally, a few hints were given to the workers. It was explained that they should keep the MTurk browser tab always open, what the engagement check after each game is, and that the submit button will only be visible if all questions are answered.

Figure 4.4: Structure of the CG test for three stimuli.

Questionnaire Instructions

For the test, the 7-point EC-ACR scale as recommended in ITU-T Rec. P.809 for the assessment of gaming QoE was used. For consistency, the scale was also used for the remaining items. The usage of the scale, especially concerning the overflow area, was explained to the participants in the introduction section. Furthermore, it was mentioned that it may happen that the quality of a scenario is not ideal, and this is intended and not a bug in the system. More information about the used questionnaire items and scales is given in section 3.3.2 and Chapter 5.

Payment

The ITU-T Rec. 808 suggests that the presence of quality control systems and conditions in which their answers would be rejected or selected for extra bonuses should be clearly explained. Thus, once more, workers were told that the quality of their answers to the questionnaire is highly important. With increased visibility (bold red text) it was stated that if it can be proved based on the analysis of internal consistency and comparison with ratings of other workers and references that workers did not answer the questionnaire properly, that they will not get paid for the HIT. As suggested in ITU-T Rec. P.808, to increase the motivation of workers to perform well, a bonus was advertised for the top 10 percent of players who also give valid answers (cf. Section 4.2.3) to the questionnaire.

Sound Check

As some games contained sound effects, the first task of a HIT was an easy soundcheck. A stereo audio file in which a speaker reads a simple mathematical equation, e.g., four plus one plus two, was included in the HIT. If the answer was given correctly (by word or number), workers were able to progress.

Worker Survey

The second section in the HIT was a short demographic questionnaire like the one in the recruitment HIT. The questions were derived from the pre-test questionnaire recommendations in ITU-T Rec. P.809.

Training

As suggested in ITU-T Rec. P.808 and P.809, before the first stimulus, a training scenario was presented. Here, workers learned the rules and controls of the game. The duration was set to 30 seconds, and a token was generated at the end of the scenarios if a participant passed the engagement check. Workers had to paste this token into the survey to proceed.

Rating Section

For each stimulus, workers were asked to play a game scenario by following a given link and then copy the verification code that appears at the end of the scenario and use it for the HIT. If the worker passed the token check, the rating section became visible. In the rating section, workers were asked to indicate how much they agree or disagree with each of the following statements by clicking on the 7-point scale below, as explained in the introduction. A dynamically generated slider provided workers always with a single item to prevent them from getting biased by their previous ratings. Once an answer was given, the next question was automatically shown. The questionnaire was implemented by adapting a software

framework for building questionnaires for the web called TheFragebogen [202]. Once all answers were given, the next game block was made available by clicking a continue button.

Quality Control

It may happen that a worker, despite the clear instructions, takes the rating process lightly or even attempts to cheat. Therefore, trapping questions (also known as gold standard questions) and consistency checks to the questionnaire [205] should be embedded. In a test using a 26-item questionnaire, three trapping and two repeated questions (consistency check items) were used. It should be avoided to add too many of these as it may show distrust to workers. The position of the additional questions was randomly assigned but kept the same for each condition. For each condition, three different types of trapping questions were randomly assigned and kept the same for each condition used: (1) obvious questions, (2) questions related to the current activity, and (3) a question related to the played game. While the first kind should be a clear sign that a quality control is embedded in the questionnaire as it was told, the latter can most likely only be answered correctly with proper attention. Table 4.1 contains the trapping questions, their type, and shows the range of accepted answers on the used 7-point Likert scale.

Table 4.1: Trapping questions (TQ) used in the questionnaire. Participants answered them using a 7-point EC-ACR scale ranging from 1 to 7 whereby a 2 corresponds to "stongly disagree" and 6 to "strongly agree".

TQ	Type	Item	Accepted Response
1	1	Please select the answer ”disagree” on the scale below.	[2.8 - 3.2]
2	2	In the game I played, I saw many colorful flowers.	[1.0 - 3.0]
3	3	The game I played is a typical ”card game” such as Solitaire.	[1.0 - 3.0]
4	1	Please select the answer ”agree” on the scale below.	[4.8 - 5.2]
5	2	Right now, I am reading a newspaper article.	[1.0 - 3.0]
6	3	In the game I played, I was able to talk to other players.	[1.0 - 3.0]
7	1	Please select the answer ”strongly agree” on the scale below.	[5.8 - 6.2]
8	2	Right now, I am answering a survey in MTurk.	[5.0 - 7.0]
9	3	In the game I played, I created my own character.	[1.0 - 3.0]

Stimuli and Conditions

Regarding the duration of a test stimulus, the ITU-T Rec. P.809 suggesting a stimulus duration of 90 seconds was followed. However, a duration of 30 seconds was used for the training scenario. The average duration of a HIT was targeted to be around 15 minutes in order to avoid fatigue. Even though gaming might be a fun activity for workers, a test should not be designed to be much longer as workers are used to the short duration of other HITs, and to reduce the risk of a decreased rating quality due to potential distractions or fatigue. A randomized condition plan containing the URLs of the games was uploaded to the CS platform, and for each condition, the desired number of participants was recruited.

Closing

Before finishing the HIT, an additional post-game questionnaire was added to assess concepts such as learnability, progress information, intuitive controls, and aesthetics. Once all questions were answered, a submit button was made visible, workers were thanked for their effort and told that they will receive the payment latest within the next five days.

Web Server

Apart from providing access to the games, an API in the webserver was used to save logging information of each played condition. The following information was stored: identification number, date and time of server and worker, game identification code containing the game name and condition (e.g., 200 ms delay), game version, if the condition was a training scenario, if the engagement check was passed, statistics of the game, and an MD5 encrypted version of these statistics. The encryption of the stats was used to prevent cheating as theoretically workers could change the key-value pairs passed through the URL for a get-request. Regarding the game statistics, the following information was stored once the engagement check was passed at the end of a condition: playing duration, number of restarts, scores, game-specific objectives, input counter used for the engagement check, key events of keyboard and mouse with timestamp and position, workers operation system and browser, as well as the height and width of the browser window.

4.2 Testing the Crowdgaming Framework

4.2.1 Experimental Design

In this section, the experiment design of six conducted CG studies will be explained. For their implementation, the previously described framework was used. The main purpose of the conducted CG tests was to gather many ratings from different users for the development of a questionnaire assessing input quality. To generate enough variance in the data, the following aspects were taken into account: a) using a large and diverse user group which represents the target group of gaming services, b) using games that cover typical mechanics and rules of popular games, and c) using common technical impairments such as a network delay.

Thus, six studies were designed in the following manner. In Study 4.1, each participant played one of the six games under three different delay conditions: 0, 150, and 300 ms. In Study 4.2, each worker played three of the six games without any impairments. In Study 4.3, each participant played one of the six games under the following three input packet loss conditions: 0, 10, 30 % packet loss rate. This type of artificially generated packet loss would correspond to a discard of packets that are sent from the player's client of a cloud gaming service (or traditional online gaming) to the cloud server (or game server) without any concealment methods. In Study 4.4, different framerates (60, 30, 10 fps) were used as independent variables for one game of the game pool per test. In Study 4.5, for either the game T-Rex or Shooting Range, different feedback elements were changed in the game design. The feedback included a) only visual feedback, b) visual and auditive feedback, and c) additional feedback about the game progress, e.g., points and medal systems. Lastly, in Study 4.6 the same games and feedback types of Study 4.5 were combined with the delay conditions used in Study 4.1.

As suggested by ITU-T Rec. P.808 [212], a balanced blocks experimental design was used for each study. This design, having one-between (the used game) and one-within subjects factor (simulated network impairment or feedback type), is often called a split-plot design [219]. Each study started with a training session followed by three test stimuli. The order of the stimuli was randomized. A visual overview of the six studies can be found in Table 4.2.

After each scenario, workers answered a pool of items assessing first the overall gaming QoE using the item proposed in ITU-T P.809 followed by 26 items measuring the concepts responsiveness,

Table 4.2: Conditions used for the six CG studies (game feedback types used: A = auditiv, V = visual, P = progress).

Study	Game	Delay [ms]	Packet Loss [%]	Framerate [fps]	Feedback
4.1	all six	0, 150, 300	0	60	V
4.2	all six	0	0	60	V
4.3	T-Rex, Shooting	0	0, 10, 30	60	AVP
4.4	T-Rex, Shooting	0	0	60, 30, 10	AVP
4.5	T-Rex, Shooting	0	0	60	V, AV, AVP
4.6	T-Rex, Shooting	0, 150, 300	0	60	V, AV, AVP

immediate feedback, and controllability, which represent quality features related to the input quality, as well as performance indication (cf. Section 3.3.2). These items are the initial item pool of the newly developed questionnaire described in Chapter 5, where also a full list of the used items can be found. As introduced in the framework overview, also three trapping questions, as well as two consistency questions, were added to each rating task in order to control the validity of provided ratings. For all items the 7-point EC-ACR scale introduced in Section 3.3 was used.

4.2.2 Demographic Information about Crowdworkers

In total, 571 workers participated in the tests, which resulted in 1713 ratings since each job consisted of three conditions. Each condition was rated by between 30 to 41 participants. A good gender-balance was reached as 245 females and 321 males participated in the tests. The majority of workers are in the age range of 26 to 50 years. More than 42% of the test participants are experienced gamers. In Table 4.3, demographic information of the workers is summarized in more detail.

Table 4.3: Demographic statistics of workers.

Gender	female	male	transgender	others	
	245	321	1	4	
Age [%]	18-25	26-35	36-50	50-68	
	12	50	31	7	
Gaming Experience [%]					
Beginner (1)	2	Intermediate (3)	4	Expert (5))	
4.9	10.4	41.5	30.4	12.8	
Hours per week spend on playing video games [%]					
0	0-1	1-5	5-10	10-20	>20
9.3	35.4	9.5	17.4	8.6	19.8
Device [%]	PC	Console	Smartphone	Others	
	50.4	31.9	15.9	1.8	

4.2.3 Data Cleansing

As suggested by ITU-T Rec. P.808, an experimenter should evaluate the submitted responses against unexpected patterns in ratings and user behavior in a session. Consequently, the experimenter may discard a response given in a session when unexpected user behavior is observed. All responses

submitted by a participant should be removed when these responses do not fulfill the abovementioned conditions more than twice.

For the data cleansing process, four data quality criteria are investigated: a) the number of wrongly answered trapping questions, b) an appropriate variance of ratings among all items, c) consistency between repeated items, and d) statistical outliers.

Thus, at first, the variance of ratings of all 81 items (27 items per condition) was calculated for each worker. The average variance of all workers was 16.54 (SD = 10.1). The data was ordered by the calculated variance over all items, and the variance of the first participant showing a difference between reference and worst condition of larger than 0.5 was taken as a threshold for the detection of variance issues. The threshold was derived to be 5.6. The reason for such a metric is that some workers might not focus on the given rating task and simply press the same answer to quickly finish the HIT. For 75 workers resulting in 225 ratings, a variance issue was detected.

Next, an Inconsistency Score (IS) was calculated. This measure investigates whether responses by workers are randomly given. The metric thus complements the variance investigation. IS is calculated using the normalized weighted Euclidean distance of the related consistency check items. The metric further considers problematic items (e.g., an ambiguously worded) by taking into account ratings of all workers and accordingly weighting the influence of each consistency check item pair. Lastly, a rejection threshold was derived by adding 1.5-times the interquartile range of the IS distribution to the 75-th percentile. For more information about the implementation of the IS, the reader is referred to [205]. Ratings of 28 workers (84 ratings) were discarded due to failure of the consistency check.

As a third metric for the data quality evaluation per worker, the trapping questions are analyzed. In total, 406 workers answered all 9 trapping questions correctly. Consequently, a great number of workers answer some trapping questions wrongly. Detailed information about the statistics can be found in Table A.2 in Appendix A showing the descriptive statistics of affected ratings. The analysis revealed that for workers who answered more than two trapping questions wrongly, many variance issues occurred. Complementary, workers who answered less than three trapping questions wrongly caused some consistency issues. As the ratings of workers who answered one trapping question wrongly still provided high-quality data in terms of variance and consistency, it was decided to keep the data of all workers who answered less than two trapping questions wrongly while removing the cases of variance and consistency issues. In total, ratings of 435 workers remained for further analysis.

Finally, an outlier detection using the outlier labeling method described in [220] was performed. As badly worded or incomprehensible items could potentially cause such outlier ratings, also statistics of the individual items are examined. However, no item with an unusually large number of outliers was detected. The average number of outliers for all 1305 ratings per item was 7.6 (min = 2, max = 16). Workers with more than three outliers for any of the 81 questionnaire items are removed. For the remaining cases, the individual ratings are declared as missing for the subsequent analysis. A total of 807 ratings was free of any outliers, whereas 213 ratings contained one item, 126 two items, and 42 ratings three items affected by an outlier. Consequently, the remaining data (1188 ratings) for each of the six studies comprises 417, 156, 135, 126, 156, 198 ratings for Study 4.1 to 4.6, respectively. The methodology and parts of the data analysis are published in [16]. However, Study 4.2 and 4.6 are not presented in the paper, and slightly different data cleansing criteria were used.

4.2.4 Study Results

In the following sections, the results of the studies will be summarized. It must be noted that for the calculation of the dependent variables controllability, responsiveness, immediate feedback, and performance indication, the mean value of the final questionnaire measuring input quality, which is explained in detail in the upcoming Chapter 5, will be used.

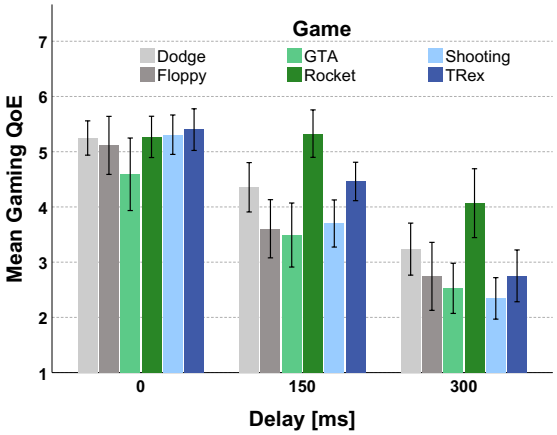
Apart from a visual representation of the data, observations are statistically investigated using a two-way Mixed ANOVA to examine the dependent variables corresponding to each study, e.g., using delay for Study 4.1, as a within-subject variable, and the game as a between-subject factor. To determine the difference between the games at each level of the dependent variable and vice versa, simple main effects are analyzed in case of evidence for an interaction effect. Even though for several cases a violation of homogeneity of covariances matrices was indicated by the Box's test, $p < .05$, the ANOVA results are considered to be valid as the ratio of the largest to the smallest sample size of investigated groups ($30/21 = 1.43$) is less than the threshold of 1.5 suggested by Pituch and Stevens [219]. For pairwise multiple comparisons, the Bonferroni correction was used. For probability values as well as confidence intervals shown in the corresponding bar plots, an alpha level of 5 percent was considered.

Study 4.1: Delay

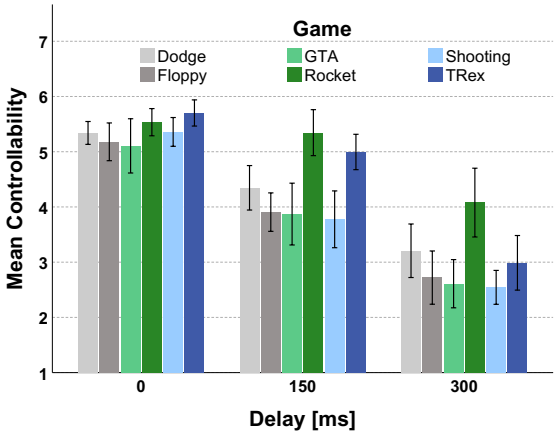
Playing a game with a delay of 150 to 300 ms delay, whose influence is not weakened by any delay compensation techniques, is expected to have a strongly negative influence on gaming QoE and aspects related to the input quality. Furthermore, the influence of delay was shown to be different for various game scenarios (cf. Section 2.2.2 and 6.1). It will now be analyzed if the collected subjective scores are in line with these findings, to judge if the used quality aspects are valid and reliable tools. In Fig. 4.5 the gaming QoE and input quality of the test conditions during Study 4.1 are illustrated.

For none of the games, a difference of ratings of the dependent variables can be seen at the reference condition, i.e., at 0 ms delay, when comparing the games with each other. Only a small reduction is visible for gaming QoE for the game GTA, possibly due to an inappropriate overlap of a game object and parts of the background image. Furthermore, as expected, a strong influence of the delay on the dependent variables can be observed. Also, the influence of delay was depending on the game which confirms findings presented in Section 2.2.2.

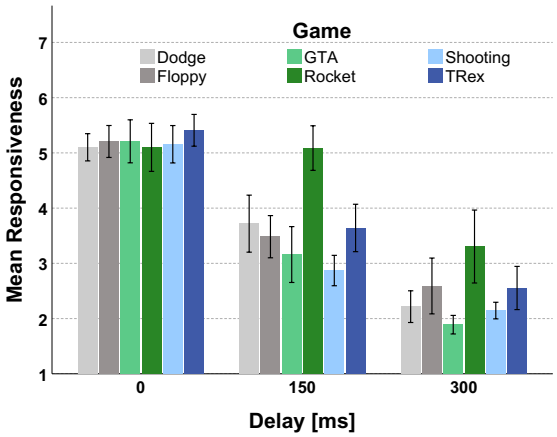
For the gaming QoE, the ANOVA yielded a significant interaction effect of game and delay, $F(10,266) = 3.82$, $p < .001$, $\eta_p^2 = .13$. The results show significant differences between the games at 150 ms, $F(5,133) = 9.50$, $p < .001$, $\eta_p^2 = .26$, as well as at 300 ms, $F(5,133) = 6.26$, $p < .001$, $\eta_p^2 = .20$. Regarding the 150 ms delay condition, the subjective ratings of gaming QoE for the GTA were significantly stronger impacted than for the jumping game T-Rex, $p = .041$, which was not the case for the 300 ms condition. The impairment of QoE ratings for T-Rex was not very strong at 150 ms as most likely the players could still jump over obstacles properly since the movements of those objects are very predictable. However, at 300 ms delay the QoE had a significant drop as the interval to react to the appearance of an obstacle was similar to the delay, leaving the players with barely any time to react. For both delay conditions, 150 ms and 300 ms, the QoE was rated significantly higher for the game Rocket Escape compared to all other games, $p < .001$, with the exception of T-Rex at 150 ms and Dodge at 300 ms. An explanation could be the fact that a player for Rocket Escape can correct potentially wrong or delayed inputs continuously, and would not directly lose as in the other games.



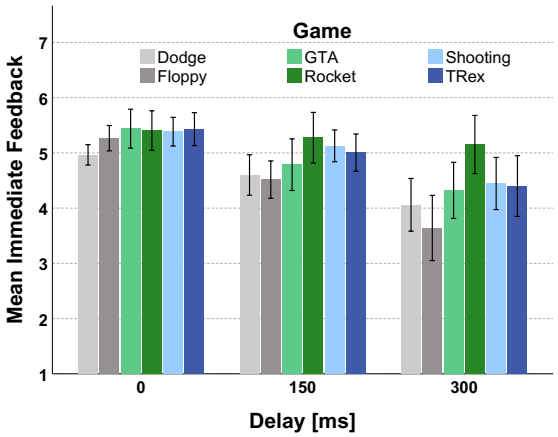
(a) Gaming QoE affected by input delay



(b) Controllability affected by input delay



(c) Responsiveness affected by input delay



(d) Immediate Feedback affected by input delay

Figure 4.5: Bar plots of means and 95 % confidence interval showing the impact of delay on gaming QoE and input quality during CG Study 4.1.

With respect to the influence of delay on QoE, statistically significant differences are found for every game. The game Rocket Escape, which was the only game that had no significantly reduced gaming QoE at the 150 ms condition, was the least affected game, Wilks's $\lambda = .84$, $F(2,132) = 13.00$, $p < .001$, $\eta_p^2 = .17$, whereas the game Shooting Range was the most affected game, Wilks's $\lambda = .55$, $F(2,132) = 54.94$, $p < .001$, $\eta_p^2 = .45$.

The ANOVA also revealed a significant interaction effect of game and delay for controllability, $F(10,266) = 2.74$, $p < .01$, $\eta_p^2 = .10$, for responsiveness, $F(10,240) = 4.99$, $p < .001$, $\eta_p^2 = .17$, as well as for immediate feedback, $F(9.39,234.72) = 2.27$, $p = .017$, $\eta_p^2 = .08$, $\varepsilon = .94$ (a Huynh-Feldt correction was applied). Simple main effects of the used games were found for a delay of 150 ms and at 300 ms for each dependent variable, as summarized in Table 4.4.

Furthermore, a strong simple main effect of delay was found for each game for all dependent variables (with an exception of Rocket Escape for immediate feedback). For controllability, the effect was weakest for Rocket Escape, Wilks's $\lambda = .81$, $F(2,129) = 15.03$, $p < .001$, $\eta_p^2 = .19$, and strongest for Shooting Range, Wilks's $\lambda = .54$, $F(2,129) = 55.77$, $p < .001$, $\eta_p^2 = .46$.

Overall, the results are in line with expectations towards the impact of delay. A strong impact of the delay was observed and both shooting games were more affected by delay than Dodge, T-Rex and Rocket Escape at 150 ms, whereas these differences diminished at 300 ms as ratings are already close

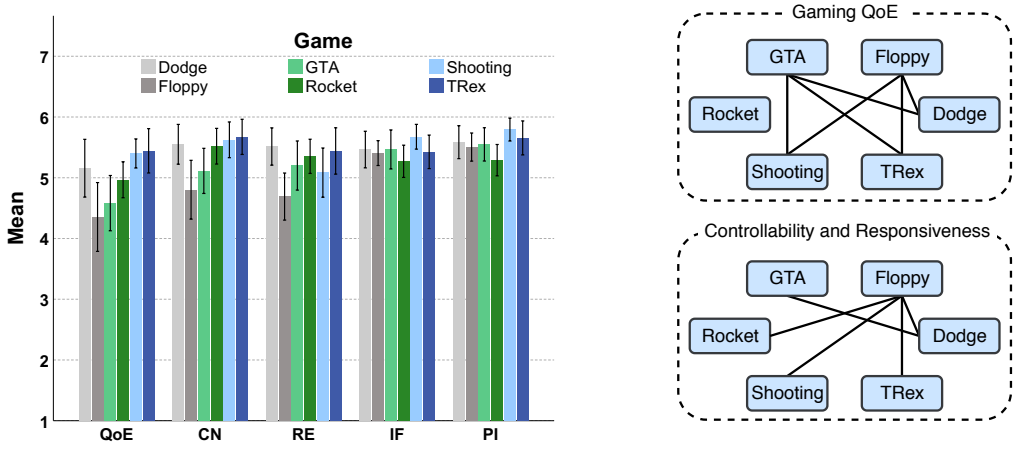
Table 4.4: ANOVA statistics of simple main effects of the used games.

Delay	Controllability			Responsiveness			Immediate feedback		
150 ms	F(5,130) 9.38	p <.001	η_p^2 .27	F(5,120) 12.18	p <.001	η_p^2 .34	F(5,125) 3.00	p .014	η_p^2 .11
300 ms	F(5,130) 6.09	p <.001	η_p^2 .19	F(5,120) 6.74	p <.001	η_p^2 .22	F(5,125) 3.91	p <.01	η_p^2 .14

to the end of the scale. When comparing the dependent variables with each other, it became visible that controllability, $r_p = .83$, $p < .001$, and responsiveness, $r_p = .81$, $p = .001$, are highly correlated with gaming QoE.

Study 4.2: Game

As in all other studies, only one game was tested per CG job, it was of interest whether a combination of games would lead to a strong variance in the ratings. A comparison of subjective ratings of each game for each of the dependent variables is shown in Fig. 4.6a whereas statistically significant differences based on pairwise comparisons using the Holm adjustment method [221] are visualized in Fig. 4.6b.



(a) Bar plots of means and 95 % confidence interval

(b) Significant differences

Figure 4.6: Influence of game on gaming QoE (QoE), controllability (CN), responsiveness (RE), immediate feedback (IF), and performance indication (PI) during CG Study 4.2.

Due to the complex block design, i.e., a worker only played three out of six games, a generalized linear mixed model analysis was carried out. Concretely, a random intercept model using the game as a fixed factor and the workers as a random effect was generated. A significant main effect of game was revealed for gaming QoE, $F(5,123.08) = 5.25$, $p < .001$, for controllability, $F(5,115.97) = 4.54$, $p < .001$, as well as for responsiveness, $F(5,120.10) = 3.22$, $p < .001$. As shown in Fig. 4.6b, the games GTA and Floppy Bird were (partially) rated significantly lower than the other games.

The results revealed interesting findings as even though no impairment was added to the games, some differences in gaming QoE ratings were observed. While this could have many reasons due to the multidimensionality of the construct, e.g., challenge provided by the game, once more the controllability and responsiveness ratings seem to explain the differences for the most part. Compared to the reference condition of the first study, the results are very reliable, as only Floppy Bird showed considerably lower ratings. A reason for this difference is unknown. Neither the self-reported gaming expertise,

M = 3.52, SD = 0.81 in Study 4.1 and M = 3.54, SD = 0.97 in Study 4.2, nor the average scores in the game, M = 7.79, SD = 5.95 in Study 4.1 and M = 9.95, SD = 9.83 in Study 4.2, can explain the findings. The rather simple nature of the browser-based games in comparison to high-end games using modern game engines could be a reason as interestingly, GTA and Floppy Bird were the most advanced games in terms of graphical appeal. Thus, the expectations of participants might be higher as for the other four games.

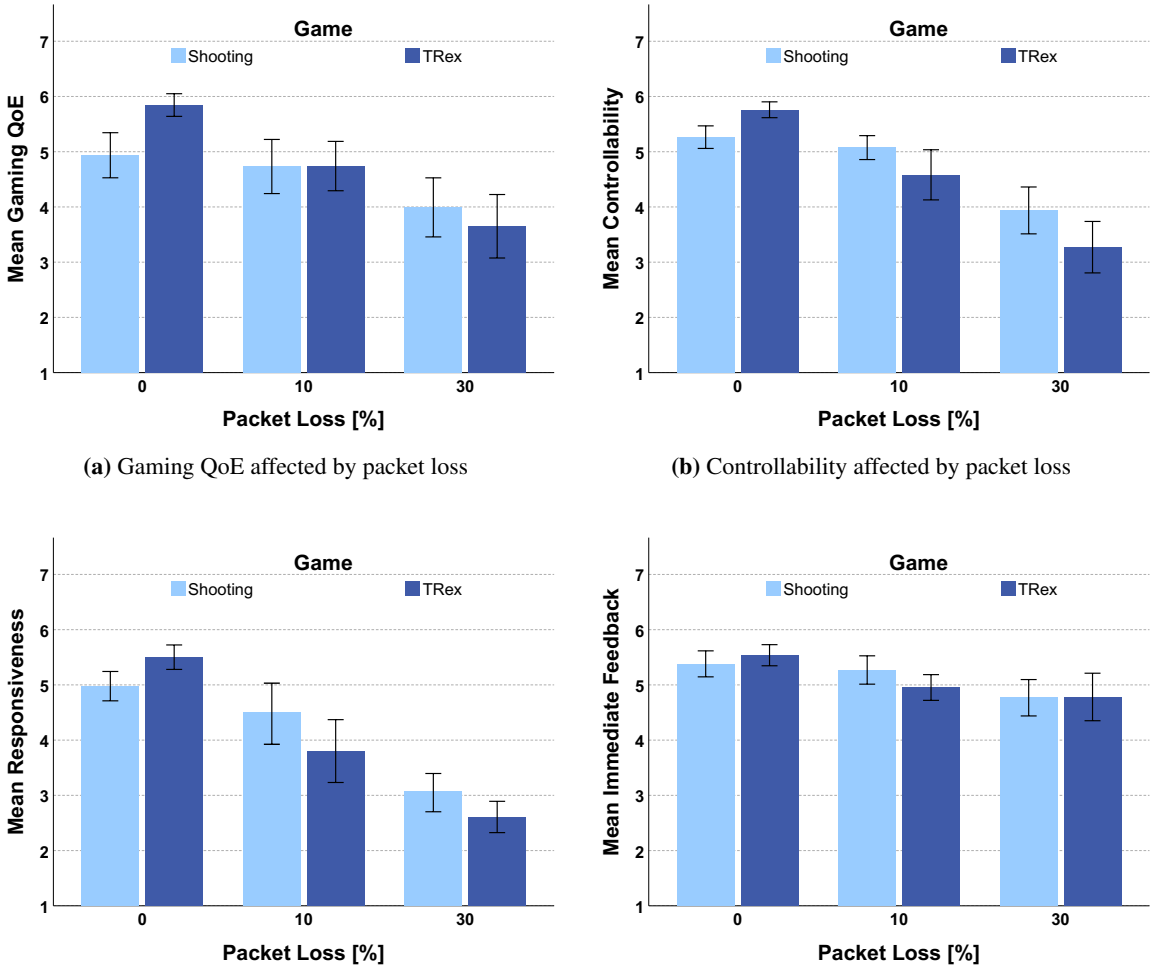
As at this point, the impact of a very dominant technical influencing factor, i.e., a network delay, as well as variation of games are covered in the presented studies, it was decided to concentrate on two out of the six games for the following studies. Therefore, the games T-Rex and Shooting Range were selected, as they produced reliable results in the first two studies and are very different in their game mechanics. It must be noted that both games got improved in terms of their game design for the remaining studies. While in the first two studies, only a timer of the remaining stimulus duration, a game score, and a performance indicator (passed obstacles or shooting accuracy) were provided to the participants, a new HUD was added to both games. The HUD showed the current game score, as well as a medal system, i.e., bronze, silver, and gold medal, reached at certain score thresholds, showing the progress in the game. Additionally, the aesthetic presentation of the game objects was slightly improved and a parameter showing the current difficulty of the game. i.e., the speed of objects was added. Finally, auditive feedback in case of performing an action, i.e., jumping or shooting, as well as when reaching a medal was added to the games.

Study 4.3: Input Packet Loss

The third study investigated the influence of packet loss linked to user inputs without any type of concealment method. Thus, it was expected to find a strong negative impact of the simulated degradations on the gaming QoE and aspects related to input quality. Fig. 4.7 shows the investigated test conditions.

The ANOVA revealed a significant interaction effect of game and packet loss for gaming QoE, $F(2,86) = 9.63$, $p < .001$, $\eta_p^2 = .18$. The effect resulted due to a simple main effect of game for the reference condition, $F(1,43) = 14.55$, $p < .001$, $\eta_p^2 = .25$, whereas no effect of game was shown for the packet loss conditions 10% and 30%. For the reference, T-Rex (M = 5.85, SE = 0.18) was rated significantly better than Shooting Range (M = 4.94, SE = 0.16). Someone can only speculate why this difference emerged for Study 4.3 but not for the previous studies. The implemented changes in the game design could be a reason. With respect to the influence of packet loss for each game, simple main effects are found for Shooting Range, Wilks's $\lambda = .68$, $F(2,42) = 9.91$, $p < .001$, $\eta_p^2 = .32$ and for T-Rex, Wilks's $\lambda = .35$, $F(2,42) = 38.64$, $p < .001$, $\eta_p^2 = .65$. No significant effect comparing the reference condition with the 10% packet loss condition was found for Shooting Range. Therefore, the impact of packet loss on gaming QoE was stronger for T-Rex. This could be caused by the fact that missing a jump in T-Rex would lead to an immediate punishment whereas in the game Shooting Range, players always had additional chances to shoot at the target.

In line with these findings, also for controllability, a significant interaction effect of game and packet loss was revealed, $F(2,80) = 10.64$, $p < .001$, $\eta_p^2 = .21$. For T-Rex, at all levels of packet loss, a simple main effect was yielded, Wilks's $\lambda = .18$, $F(2,39) = 89.86$, $p < .001$, $\eta_p^2 = .82$, whereas for Shooting Range, Wilks's $\lambda = .49$, $F(2,39) = 20.54$, $p < .001$, $\eta_p^2 = .51$, pairwise comparisons only showed a significant reduction of controllability ratings between the reference condition and 30% packet loss as



(a) Gaming QoE affected by packet loss

(b) Controllability affected by packet loss

(c) Responsiveness affected by packet loss

(d) Immediate Feedback affected by packet loss

Figure 4.7: Bar plots of means and 95 % confidence interval showing the impact of packet loss on gaming QoE and input quality during CG Study 4.3.

well as between the 10% and 30% packet loss condition. A main effect of the used games was revealed at the reference condition, $F(1,40) = 15.71$, $p < .001$, $\eta_p^2 = .28$, as well as for the 10% packet loss condition, $F(1,40) = 6.96$, $p = .012$, $\eta_p^2 = .15$, and 30% packet loss condition, $F(1,40) = 5.32$, $p = .026$, $\eta_p^2 = .12$. However, while T-Rex was rated slighly better for the reference condition, the opposite was the case for the packet loss conditions.

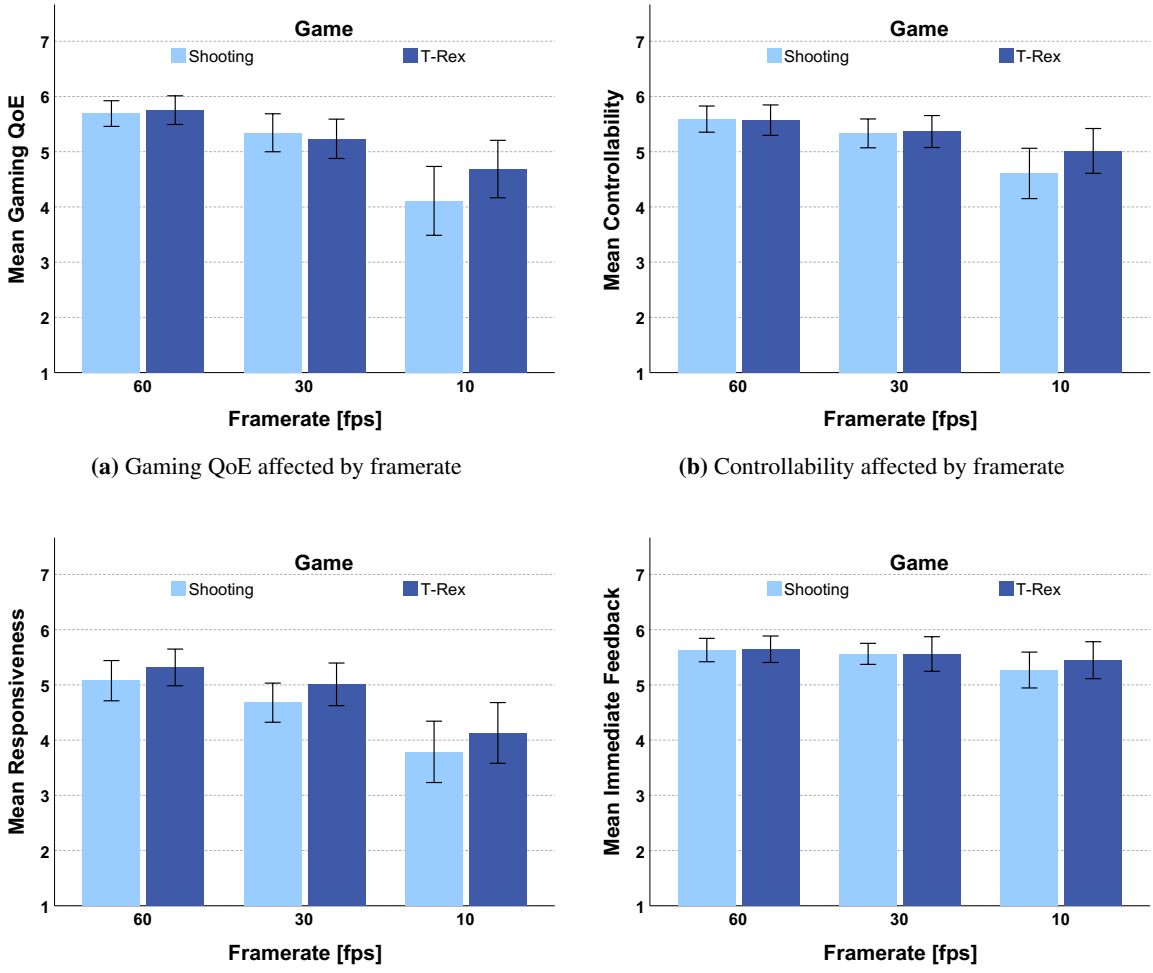
For responsiveness, also an interaction effect was shown, $F(2,70) = 8.00$, $p < .001$, $\eta_p^2 = 0.19$. However, the reason for this effect was based on the reference condition, as T-Rex was rated significantly better than Shooting Range, $F(1,35) = 12.18$, $p < .001$, $\eta_p^2 = 0.26$. This was not the case in the previous study nor for the other packet loss conditions. Furthermore, a main effect of packet loss was revealed for all conditions of T-Rex, Wilks's $\lambda = 0.13$, $F(2,34) = 118.85$, $p < .001$, $\eta_p^2 = 0.87$. For Shooting Range, also a main effect of packet loss was shown, Wilks's $\lambda = 0.28$, $F(2,34) = 44.00$, $p < .001$, $\eta_p^2 = 0.72$. However, pairwise comparisons between the reference condition and the 10% loss condition showed no significant difference, $p = .088$.

Lastly, for immediate feedback, a main effect of packet loss was exposed, $F(2,70) = 17.16$, $p < .001$, $\eta_p^2 = .33$. Pairwise comparisons indicated significant differences for all comparisons of packet loss conditions with the exceptions of the reference compared to 10% packet loss for Shooting Range as well as 10% compared to 30% packet loss for T-Rex.

In summary, the results of the study fulfill the expected impact of high packet loss. Furthermore, a much stronger impact on controllability and responsiveness compared to immediate feedback was revealed. The degradation of controllability and responsiveness ratings was comparable to Study 4.1.

Study 4.4: Framerate

In the fourth study, the influence of framerates on gaming experience was investigated for the games Shooting Range and T-Rex. Three levels of framerates (60, 30, 10 fps) were simulated by changing the game engine drawing rate which would correspond to a change of the encoding framerate of a cloud gaming system while assuming no other cause of frame losses. Fig. 4.8 shows the gaming QOE and input quality features of the test conditions.



(a) Gaming QoE affected by framerate

(b) Controllability affected by framerate

(c) Responsiveness affected by framerate

(d) Immediate Feedback affected by framerate

Figure 4.8: Bar plots of means and 95 % confidence interval showing the impact of framerate on gaming QoE and input quality during CG Study 4.4.

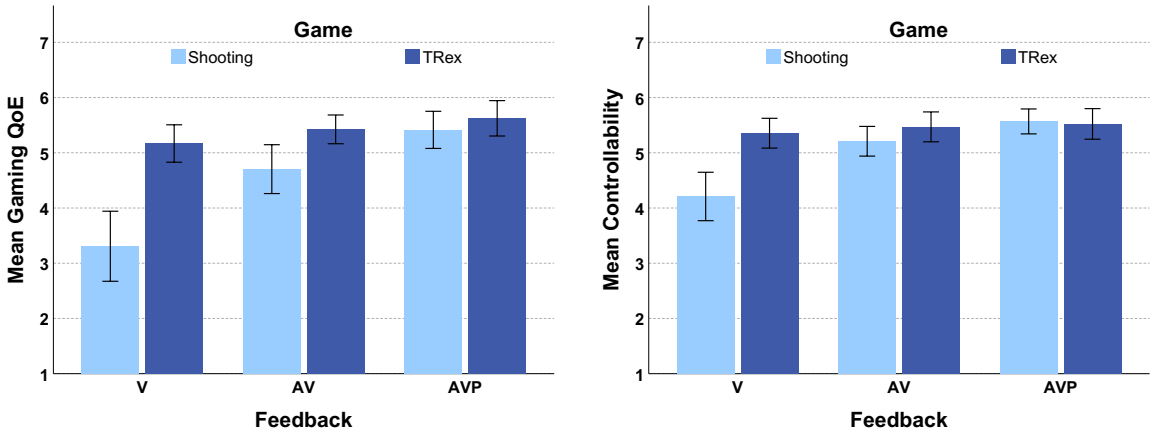
The ANOVA yielded a significant main effect of framerate for gaming QoE, $F(1.44,57.74) = 27.79$, $p < .001$, $\eta_p^2 = .41$, $\varepsilon = .72$ (Greenhouse-Geisser correction was applied). However, for Shooting Range no significant difference between 60 and 30 fps was found, whereas for T-Rex no difference was shown for 30 and 10 fps. Also, a significant main effect of framerate was shown for the game Shooting Range regarding the ratings for controllability, $F(1.75,68.14) = 14.18$, $p < .001$, $\eta_p^2 = .27$, $\varepsilon = .87$, as well as for responsiveness, $F(1.75,70.10) = 20.35$, $p = 0$, $\eta_p^2 = 0.34$, $\varepsilon = .88$ (Huynh-Feldt correction was

applied). However, significant differences did only emerge for the 10 fps conditions compared to both other conditions. For immediate feedback, no significant main effect was found.

While the ratings of gaming QoE are in line with findings from traditional lab studies showing that non-expert gamers do not rate 60 fps and 30 fps much differently, it appears slightly surprising that the impact of reduced framerates was rather low for the controllability. It seems that the caused visual discontinuity, i.e., no smooth movements of game objects, reduced the gaming QoE but had only a minor impact on the interaction with the games.

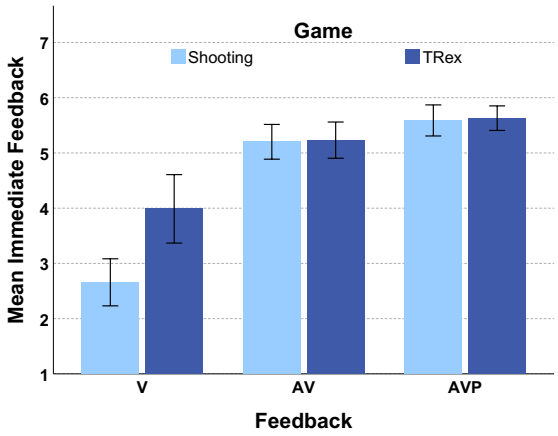
Study 4.5: Feedback Type

Despite the four other studies that were mostly focused on the network degradation, the fifth study investigates changes in the game design. Three types of feedback were developed for the games Shooting Range and T-Rex: minimum visual feedback showing only a timer and game elements (V), proper audio-visual feedback including also points and scores (AV), and audio-visual feedback with progress information such as medal systems (AVP). Fig. 4.9 the bar plots of the dependent variables with exception of responsiveness, as only a small influence of the game design on this aspect was expected.

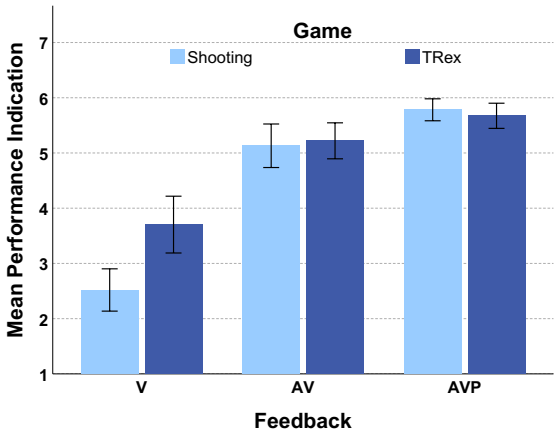


(a) Gaming QoE affected by feedback type

(b) Controllability affected by feedback type



(c) Immediate Feedback affected by feedback type



(d) Performance Indication affected by feedback type

Figure 4.9: Bar plots of means and 95 % confidence interval showing the impact of feedback type on gaming QoE, controllability, immediate feedback, and performance indication during CG Study 4.5.

For both games, adding more feedback elements to the game resulted in enhancement of the ratings for the dependent variables. The enhancement was stronger in the game Shooting Range as in the version with only visual feedback, users did not have a good insight whether they were successful on shooting the targets (missing bullet hole). As they had to hit a target three times, which was not the case in T-Rex, this had stronger impact on the gameplay. This difference between the games when comparing the audio feedback conditions with the other feedback types was also confirmed by the two-way Mixed ANOVA which revealed an interaction effect for gaming QoE, $F(1.65,82.37) = 17.51$, $p < .001$, $\eta_p^2 = 0.26$, $\varepsilon = .82$, controllability, $F(1.66,81.48) = 18.17$, $p < .001$, $\eta_p^2 = 0.27$, $\varepsilon = .83$, immediate feedback, $F(1.68,82.09) = 10.17$, $p < .001$, $\eta_p^2 = 0.17$, $\varepsilon = .84$, performance indication, $F(2,96) = 8.79$, $p < .001$, $\eta_p^2 = 0.15$, $\varepsilon = .80$, as well as for responsiveness, $F(1.94,96.92) = 7.25$, $p < .01$, $\eta_p^2 = 0.13$, $\varepsilon = .80$ (for $\varepsilon < 1$, a Huynh-Feldt correction was applied). While for T-Rex, no significant influence of the feedback type was found for gaming QoE, controllability, and responsiveness, a significant influence for each condition was found for immediate feedback, Wilks's $\lambda = .55$, $F(2,48) = 19.97$, $p < .001$, $\eta_p^2 = .45$, and performance indication, Wilks's $\lambda = .42$, $F(2,47) = 32.68$, $p < .001$, $\eta_p^2 = .58$. For Shooting Range, a significant simple main effect was found for each dependent variable. While for performance indication, Wilks's $\lambda = .21$, $F(2,47) = 89.45$, $p < .001$, $\eta_p^2 = .79$, the strongest effect was observed and each condition was statistically significant different from each other, the weakest effect was shown for responsiveness, Wilks's $\lambda = 0.58$, $F(2,49) = 17.48$, $p < .001$, $\eta_p^2 = 0.42$, where only the visual feedback condition was different to both others.

Summarising the above, it can be said that the implemented game design changes with respect to the feedback types showed an impact on the performance indication feature but had a low influence on the input quality (only the missing bullet holes were interpreted as issues with the responsiveness and controllability of Shooting Range). Furthermore, one can conclude that adding auditive feedback was more impactful than adding additional progress information.

Study 4.6: Feedback Type and Delay

In the last CG test, feedback type and delay are used as independent variables. The study design consists of eight blocks. In the first block, the game Shooting Range was played whereas in the second block, the game T-Rex was used. For two of the remaining four blocks per game, a delay of 150 ms was used and for the other two a delay of 300 ms was used. Each study started with a reference condition, i.e., no delay and only visual feedback, followed by the same feedback type but with a delay of either 150 ms or 300 ms as a second condition, and lastly with the same delay but audio-visual feedback (AV) or audio-visual feedback with progress information (AVP). The order of the second and third condition was randomized. It must be noted that compared to the fifth study, the visual feedback for Shooting Range was improved as the bullet animation was added for each condition. Thus, the aim of this study was to investigate whether adding sound changes impact of delay on the dependent variables and whether additionally adding progress feedback changes the impact of a delay.

Due to the low sample size (7 to 11 ratings per condition) caused by the block design or rather by the high number of independent variables, a Wilcoxon signed-rank test was performed for each condition combination. The test determined that there was a statistically significant increase in performance indication for T-Rex for the AV feedback condition ($M = 5.36$, $SD = 0.64$) compared to the visual feedback condition at 150 ms ($M = 3.88$, $SD = 1.14$), $z = -2.380$, $p = .017$, as well as for the AVP feedback condition ($M = 5.62$, $SD = 0.56$) compared to the visual feedback condition at 150 ms

(M = 4.18, SD = 0.93), z = -2.80, p < .01. For the latter, also a significant increase of immediate feedback for the AVP condition (M = 5.25, SD = 0.59) compared to the visual feedback condition (M = 3.93, SD = 1.07) was found, z = -2.55, p = .011. However, no significant difference was revealed for gaming QoE when comparing the AVP condition (M = 4.52, SD = 1.23) with the visual feedback condition (M = 4.28, SD = 0.7), z = -0.56, p = .57, nor for the other input quality features. Regarding the 300 ms conditions for T-Rex, once more a significant improvement of performance indication was found when comparing the AVP feedback condition (M = 5.45, SD = 0.82) with the visual feedback condition (M = 3.78, SD = 0.81), z = -2.37, p = .018. Also, for immediate feedback, a significant difference was shown when comparing the AVP feedback (M = 4.9, SD = 1.05) with the visual feedback (M = 3.29, SD = 0.31), z = -2.20, p = .028. While for Shooting Range, a significant improvement of performance indication at 150 ms was shown comparing the AVP feedback (M = 5.33, SD = 0.79) with the visual feedback (M = 4.93, SD = 0.36), z = -2.51, p = .012, only a tendency to an effect was determined for the 300 ms conditions comparing the AVP feedback (M = 5.08, SD = 0.57) and visual feedback (M = 4.38, SD = 1.34), z = -1.68, p = .092. No effect for the other dependent variables was found for Shooting Range. In summary, it can be concluded that the mediating effect of feedback type, i.e., adding auditive feedback and progress information, on the impact of delay on gaming QoE and related aspects is rather limited.

4.3 Test Environment Comparison

While the presented results of the six CG studies fulfilled the expectations with respect to the influence of the simulated impairments on gaming QoE and input quality features, it remains still an open question as to how closely they resemble results gathered for the same conditions in a lab environment. Therefore, additional ratings using the crowd gaming framework in three CS tests as well as a similar setup in two lab studies were collected.

4.3.1 Data Collection

In the first lab study, referred to as Dataset 4.3 (Lab1) in the following, Sabet et al. [210] investigated the adaptability of players to delay by generating an artificial pattern of changing delays in three games. The investigation is done both subjectively and objectively by tracking a player's in-game performance. The study design adhered to the guidelines provided in ITU-T Rec. P.809 and assessed the gaming QoE, PX using the iGEQ, as well as input quality features used in the CG studies presented in the previous section. For the study, three games developed for the CG framework were used. However, two of them were adapted for the purpose of the adaptability research. Thus, only the game Rocket Escape for the reference condition and for a delay of 300 ms can be compared to the CG data. The corresponding CG data for this game was presented in Study 4.1 in Section 4.2, and is referred to as Dataset 4.3 (CS1) in the following. In 2020, a second lab study was particularly designed for the comparison of both test environments, i.e., crowdsourcing and lab environment. The study design adhered once more to the guidelines provided in ITU-T Rec. P.809 and apart from gaming QoE and input quality features, also the video discontinuity, player performance, and acceptability were assessed using the games T-Rex and Shooting Range. As independent variables, input delay (0, 100, 200, and 300 ms) and framerates (60, 30, and 10 fps) were used. In addition, during the 200 ms delay condition, also a uniform jitter of 50 ms or 100 ms was added using a random number generator. To follow the

comparison with corresponding CS data in the following, the delay and jitter conditions are combined to the Dataset 4.3 (Lab2), whereas the framerate conditions are combined to the Dataset 4.3 (Lab3).

To complement the data of the second lab study, two additional CS tests were conducted adhering to the CG framework presented in this chapter. In one CS test, referred to as Dataset 4.3 (CS2), the framerate conditions and games included in Dataset 4.3 (Lab2) were used. In the other CS test, referred to as Dataset 4.3 (CS3), the remaining conditions were investigated. However, in the latter test, also some additional conditions were added to investigate several latency compensation techniques using game adaptation which mitigates the influence of delay on QoE [211]. These conditions are excluded from a comparison of the test environments in the following. In Table 4.5 a summary of all conditions used in each dataset is presented. Lastly, Table 4.6 provides experimental details and demographic information about participants of the conducted studies as well as some statistics about the data cleansing. For the latter, the same method as described in Section 4.2.3 was considered. Regarding the demographic information it can observed that participants in the lab studies are younger but that the expertise of participants in each dataset was similar.

Table 4.5: Overview of test conditions of subjective test datasets for test environment comparison.

Dataset 4.3 (comparison)	Condition (CID)	Game	Delay [ms]	Jitter [ms]	Framerate [fps]	N (lab)	N (CS)
CS1 vs. Lab1	1	Rocket Escape (RE)	0	0	60	27	20
CS1 vs. Lab1	2	Rocket Escape (RE)	300	0	60	27	20
CS2 vs. Lab2	3	Shooting Range (SR)	0	0	60	24	24
CS2 vs. Lab2	4	Shooting Range (SR)	100	0	60	26	29
CS2 vs. Lab2	5	Shooting Range (SR)	200	0	60	25	29
CS2 vs. Lab2	6	Shooting Range (SR)	200	50	60	25	30
CS2 vs. Lab2	7	Shooting Range (SR)	200	100	60	23	26
CS2 vs. Lab2	8	Shooting Range (SR)	300	0	60	24	28
CS2 vs. Lab2	9	T-Rex (TR)	0	0	60	26	31
CS2 vs. Lab2	10	T-Rex (TR)	100	0	60	26	29
CS2 vs. Lab2	11	T-Rex (TR)	200	0	60	26	26
CS2 vs. Lab2	12	T-Rex (TR)	200	50	60	25	28
CS2 vs. Lab2	13	T-Rex (TR)	200	100	60	25	28
CS2 vs. Lab2	14	T-Rex (TR)	300	0	60	26	32
CS3 vs Lab3	15	Shooting Range (SR)	0	0	60	24	28
CS3 vs Lab3	16	Shooting Range (SR)	0	0	30	26	31
CS3 vs Lab3	17	Shooting Range (SR)	0	0	10	23	33
CS3 vs Lab3	18	T-Rex (TR)	0	0	60	26	19
CS3 vs Lab3	19	T-Rex (TR)	0	0	30	26	27
CS3 vs Lab3	20	T-Rex (TR)	0	0	10	23	21

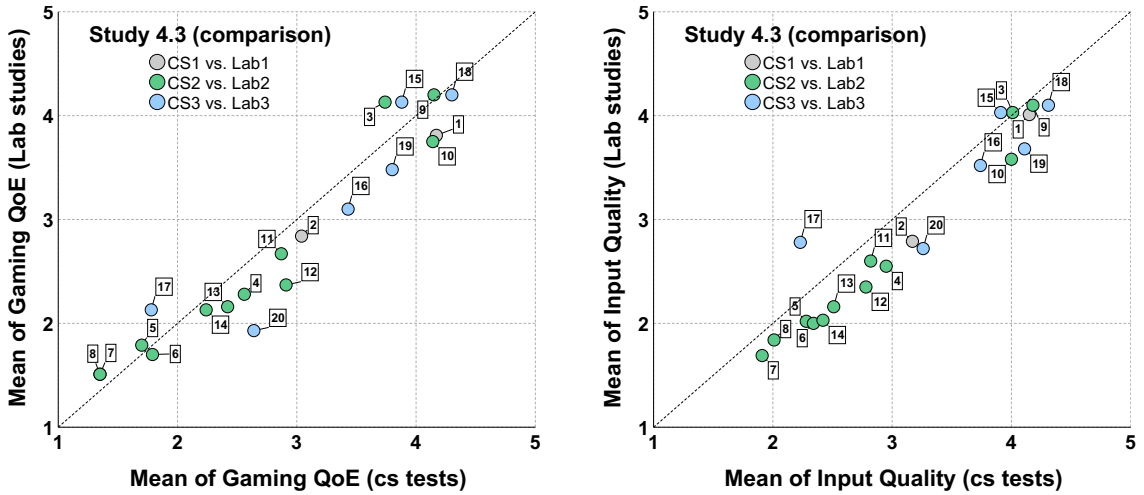
4.3.2 Analysis

To allow a comparison of subjective ratings between the CG approach and lab studies, the Dataset 4.3 (Lab1), Dataset 4.3 (Lab2), and Dataset 4.3 (Lab3) are combined to represent 20 conditions investigated in the lab environment. Also, the Dataset 4.3 (CS1), Dataset 4.3 (CS2), and Dataset 4.3 (CS3) are combined to represent the corresponding data collected in a CS environment. Please note that the ratings are converted to the typically used 5-point range for video quality according to [188] for the purpose of this comparison.

Table 4.6: Overview of subjective test datasets for test environment comparison.

Dataset 4.3 (comparison)	CS1	CS2	CS3	Lab1	Lab2 & Lab3
Date (year)	2018	2020	2020	2019	2020
Environment	CS	CS	CS	Lab	Lab
N conditions	2	12	6	2	18
Samples assessed	70	468	231	54	416
Samples with wrong trapping questions	24	84	51	-	-
Samples after full data cleansing	40	340	159	54	398
Demographic Information of Test Participants					
Age (M / SD)	33.43 / 7.73	37.98 / 10.52	35.25 / 9.07	26.25 / 3.79	25.15 / 4.94
Gender (f/m/other)	20 / 14 / 1	28 / 28 / 1	16 / 37 / 0	10 / 17 / 0	7 / 19 / 0
Expertise (M / SD)	3.54 / 1.04	3.46 / 0.79	3.75 / 0.98	3.55 / 0.75	2.84 / 1.18

The descriptive statistics for the gaming QoE and input quality of the 20 conditions are provided in Table A.3 in Appendix A. For a visual representation of the data, the scatter plots of the mean values for each condition are shown in Fig. 4.10. Each data point is labeled with the condition ID (CID) that was assigned in Table 4.5.



(a) Comparison of gaming QoE ratings

(b) Comparison of input quality ratings

Figure 4.10: Comparison between assessed MOS from lab studies with assessed MOS from CS tests of (a) gaming QoE, and (b) input quality.

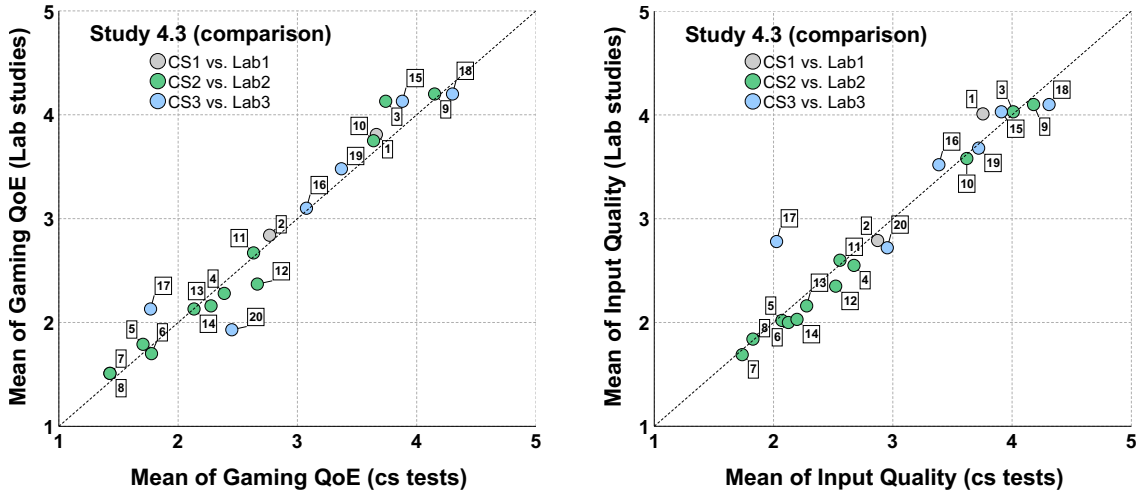
The scatter plots show that the mean values are similar between lab and CS results for the majority of conditions. For a statistical comparison, the MOS values per condition obtained from the CS tests are compared with those from the corresponding laboratory-based experiments in terms of Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Correlation Coefficient (SRCC), and Root Mean Square Error (RMSE). The analysis revealed a low RMSE of 0.31 for gaming QoE and 0.33 for input quality, as well as a very high correlation between the MOS values obtained in CS tests and lab studies for gaming QoE, PLCC = .955 and SRCC = .953, and for input quality, PLCC = .959 and SRCC = .922. These results are a first indicator illustrating a high comparability of subjective ratings resulting from CS tests and lab studies. However, in the scatter plots, it can be observed that there is a

shift, i.e., a bias and a different gradient, between CS and lab scores throughout all datasets with the exception of the reference conditions. Crowdworkers tend to rate the gaming QoE and input quality slightly higher as participants in the corresponding lab studies. Thus, first order mapping functions adjusting the crowdsourcing MOS values to the lab MOS values based on a linear regression were applied according to the following equations:

$$GamingQoE\_mapped = 0.359 + 0.792 \cdot GamingQoE\_crowd \tag{4.1}$$

$$InputQuality\_mapped = 0.020 + 0.90 \cdot InputQuality\_crowd \tag{4.2}$$

For the latter, a very similar mapping was applied in [222] to adjust ratings of speech quality assessed using a crowdsourcing approach to corresponding lab data. The resulting ratings of both test methods when applying the mapping are depicted as scatter plots in Fig. 4.11. The mapping decreases the RMSE between lab studies and CS test ratings to 0.20 for gaming QoE and 0.21 for input quality. Furthermore, the correlation between the MOS values increased for gaming QoE, PLCC = .978 and SRCC = .976, as well as for input quality, PLCC = .968 and SRCC = .939.



(a) Comparison of gaming QoE ratings

(b) Comparison of input quality ratings

Figure 4.11: Comparison between assessed MOS from lab studies with mapped MOS from CS tests of (a) gaming QoE, and (b) input quality.

To complement the previously presented analysis, in the following the collected data will be analyzed regarding its distribution of the ratings, i.e., by performing hypothesis testing. However, the three CS studies are using different groups of test participants, types of degradations, games, as well as numbers of ratings per condition. This incomplete block design makes it impossible to use traditional parametric tests such an an ANOVA for the analysis of significant differences between the means. Thus, a generalized linear mixed model analysis was carried out. A random intercept model, i.e., a generalized linear mixed effect model (GLMM), using the *test environment* (CS and Lab) and *test condition* (CID) as a fixed factor, and the *participants* as a random effect was generated. The GLMM test statistics for the main effects of condition and test environment as well as their interaction effect are summarized in Table 4.7.

For gaming QoE, irrespective of the applied data mapping, no main effect of test environment was found. The overall mean value for the gaming QoE (not mapped) in the lab, M = 2.79 (SD = 0.62), is only slightly lower compared to CS test results, M = 2.91 (SD = 0.68).

Table 4.7: Test statistics of GLMM analysis for test environment comparison.

	Gaming QoE				Gaming QoE (mapped)			
Effect	F	df1	df2	p	F	df1	df2	p
Condition	112.42	19	650.61	<.001	128.92	19	684.06	<.001
Environment	2.49	1	90.98	.12	0.20	1	89.50	.66
Condition x Environment	2.74	19	650.61	<.001	1.69	19	684.06	.03
	Input Quality				Input Quality (mapped)			
Effect	F	df1	df2	p	F	df1	df2	p
Condition	101.71	19	665.58	<.001	117.37	19	676.75	<.001
Environment	9.05	1	107.61	<.001	0.01	1	105.82	.91
Condition x Environment	2.44	19	665.58	<.001	2.01	19	676.75	.01

For input quality, the overall mean value of M = 2.93 (SD = 0.57) in the lab is significantly lower compared to the mean value of M = 3.15 (SD = 0.64) resulting from the CG tests. However, when applying the data mapping, no significant effect of the test environment for input quality, $F(1,106) = 0.01$, $p = .91$, can be observed.

However, an interaction effect of test environment and test condition was found for gaming QoE as well as for input quality. When considering the mapped data, for gaming QoE, the reference condition (CID3) of the game Shooting Range in the Dataset 4.3 (CS2) was rated significantly higher in the lab study, $p = .04$. It must be noted though that this effect is very small. Additionally, both conditions with a framerate of 10 fps (CID17, CID20) were rated significantly different when comparing both test environments. For the mapped input quality results, also for the 10 fps condition (CID17) of the game Shooting Range, the Lab MOS was significantly lower than the CS MOS.

As a final investigation with respect to the comparability of the ratings obtained from both test environments, it is interesting to statistically analyze whether one would draw the same conclusions when comparing the individual test conditions with each other, irrespective of the test environment used. Thus, for each test environment, once more a GLMM analysis using the *test condition* (CID) as a fixed factor, and the *participants* as a random effect was performed. A summary of all pair-comparisons for the gaming QoE and input quality is provided in Table A.4 and A.5 in Appendix A.

For gaming QoE, overall the majority of pair-comparisons between the conditions lead to the same conclusion for both test environments. For 5 of the 37 possible comparisons, which are all high delay conditions including jitter, the participants in the lab studies, unlike the crowdworkers, did not rate the gaming QoE significantly different. However, it must be noted that for Shooting Range, the differences between the mean values of these conditions are very small, e.g., a difference of 0.35 for the CS MOS of the 200 ms condition (CID5) compared to the 300 ms condition (CID8). As the ratings for the latter conditions are already close to saturation, their standard deviation is very small. Consequently, despite the small deviations in the means this still resulted in statistically significant differences between the conditions. For T-Rex, contrary to the lab results, the gaming QoE for a constant delay of 200ms in addition to a jitter of 50ms (CID12) was rated noticeably higher in the CS tests. This led to statistically significant differences when comparing CID12 to CID13 (a constant delay of 200ms in addition to a jitter of 100ms) and CID14 (a constant delay of 300ms).

For the input quality, only for the comparison of CID6 (a constant delay of 200ms and a jitter of 50ms) and CID8 (a constant delay of 300ms) statistically significant differences were revealed in the CS data but not in the lab results.

4.3.3 Discussion

In sum it can be concluded that in most cases the CS tests resulted in comparable results to those gathered in the lab studies. This was not only shown by a very low RMSE and high correlation, but also due to a similar spread of the data in terms of standard deviations as well as regarding the usage of the rating scale range.

Interestingly, the analysis revealed with one exception that the reference conditions are not rated significantly different when comparing both test environments. Consequently, one may argue that crowdworkers only rated conditions with impairments more positively than participants in lab studies. An open question remains about what caused the systematic shift in the ratings for the impaired conditions, e.g., a delay condition. A possible reason could be various user factors that cannot be analyzed with much confidence due to the rather low amount of available data points. Furthermore, it could be that crowdworkers actually expected a reduced gaming experience as they might have experienced network issues in online games in their test environment before, whereas participants in the lab might not expect such issues as they are in a well-controlled environment. In this respect, a CS test may reflect the reality of a gaming session better than a lab study.

Regarding the different types of investigated degradations, it was shown that for two games, the 10 fps condition was not rated consistently when comparing both environments. This might be due to different display devices being used. Thus, further research is required to investigate whether the comparability between lab and CS results is limited for particular types of (visual) degradations. This also applies to the resolution of parameter changes with respect to deriving the same conclusions for both test environments when comparing different conditions. While it appears that the majority of the delay conditions were rated consistently similar comparing both test environments, for comparisons of delay conditions with only minor differences, some inconsistencies between the CS MOS and lab MOS were shown.

However, the overall close similarity between the CS tests and lab studies results, the use of the rating scale range and distribution of ratings, and also the monotony of the ratings for expected trends, e.g., a reduction of gaming QoE and input quality ratings for higher delays, showed that the crowdgaming method is well suited for the purpose of this thesis.

4.4 Summary

In this chapter, a new method for the assessment of gaming QoE using a crowdsourcing approach was presented. In comparison to passive tests, including an engagement check in an interactive test was very useful. It helped to filter data from workers who did not play the game as expected and ensured that workers learned to control the games during a training scenario. A training section was crucial to gather high-quality data. The same applies to the quality control items added to the questionnaire. As suggested in ITU-T P.808, the number of additional items added for reliability checks should not be larger than 10% of the number of items in the questionnaire. In the case of a short questionnaire, especially the game content related trapping questions should be considered. While the dropout rate of

about 30% in our tests appears to be high, this value is also in line with CS tests for the assessment of speech quality we conducted in the past.

The presented work shows that with a proper stimulus design and controlling the environment and participants' behavior, results obtained by crowdgaming studies can resemble those gathered from lab studies for the tested conditions, i.e., artificially added delay and packet loss as well as frame losses and game design changes. Someone can reason that consequently, the developed framework should also work for various other conditions. Additionally, it can be confirmed that expected influences of the simulated degradations on the measured quality features and overall gaming QoE are observed in all CG studies. Due to their high correlation with the gaming QoE, the controllability and responsiveness seem to be good predictors of the former.

Interestingly, for a delay of 150 ms, the impact of delay on Shooting Range was much stronger than on T-Rex for the controllability but not that strong for responsiveness. Additionally, the impact of packet loss was generally stronger on responsiveness compared to the controllability for both games. Thus, both concepts appear to be of value for understanding the judgement process of players.

Furthermore, within the six studies, repeated conditions such as the reference condition showed very similar results which confirms high consistency. It is also encouraging that the scales, especially those of the input quality features, were used on a good range of low and high ratings. Thereby, it can be concluded that the crowdgaming method is well suited for the development or validation of questionnaires and that the work of the ITU-T Rec. P.808 and P.809 are of great use for crowdgaming tests.

Due to the satisfying similarity of data assessed using the CG method compared to traditional lab studies, the framework described in this chapter was proposed as a new recommendation based on the work item P.CrowdG to the ITU-T SG12 [33], [34].

Chapter 5

Development of the Gaming Input Quality Scale (GIPS)

In this chapter, the development of the Gaming Input Quality Scale (GIPS) is described. The steps of the development and validation process of the GIPS questionnaire are visualized in Fig. 5.1. Each step will be explained in more detail in the following sections.

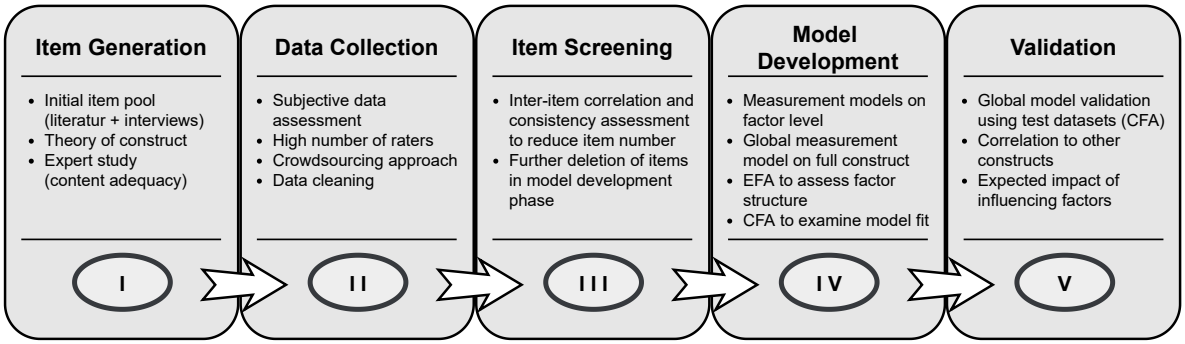


Figure 5.1: Procedure of development process of the Gaming Input Quality Scale (GIPS).

The GIPS aims to measure the input quality of a player interacting with a gaming system. In particular, the questionnaire should address the impact of common network degradations during cloud gaming such as delay and packet losses rather than the emotional state of a player or design flaws of a game. Participants in conducted studies, which are presented in Section 3.2 [15] and Section 6.1 [14] stated during post-test interviews and questionnaires that the playability of a game scenario, i.e., referred to as interaction quality in this thesis, is of high importance for their overall Quality of Experience judgement. However, as indicated in Section 3.3, no validated questionnaire is available to measure aspects related to the input quality of cloud gaming services. Thus, there was a need for the development of a psychometrically validated, and reliable instrument for assessing this important quality feature which also completes the last missing tool for assessing all major aspects covered by the cloud gaming taxonomy. Measuring a psychological construct such as the input quality is a challenging task as it represents an unobservable construct, a so-called *latent variable*, which cannot be measured directly but rather in an indirect way. As the term input quality might not be intuitive for test participants, it must be determined which items, i.e., questions of a questionnaire, adequately represent it and which can measure it reliably. As these items can be directly measured, they are called observable, manifest, or *indicator variables*. Finally, the input quality might be a composition of several

different components, which are referred to as factors in the following, rather than a single, solitary concept [223]. Please note that these factors should not be confused with quality influencing factors described in Section 2.2.2. Instead, considering the taxonomy terminology, they would be considered as quality features.

5.1 Item Generation

The creation of items to assess a construct under examination, in this case the input quality, can be done either inductively, deriving scales from generated items, or deductively, generating items based on a theoretical definition [224]. For the development of GIPS, the deductive approach was followed. Based on the post-test interviews and questionnaires used in the studies presented in Section 3.2 and Section 6.1, as well as by taking a variety of existing questionnaires presented in Section 3.1.2 into account, the factors linked to input quality described in the following were derived and considered for the GIPS.

Responsiveness

Responsiveness describes the temporal aspects of the feedback a player receives after performing an action, e.g., a mouse click or a keystroke. The response of the game (system) should be available immediately after the player performs an action (input event).

Controllability

The perceived controllability is the degree to which a player is able to control a game using the given input device and available interaction possibilities. It describes whether the performed input action resulted in the desired outcome. The controllability does not relate to the learnability of the controls nor to autonomy (freedom or power over something).

Performance Indication

Performance Indications offer the player all necessary insights about the progress of the game. The player should be aware of how well she/he is doing and what consequences each action has. For this purpose, the design and the content of the interface are important to ensure ideal feedback.

It must be mentioned that the last factor, performance indication, was assumed to be only a mediating factor, whereas the other two as direct factors of the input quality. Thus, performance indication, since it fits conceptually best in terms of the taxonomy terminology, will be considered as part of playing quality in the following of the thesis.

To begin with, a preliminary item pool consisting of 41 items for the concepts *responsiveness*, *controllability*, and *performance indication* was constructed. While one aim during the scale construction was to achieve a measurement tool consisting of a low number of items, a certain redundancy in the initial item pool was desirable as it serves to uncover sub-dimensions or closely related but distinct constructs [223]. Keeping a measure short is an effective means of minimizing response biases caused by boredom or fatigue [225].

The item pool was to a large extent derived from available questionnaires, which are partially also created for domains other than gaming, and by adding some self-created items. For its generation, the

guidelines presented in [224], [226] such as assessing only a single issue in one question, using simple and short items, using a language that is familiar to the target group, and to avoid negatively-worded items were largely followed. To adhere to these guidelines, some of the items taken from published questionnaires were modified as they were not designed for the context of gaming. The initial item pool, including the sources of the individual items, can be found in Table C.1 in Appendix C.

Assuring content adequacy before collecting user ratings for the construction of a questionnaire helps to strengthen the construct validity of the measurement tool as it enables removing items that may be conceptually inconsistent [224]. Thus, a pretesting was performed to investigate the items for content adequacy by applying a card sorting method. Six experts in questionnaire development or gaming were invited. The experts were between 24 and 50 years old, five experts were male, and three hold a doctoral degree while the other three were in their last year of a master's degree or completed it. Asked about whether they are an expert in gaming, either at playing or in research, three agreed to this statement. Five of the experts already developed a questionnaire by themselves, and four had good knowledge of network impairments during a gaming experience. After showing some examples of bad network conditions in games developed for the crowdsourcing studies described in Chapter 4, the aforementioned definitions of responsiveness, controllability, and performance indication were given to the expert. The experts then assigned each item to one of the concepts (or to the category "others"). They were then asked to mark items with wording issues (inconsistent terminology, understanding, missing words, or typos), to judge the relevance to the construct the item was assigned to (ordering), and to group items with very similar in their meaning. Items that were assigned by more than two experts to an unforeseen category were removed. Finally, also items marked with low relevance or very strong similarity were removed, and the wording was improved as suggested. While this technique does not guarantee validity of the scale, it provides evidence that the items represent a reasonable measure of the construct under examination [224]. The procedure resulted in 26 items used for a large-scale subjective test described in the following subsection.

5.2 Data Collection

For the development of a questionnaire by means of a factor analysis, a high number of ratings is required. In case of an insufficient sample size, the detected patterns of covariation might be unstable as correlations among questionnaire items can result solely by chance. In their content analysis on new scale development articles, Worthington and Whittaker conclude that a common rule of thumb of using a sample size of at least 300 is generally sufficient and that sample sizes of 150 to 200 are likely to be adequate if the dataset consists of a 10:1 participant-to-item ratio per factor with factor loadings at approximately |.4| [227]. A second important aspect of the data collection is to provide enough variance in user ratings for all concepts of the questionnaire. Apart from variability caused by the raters themselves, also a variety of different contents as well as typical technical influencing factors should be addressed. Lastly, the test participants should represent the target group of services for which the questionnaire will be used. In the case of GIPS, they should possess an interest in gaming activities and should be capable of controlling games properly. Specific gender or age requirements are not necessary, as the gaming market is very diverse (cf. Chapter 1).

For all these requirements, the crowdgaming framework described in the previous chapter is an ideal candidate as it allowed the collection of a high number of ratings from a diverse user group,

the simulation of typical degradations, i.e., delay, packet loss, and frame losses, as well as enough variability in the game content due to the available six JavaScript-based games. The games are different in their interaction concepts (based on game bricks classification [228]) and used input devices (mouse or keyboard). With respect to the performance indication concept, the method also allowed changes in the game design, i.e., generating different types of feedback.

Despite the games being of rather abstract nature, arguably they are well suited to investigate the input quality due to the fact that they cover the most common game mechanics. However, the games might not be suitable for research about immersion. Gerling et al. investigate how the fidelity of graphics affects PX [229]. The authors used two games with different mechanics, and compared the PX between a low-fidelity graphics (abstract) and high-fidelity graphics (stylized) version of each game. They showed that high-fidelity graphics result in an overall increase in the sense of immersion but that a game with suitable mechanics does not require high-fidelity graphics to result in a good PX, which also includes the feeling of control over the game.

Therefore, the dataset created as a result of the six crowdgaming studies presented in the previous chapter will be used for the development of the GIPS. In addition to the data cleansing method described in Section 4.2.3, also all ratings provided by participants with one or more wrongly answered trapping questions were removed.

For the input quality, the ratings from the studies 4.1 to 4.4 presented in Section 4.2.4, which used the independent variables delay, game content, framerate, and packet loss, are used. The 866 responses were randomly divided into two groups. The first group, referred to as *training dataset* in the following, contained approximately 40 % of the data, 235 samples from 171 different participants, and was used to derive the factor structure of the questionnaire. The remaining data, 379 samples from 214 different participants, referred to as *test dataset* in the following was used to validate the questionnaire.

For the performance indication factor, the ratings of the studies 4.5 and 4.6 are used, as in the other crowdgaming studies this factor was not targeted. Again, the data was split into a traning dataset consisting of 152 ratings (from 94 different participants) and a test dataset including 188 ratings (from 107 different participants).

5.3 Item Screening

While after the steps presented in Section 5.1 the fundamental concept of the input quality was established, there is still an unfeasibly high number of items for its assessment required. Thus, an item screening process is mandatory. For the development of the GIPS, the method used in [205], [230], which is based on the multi method approach published by Homburg and Giering in [231], was applied. Additionally, the guidelines presented in [223], [224], [232], [233] are followed. Before reporting on these steps, some principles of a factor analysis should be covered first. The basic assumption of a factor analysis is that for a collection of observed, correlated variables, i.e., items in the context of a questionnaire, there is a set of underlying variables called factors that can explain the interrelationships among those variables [233]. In psychology research, two techniques are commonly used to reduce the set of observed variables to a smaller, more parsimonious set of variables with as little a loss of information as possible: a) Principal component analysis (PCA), and b) Exploratory factor analysis (EFA). Whereas a PCA involves extracting linear composites of observed variables, an EFA is based on a formal model predicting observed variables from theoretical latent factors.

In the context of scale construction, the EFA is more frequently recommended to assess the underlying factor structure and refine the item pool [232]. This is due to the fact that a PCA does not take the unique variance, i.e., the variance specific to a particular item or error variance which comes from measurement errors [234], into account. However, as is it usually reasonable to assume that a set of items was not measured perfectly, the EFA is preferred. In the present work, we assume that there is a construct called input quality (or rather its factors) that explain a correlation among the corresponding items. An EFA aims to find the factors with the strongest correlation of items within a factor while reducing the correlation between the factors as much as possible.

Another fundamental method for item screening and questionnaire development is the Confirmatory factor analysis (CFA) which is typically used as a successor to the EFA. The goal of the CFA is to test an existing theory. It hypothesizes an a priori model of the underlying structure of the target construct and examines if this model fits the data adequately and assesses the relationships among items and scales [224]. Statistically, this is reached when the difference between the estimated and observed covariance matrices are very small. Based on the previously described methods and guidelines, the following procedure is applied:

1. Reduction of the number of items by evaluating the inter-item correlations and consistency analysis on the factor level (or on a construct consisting of multiple factors)
2. Further improvement of the item section by performing an EFA on the factor level while assuming a low-factor solution
3. Examine whether the priori model of the underlying structure of the target construct fits the data adequately using a CFA
4. Investigating the internal consistency for the final structure of each factor

Inter-item Correlation Analysis

Prior to conducting the factor analysis, the inter-item correlations are examined for each latent factor. The correlation tables are summarized in Appendix C. As described in [224], items should be considered for deletion if they correlate with less than .4 with all other variables as this indicates that the corresponding item is not drawn from the appropriate domain, thus leading to an unreliable measure. For all three factors, the majority of the items had absolute inter-item correlations above 0.70 which shows that the initial item pool was well selected and suitable for further analysis. However, for controllability, a low correlation of item CN6 ($r = .49$) with other items and moderate correlations for the items CN4, CN5, CN7, and CN9 ($r = .80$ to $.82$) are observable. For responsiveness, the items RE4, RE6, and RE8 show low correlations ($r \leq .56$) with the majority of other items but are correlated with each other. This indicates a possible two-factor solution for the selected items. For the performance indication items, the item PI6 had a low correlation ($r = .51$) with other items while the items PI8 ($r = .63$), and PI9 ($r = .63$) had a moderate correlation.

Initial Consistency Assessment

As suggested by Homburg and Giering in [231], already before the factor analysis it is beneficial to assess the internal consistency of each construct using the well-known reliability coefficient called Cronbach's α. A Cronbach's α value greater than 0.7 indicates a "strong item covariance or homogeneity and suggests that the sampling domain has adequately been captured (Churchill, 1979)"

[224]. Detailed information about the calculated Cronbach's α values are summarized in Table C.5 in the Appendix C. The analysis showed an initial Cronbach's α of 0.969 for the controllability items. While this is already very satisfying, the coefficient was further improved to 0.975 by removing the item CN6. Even though for the items CN5 and CN9 a slightly lower correlation with other items was found, removing these items would not improve the internal consistency, and thus the items remained for the upcoming EFA. For responsiveness, even though the initial Cronbach's α value of 0.964 for all responsiveness items is very good, a potential two-factor solution is also confirmed by the consistency analysis. When separating the weaker correlated items found by the inter-item correlation analysis, a Cronbach's α of 0.968 resulted for the items RE1, RE2, RE3, RE5, and RE7. It must be noted that removing RE2 would slightly increase the Cronbach's α but as the consistency is already very satisfying, the item remained for further analysis. Consequently, RE4 and RE6 remained as a possibly separated factor (Cronbach's α of 0.891) while the item RE8 was removed entirely. Lastly, for performance indication, the initial Cronbach's α of 0.95 was already very satisfying. However, removing the items PI6 and PI9 would further increase the Cronbach's α coefficients to 0.96. Apart from a suitable item selection, it must be noted that this may partly be caused by the large number of items as the Cronbach's α value also depends on the number of items of a construct.

5.4 Model Development

Measurement models for each input quality factor

When performing a factor analysis, we distinguish between two types of rotation methods. Rotations can be orthogonal, like varimax rotation, or oblique, like the promax rotation. With oblique factor rotations, the new factors are correlated while as for orthogonal rotation, the factors are not correlated [235]. When using an EFA on the factor level, the goal is to ensure that the items for each construct actually form only one factor (or a number as low as possible). In this step by step process of item reduction, one searches for issues resulting in case a solution requiring more than one and orthogonal factor. Therefore, in the following, a Maximum-Likelihood (ML) factor analysis using a varimax rotation is calculated for each factor (cf. [230]) using the training dataset. Here, ideally, a one-factor solution should be confirmed. The ML method allows for testing the fit of the hypothesized factor structure via the χ^2-goodness-of-fit test. For calculating the EFAs, IBM's SPSS 25 was used whereas measurement models are developed using IBM's AMOS 27. A measurement model can be defined as:

Measurement model

A measurement model, which is essentially a CFA, depicts the pattern of observed variables for the latent constructs in the hypothesized model and examines the reliability of the observed variables as well as the extent of interrelationships and covariation among the latent constructs [236].

Multiple criteria can be used to determine the factorability of the data. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy should reach a minimum of 0.5 whereas values between 0.5 and 0.7 are mediocre, values between 0.7 and 0.8 are good, values between 0.8 and 0.9 are great and values above 0.9 are superb [233]. Next, Bartlett's test of sphericity, which hypothesis that the correlation matrix is an identity matrix, should be reported. A significant result ($p < .05$) indicates that the matrix is not an identity matrix, i.e., the variables do relate to one another enough to run a meaningful EFA.

The goodness-of-fit test should not be significant on an alpha level of 5 percent and exceed a χ^2 of 0.6. This would support to confirm the null hypothesis, which assumes that the hypothesized factor structure fits the data structure. Critical items can be identified and potentially get deleted based on the communalities statistics, i.e., the extent to which an item correlates with all other items.

Using all remaining items of controllability resulted in a significant goodness-of-fit test, indicating a poor fit for the single factor solution. Thus, based on low communalities, the items CN5, CN8, and CN9, as partially suspected based on the initial consistency assessment, were removed. After their removal, Bartlett's test of sphericity revealed the desirable significant χ^2 statistic, $\chi^2(6) = 1272.68$, $p < .001$. The KMO exceeded the minimum value of 0.50 and at 0.867 was regarded as great. The goodness-of-fit test showed that the assumed one-factor structure fits the data well, $\chi^2(2)$=4.83, p=.09.

For responsiveness, for no combination of items without excluding RE4 and RE6, a one-factor solution could be found as indicated by the significance of goodness-of-fit statistics. To identify the items not loading on the same factor, a ML factor analysis with two factors was calculated as this assumption was visible due to the inter-item correlation and consistency analysis. Here, a solution using the items RE2, RE3, and RE5 for one factor and RE4 and RE6 as a second factor was found, Bartlett's $\chi^2(10)$=973.64, $p < .001$, KMO of 0.755, goodness-of-fit $\chi^2(1)$ of 0.10, $p = .92$. As the two items RE4 and RE6 conceptually describe the temporal feedback of the game, this factor will be called immediate feedback for the remainder of this work whereas the other items form the factor responsiveness.

Lastly, for performance indication, as indicated by the lowest inter-item correlation, the items PI8 (communality of .49) and PI7 (communality of .72) were removed due to the significant goodness-of-fit test results, $p < .001$. Even though a 5-item solution was revealed, $\chi^2(2)$ of 4.69, $p = .096$, removing PI2 lead to a very satisfying one-factor solution using the items PI1, PI3, PI4, and PI5 resulting in a KMO of 0.86, Bartlett's $\chi^2(6)$=743.67, $p < .001$, goodness-of-fit $\chi^2(2)$ of 2.48, $p = .29$.

To finalize the construction of input quality factors and the modifying performance indication factor, a CFA for calculated for the found items using the training dataset. For each construct, an excellent fit was confirmed by a selection of fit indices. Although there are many fit indices that can be used, some of the most popular and useful are the Comparative fit index (CFI), Standardized root mean square residual (SRMR), and Root mean square error of approximation (RMSEA) [223] as well as the Minimum discrepancy (CMIN) known as χ^2 value of the model, minimum discrepancy per degree of freedom (CMIN/df), and P of Close Fit (PClose). Furthermore, for reliability measures, the Composite reliability (CR), Maximal reliability (MaxR(H)), also known as McDonald construct reliability, as well as Item reliability (IR) are often used. For convergent validity measures, Average variance extracted (AVE) should be greater than 0.5. Table 5.1 and Table 5.2 show the model fit measures as well as model validity and item reliability measures for all constructs. The provided thresholds in both tables for these criteria are derived from [237]. With exception of RMSEA for controllability (RMSEA = 0.079 reached an acceptable level), all fit and model validity indices are excellent. More information about the various fit indices can be found in [205], [230], [238], [239]. The measurement models including the standardized factor loadings and item reliabilities for each factor are shown in Fig. 5.2. The straight line pointing from a latent variable to the indicator variables suggest a causal effect of the latent variable on the observed variables whereas the double-headed arrows between latent variables indicate a correlation among them [236]. As performance indication as a mediating factor is treated as a single factor, the model development phase will only continue with the input quality factors.

Table 5.1: Model fit measures based on CFA on factor level.

Measure / Factor	CMIN	CMIN/df	CFI	SRMR	RMSEA	Pclose
Controllability	4.89	2.45	0.99	0.01	0.08	0.22
Responsiveness and Immediate Feedback	5.47	1.37	0.99	0.02	0.04	0.50
Performance Indication	2.53	1.26	0.99	0.01	0.04	0.41
Threshold Excellent	–	<3	>0.95	<0.08	<0.06	>0.05
Threshold Acceptable	–	>3	<0.95	>0.08	>0.06	<0.05
Threshold Terrible	–	>5	<0.90	>0.10	>0.08	<0.01

Table 5.2: Model validity and item reliability measures based on CFA on factor level.

Measure / Factor	CR	AVE	MaxR(H)	IR_min	IR_max
Controllability	0.97	0.88	0.98	0.73	0.94
Responsiveness	0.94	0.83	0.96	0.67	0.92
Immediate Feedback	0.90	0.82	0.97	0.68	0.97
Performance Indication	0.96	0.86	0.98	0.74	0.95
Threshold Acceptable	>0.6	>0.5	>0.7	>0.4	>0.4

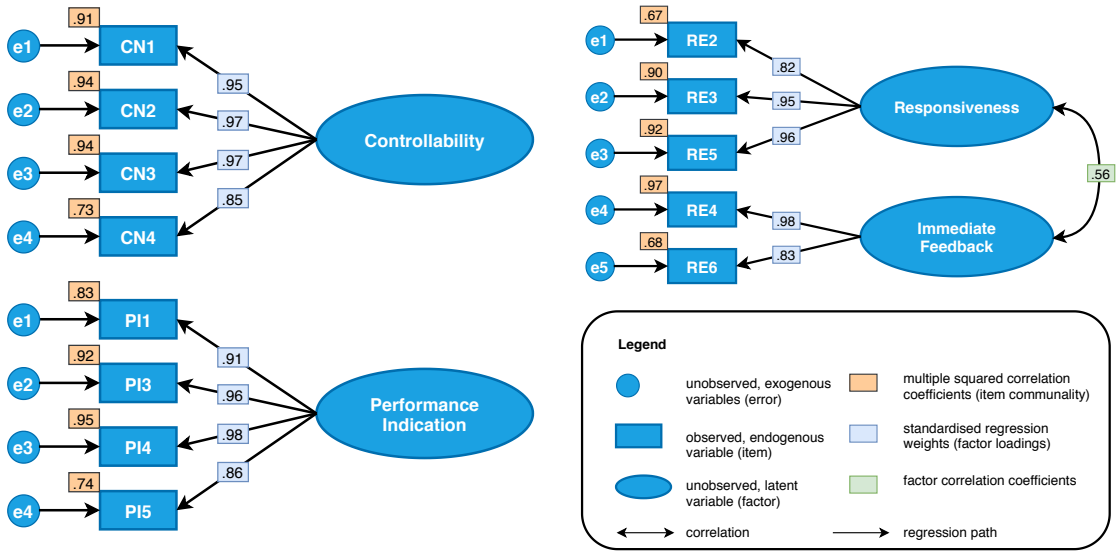


Figure 5.2: Measurement models of input quality aspects as well as performance indication.

Global Measurement Model for Input Quality

After a satisfying structure of each individual construct was found in the previous steps, an EFA with ML and promax rotation was performed for all remaining input quality items using the training dataset.

A promax rotation was used since on the global level, i.e., considering all factors that contribute to the input quality at the same time, an orthogonal rotation is often not an ideal choice. The reason is that any factor forming the input quality might be to some extent related to other factors, and thus, arbitrarily forcing those factors to be orthogonal may distort the findings. Even though an orthogonal rotation is generally less prone to sampling errors, and thus more replicable, using a large sample size usually addresses this concern [235].

For the analysis, first, a three-factor solution assuming the factors controllabilty, responsivness, and immediate feedback was investigated. The items RE2 and RE6 showed slightly lower communalities

(< .7) and CN4 had moderate cross loading to a second factor (.52 together with RE2, RE3, and RE5). While removing RE2 or RE6 was not advantageous, the removal of CN4 led to an appropriate 3-factor solution, $\chi^2(7) = 9.85$, $p = .20$.

However, the pattern matrix showed a loading larger than 1 for RE6. This might be caused by the promax rotation used or due to the low item number for the corresponding factor. Such a phenomenon is also called a Heywood case. To resolve this issue, a Principal axis factoring (PAF) analysis was used instead of the ML method, which is particularly prone to the occurrence of Heywood cases [230]. Bartlett's test of sphericity revealed the desirable significant χ^2 statistic, $\chi^2(28) = 2311.31$, $p < .001$. The KMO exceeded the minimum value of 0.50 and at 0.89 was regarded as close to superb. The goodness-of-fit test showed that the assumed three-factor structure fits the data well, $\chi^2(7) = 9.85$, $p = .20$, explaining 86.6 percent of the variability in the data. In Fig. 5.3 the pattern matrix illustrating the regression coefficients for each variable on each input quality factor is shown.

	Factor		
Item	1	2	3
RE2	.85		
RE3	.77		
RE5	.83		
CN1		.88	
CN2		.91	
CN3		.86	
RE4			.86
RE6			.92

Table 5.3: Pattern matrix of input quality factors (cross loadings lower .3 are omitted). Interpretation of factors: 1) responsiveness, 2) controllability, 3) immediate feedback.

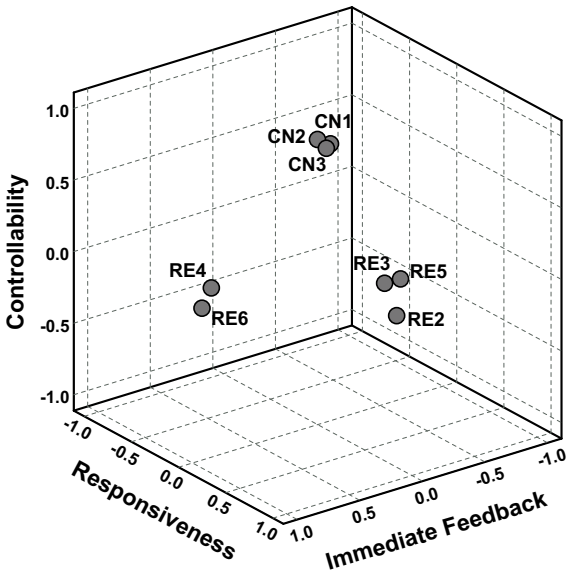


Figure 5.3: Factor plot of input quality in rotated factor space based on PAF analysis.

It can be observed that no cross-loadings larger than 0.3 exist. Furthermore, the lowest loading is greater than the desired minimum of 0.5 and the average loadings per factor is higher than 0.7. This is an indication of a good convergent validity. However, it must be noted that a high correlation between the controllability and responsiveness factor (r=.82) was found, whereas the correlation to immediate feedback factor was .60 and .52, respectively. Consequently, discriminant validity is not ideal for the suggested structure but based on the EFA result sufficient enough. This is also illustrated in Fig. 5.3 showing the factor plot of input quality. A dataset containing conditions such as different input devices or various packet loss strategies could increase the distinction between those factors.

After a theoretically-based factor structure is borne out of the EFA, a confirmatory factor analysis was conducted using the training dataset. A CFA investigates whether underlying latent factors truly "cause" the variance in the observed variables (items). The global model was evaluated using the fit indices introduced before. The model was constructed in IBM's AMOS and reached an excellent fit. The model fit measures are presented in Table 5.4 and the measurement model itself is presented in Fig. 5.4. Table 5.4 shows that the cut-off values recommended by Hu and Bentler [237], CFI $\geq .95$, and SRMR $\leq .08$, are fulfilled excellently.

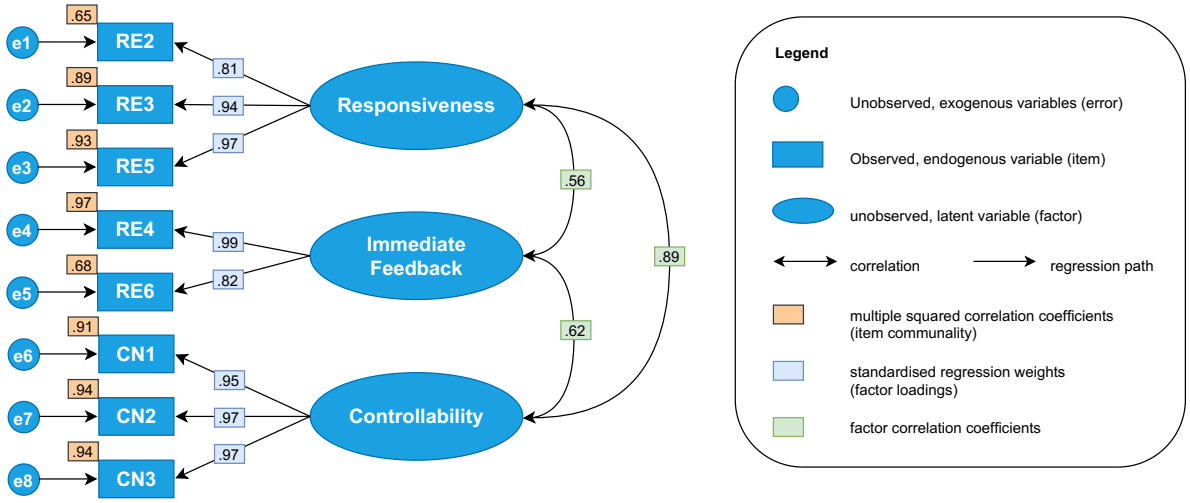


Figure 5.4: Global measurement model of Input Quality.

Table 5.4: Model fit measures of Input Quality model.

Measure	Estimate	Threshold	Interpretation
CMIN	27.946	–	–
CMIN/DF	1.644	Between 1 and 3	Excellent
CFI	0.995	>0.95	Excellent
SRMR	0.018	<0.08	Excellent
RMSEA	0.052	<0.06	Excellent
PClose	0.415	>0.05	Excellent

Additionally, the model reliability and validity measures summarized in Table 5.5 satisfy the recommended thresholds. As now multiple factors are used for the CFA, also the Maximum shared variance (MSV) and MaxR(H) are reported. The CR of all latent constructs is greater than 0.70 and the AVE exceeded 0.50 showing good construct reliability and convergent validity, respectively. Finally, discriminant validity is acceptable, indicated by MSV being smaller than AVE as well as the square root of the AVE being greater than the other inter-construct correlations. The latter fit comparison is also known as the Fornell-Larcker criterion [240].

Table 5.5: Model reliability and validity measures of input quality model. The square root of AVE is shown on diagonal in bold faces.

	CR	AVE	MSV	MaxR(H)	Inter-construct correlations		
					RE	IF	CN
Responsivness (RE)	0.934	0.826	0.797	0.960	**0.909**		
Immediate Feedback (IF)	0.904	0.825	0.387	0.975	0.558	**0.909**	
Controllability (CN)	0.975	0.928	0.797	0.976	0.893	0.622	**0.963**

Final Consistency Assessment

After the exploratory and confirmatory factor analyses have been conducted, resulting in the removal of unsuitable items and establishment of an appropriate model structure, the internal consistency reliabilities for the complete GIPS were calculated. As the unidimensionality of individual scales has been established through the factor analyses previously conducted, items could get deleted if

this would improve or not negatively impact the reliability of the scale [224]. However, no major improvement was found for the remaining items and the number of items per factor is already low. This was initially desired as the questionnaire will likely be used very often in combination with other questionnaires. Too many items may have a negative effect on the quality of measurements. The reliability analysis showed that the GIPS reaches a very good internal consistency. The Cronbach's α for responsiveness, immediate feedback, and controllability are .932, .891, and .975, respectively. The additionally developed scale measuring performance indication achieved a Cronbach's α of .958. Thus, the sampling domain has adequately been captured. It should be emphasized that even though the factors included only two to three items each, the content adequacy assessment and factor analyses helped retain items that were consistent with the corresponding construct domain [224].

5.5 Validation of GIPS

To ensure that the model structure is not sample-dependent, the developed GIPS is validated in two steps. First, the test dataset, which resulted from the crowdgaming studies described in the previous chapter and which comprises 379 samples, was used for the validation. Second, three additional datasets, in which only the 8 remaining items of the input quality factors are used, were created. These three dataset involve the two datasets, Dataset 4.3 (CS) and Dataset 4.3 (Lab), which were presented in Section 4.3, as well as a large-scale dataset called G.1072, which was used for the development of the opinion model predicting gaming QoE for cloud gaming services, ITU-T Rec. G.1072. For the latter, apart from various video encoding settings, also network impairments such as delay and packet loss are included. However, the packet loss was simulated at the down-stream of a cloud gaming service, thus, resulting in jerkiness in the video rather than loss of input packets as it was simulated in the crowdgaming studies. The test design of this dataset was already explained in Section 3.3.2. However, it must be noted that conditions only causing spatial video artifacts, e.g., low bitrates or low resolutions were not considered for die GIPS validation. More information about the datasets can be found in Table 5.6 as well as in Section 7.1. Lastly, it should be mentioned that for the validation analysis no outliers for the individual ratings should be included. Thus, a slightly stricter data cleansing was performed compared to the analysis presented in Section 4.2.3 and [34].

Table 5.6: Details about the datasets used for the validation of GIPS.

	Test Dataset	Dataset 4.3 (CS)	Dataset 4.3 (Lab)	G.1072
Samples	379	656	272	1531
Participants	214	131	20	142
Environment	home	home	lab	lab
N_games	6	2	2	9
Framerate [fps]	60, 30, 10	60, 30, 10	60, 30, 10	60, 30, 10
Jitter [ms]	0	[0:50:100]	[0:50:100]	0
Delay [ms]	[0:150:300]	[0:100:300]	[0:100:300]	25, 50, 100, 200, 400
Up-link packet loss [%]	0, 10, 30	0	0	0
Down-link packet loss [%]	0	0	0	0, 0.1, 0.2, 0.5, 1, 2

The model fit measures, as well as the reliability and validity measures of the input quality model for the validation dataset, are shown in Table 5.7 and 5.8, respectively. It can be observed that the model largely matches the results of the training sample. The fit measures are all excellent with the exception of RMSEA being in an acceptable range for the Dataset 4.3 (Lab). The recommended combination of CFI > 0.95 and SRMR < 0.08 by Gaskin [241] is fulfilled for all four datasets. However, discriminant validity issues are revealed for responsiveness in the Dataset 4.3 (CS) and G.1072 dataset, as the square root of the AVE is less than its correlation with controllability and immediate feedback. This was partially expected as all factors were assumed to be related. Though Malhotra and Dash argue that AVE is often too strict, and reliability can be established through CR alone [239], [241]. A reason for the slightly too low badness-of-fit measures RMSEA for the G.1072 dataset might be the high number of dependent variables and the different types of packet loss being used. It must be noted that removing the items RE2 and CN1 would result in an excellent model fit for the G.1072 dataset (and all others) but would not solve the discriminant validity issues. For completeness, the fit measures for G.1072 are also provided in Table 5.7.

Table 5.7: Model fit measures of input quality model for the validation dataset.

	Test Dataset	Dataset 4.3 (CS)	Dataset 4.3 (Lab)	G.1072 (8 items)	G.1072 (6 items)
CMIN	37.66	52.62	42.33	210.38	30.84
CMIN/DF	2.22	3.10	2.49	12.38	5.14
CFI	0.994	0.995	0.992	0.990	1.00
SRMR	0.02	0.01	0.01	0.01	0.00
RMSEA	0.057	0.057	0.074	0.086	0.052
PClose	0.30	0.25	0.07	0.00	0.39
IRmin	0.58	0.73	0.81	0.67	0.88
IRmax	0.94	0.98	0.95	0.95	0.95

A key idea of a questionnaire validation is to investigate how the focal construct (and its specific factors) is related to other constructs. Based on the theory of the construct, it can be stated what the construct should be positively, negatively, and relatively independent related to. This so-called nomological network is important for the validation process, as the fulfillment of expected relationships to other established measures is an indicator for convergent and divergent validity [223]. In the previous chapter, in which the developed crowdgaming framework was tested, the relationship of the three input quality factors with the overall gaming QoE was already illustrated in great detail. The input quality rating followed the expected trends as the subjective scores degraded in case of a simulated network delay, reductions of framerates, or the presence of input packet losses (up-link). Furthermore, the expected differences between the sensitivity of games were shown. In addition to these findings, which support the predictive validity of the GIPS, no significant changes in scale scores over repeated occasions could be observed. Thus, the reliability and stability of the operationalized construct were shown. In the G.1072 dataset, the GIPS was used in addition to various other quality aspects which are part of the taxonomy presented in Chapter 2. As in the G.1072 data also ratings of the iGEQ, which measures the PX, are contained, the correlations of the input quality factor and iGEQ can be analyzed. It is expected that for the positive PX aspects such as immersion or positive affect, a positive correlation can be found, whereas a negative correlation with challenge, negative affect, and tension exists. The Pearson correlation coefficients are summarized in Table 5.9. It can be observed that the assumptions

Table 5.8: Model reliability and validity measures of input quality model for the validation datasets. The square root of AVE is shown on diagonal in bold faces.

Dataset		CR	AVE	MSV	MaxR(H)	RE	IF	CN	Cronbach's α
Test Dataset	RE	0.95	0.87	0.80	0.97	**0.93**			.95
	IF	0.85	0.74	0.36	0.92	0.58	**0.86**		.82
	CN	0.96	0.90	0.80	0.96	0.90	0.60	**0.95**	.96
Dataset 4.3 (CS)	RE	0.95	0.87	0.92	0.97	**0.93**			.95
	IF	0.97	0.94	0.46	0.98	0.68	**0.97**		.97
	CN	0.97	0.92	0.92	0.98	0.96	0.64	**0.96**	.97
Dataset 4.3 (Lab)	RE	0.95	0.87	0.79	0.97	**0.93**			.95
	IF	0.96	0.92	0.79	0.97	0.89	**0.96**		.96
	CN	0.97	0.90	0.79	0.97	0.89	0.80	**0.95**	.96
G.1072 (8 items)	RE	0.93	0.83	0.97	0.95	**0.91**			.93
	IF	0.97	0.95	0.97	0.97	0.98	**0.97**		.97
	CN	0.97	0.92	0.92	0.97	0.96	0.93	**0.96**	.97
G.1072 (6 items)	RE	0.95	0.90	0.97	0.95	**0.95**			.95
	IF	0.97	0.95	0.97	0.97	0.99	**0.97**		.97
	CN	0.96	0.92	0.93	0.96	0.97	0.93	**0.96**	.96

are largely supported. Even though for challenge only negligible correlations and for flow only low correlations are found, all other PX aspects are moderately or highly correlated with the input quality aspects.

Table 5.9: Correlations (Pearson's r) between GIPS sub-scales (highlighted in bold) and iGEQ sub-scales (N = 1531). All shown correlations higher than .07 are significant at the 0.01 level (2-tailed).

	RE	**CN**	**IF**	IM	CO	FL	PA	CH	NA
Responsiveness (RE)	1	-	-	-	-	-	-	-	-
Controllability (CN)	0.90	1	-	-	-	-	-	-	-
Immediate Feedback (IF)	0.93	0.90	1	-	-	-	-	-	-
Immersion (IM)	0.62	0.60	0.60	1	-	-	-	-	-
Competence (CO)	0.62	0.68	0.62	0.69	1	-	-	-	-
Flow (FL)	0.43	0.43	0.41	0.71	0.57	1	-	-	-
Positive Affect (PA)	0.69	0.71	0.68	0.76	0.83	0.64	1	-	-
Challenge (CH)	-0.17	-0.23	-0.20	0.08	-0.12	0.19	-0.06	1	-
Negative Affect (NA)	-0.53	-0.55	-0.53	-0.58	-0.57	-0.48	-0.64	0.02	1
Tension (TE)	-0.69	-0.73	-0.69	-0.58	-0.65	-0.40	-0.71	0.28	0.69

To finalize the analysis, also the model derived for the performance indication factor was validated using the test dataset. The fit indices are summarized in Tables 5.10. All indicators show an excellent fit. In addition, the reliability and validity measure also fulfill the previously mentioned criteria: CR = 0.95, AVE = 0.83, MaxR(H) = .96, IR_min = 0.77, and IR_max = 0.92. An analysis of the full post-game questionnaire used for the creation of the G.1072 database, in which also the performance indication factor is included, will be given in Section 7.2.

Finally, Table 5.11 contains the list of the items remaining in the final version of the GIPS.

Table 5.10: Model fit measures of performance indication model for the test dataset.

Measure	Estimate	Threshold	Interpretation
CMIN	3.27	–	–
CMIN/DF	1.63	Between 1 and 3	Excellent
CFI	0.998	>0.95	Excellent
SRMR	0.01	<0.08	Excellent
RMSEA	0.058	<0.06	Excellent
PClose	0.34	>0.05	Excellent

Table 5.11: List of final items of the GIPS. The index *i* indicates an inverted item. In addition, the position of the item in the used questionnaire (No.), as well as the initial source of the items (Ref), is provided. An asterisk attached to the source indicates an adaptation of the item by experts.

Code	Label	No.	Ref
Controllability			
CN1	I felt that I had control over my interaction with the system.	3	[115]
CN2	I felt a sense of control over the game interface and input devices.	6	[165],[242]*
CN3	I felt in control of my game actions.	9	[163]*
Responsiveness			
RE2$_i$	I noticed delay between my actions and the outcomes.	4	[243]*
RE3	The responsiveness of my inputs was as I expected.	7	[163]*
RE5	My inputs were applied smoothly.	13	self
Immediate Feedback			
RE4	I received immediate feedback on my actions.	10	[244],[165]
RE6	I was notified about my actions immediately.	16	[165]*
Performance Indication			
PI1	I could easily assess how I was performing in the game.	2	[149]
PI3	I was aware of how well I was performing in the game.	8	[244]
PI4	It was clear to me how my performance was going.	11	[168]*
PI5	I was informed about my progress in the game.	14	[149]*

5.6 Summary

In this chapter, the development of the Gaming Input Quality Scale (GIPS) is described. The GIPS aims to measure the input quality of a player interacting with a gaming system. As no measurement tool for the input quality could be identified based on the literature presented in Chapter 3, a psychometrically validated instrument for assessing this quality aspect was required. Instead of using self-developed single item scales such as the one used in [80] to assess the perception and annoyance of delay regarding control input, controllability of a game, or the reactiveness item used in the study presented in Section 3.2 [15] and Section 6.1 [14], a validated questionnaire will offer the means to collect reliable and comparable study results.

The development of the GIPS included the following steps

- An item generation resulting in an initial item pool of 41 items based on various available questionnaires and expert interview
- A data collection using the crowdgaming method presented in Chapter 4

- A multi-method approach published by Homburg and Giering applying an item screening and development of a measurement model using an EFA and CFA
- A validation of the identified items and its underlying items structure based on a test dataset and three other independent datasets

Finally, the GIPS comprises the factors responsiveness and immediate feedback, as well as controllability measured with a total of 8 items. The low number of items was targeted as the GIPS should be combined with other questionnaires, by which a very high number of resulting items would provoke fatigue of respondents. In addition, a factor called performance indication was derived which may serve as a moderating variable between network impairments and input quality.

The GIPS achieved an excellent model fit as well as reliability and validity measures. However, it must be noted that for the validation dataset, discriminant issues were shown between controllability and responsiveness. As those datasets do not include noticeable packet loss on the control stream, it can still be concluded that both factors present valuable insight for studies in which they may be more distinct. The factors were least distinct for the G.1072 dataset, which is significantly reduced compared to the one used later in Chapter 7. Finally, it was shown that the GIPS factors change accordingly to all expectations in case of common network and encoding impairments as shown in Chapter 4, and are correlated as expected with PX aspects assessed using the iGEQ.

Chapter 6

Impact and Classification of the Game Content

As summarized in Chapter 2, there is a multitude of influencing factors on gaming QoE. While already many important findings are achieved by the research community, which also led to the ITU-T Rec. G.1032, there are still many issues to address. One very dominant influencing factor is the *game* used as a test stimulus. Thus, in this chapter, research about this important factor is presented.

The first section targets the comparison of different game scenarios with regards to their delay sensitivity. The work is mainly based on a publication presented in [14].

Section 6.2 continues this work and presents a first version of a game content classification with respect to the impact of network and encoding parameters. This research is based on work presented in [17], [18], [28], and also resulted in an Annex for the ITU-T Rec. G.1072 [203].

6.1 Impact of Game Scenario (Study 6.1)

As described in Section 2.2.2, several previous studies indicated that not every game is equally sensitive towards delays. Delay sensitivity is defined as a change in the gaming QoE in relationship to a change in delay. To classify games, for marketing purposes, the well-known classification of genres, which categorizes games into different classes such as action, fight, flight, shooter, strategy, and sports games, is often used. However, some researchers critically discussed the usage of the genre classification since it is not accurate enough due to the strong overlapping of genres [245], [246]. It would be questionable to argue that one genre is more sensitive than another one when it is possible that the same tested game belongs to both genres. In this section, an even more critical view is considered, namely that the game scenario, i.e., a timely limited section of a game (referred to a scene in the video community), should be classified with its underlying characteristics for research studies instead of a game itself in order to achieve comparable and generally valid research results. As a consequence, this also means that the genre classification should not be used to model the impact of delay on gaming QoE.

In their research about the influence of delay on gaming QoE in [80], Sackl et al. proposed a list of metrics in order to classify games with a special focus on delay sensitivity. The listed metrics contained the required number of actions, maximum successful time, reaction time, and predictability of actions. In addition to these metrics, some other attempts have been undertaken with the aim of identifying key game characteristics responsible for making games sensitive to delay. Claypool determined the sensitivity of games to delay based on two factors, interaction and perspective [247]. Interaction is

characterized by the deadline and precision model. Deadline, referred to as temporal accuracy in the following, is defined as the time required to complete an action [247]. Precision, referred to as spatial accuracy, is the degree of accuracy required to complete the interaction successfully [247]. The game perspective specifies whether the game is based on an omnipresent model or an avatar model, in which the first-person and third-person perspectives are available. Claypool showed that the first-person avatar is significantly more sensitive to delay compared to third-person avatar perspective in a shooting game, while the used omnipresent game was less sensitive than avatar games [248]. While an investigation related to delay sensitivity is still missing, Djaouti et al. [228] defined and proposed a set of game rules in order to classify games. These elementary rules, named "gameplay bricks," are key elements of a game that describe its gameplay. The authors presented ten bricks split into 1) rules determining the goals, including avoid, match, and destroy, and 2) rules defining the means and constraints to reach the goals consisting of create, manage, move, random, select, shoot, and write. For instance, a simple racing gameplay could be defined by the rules of avoid (obstacles), match (racing line) and move (control the car). The previously described research by Claypool and Djaouti et al. will be considered for the selection of game scenarios described in the following section.

6.1.1 Method

To prove the hypothesis that a change in characteristics of a scenario within the same game can lead to strong differences in regard to the delay sensitivity, an empirical study was conducted.

Therefore, three games, each with two different scenarios, were selected in a way that the games are different in their elementary rules. The scenarios differed in only a very limited set of characteristics. In particular, the following criteria for the selection of the games and scenarios were considered. Firstly, the games should differ in their elementary rules (gameplay bricks) as it is assumed that certain rules influence the delay sensitivity. Secondly, a single-player mode was desirable to avoid social influences. Furthermore, a low complexity regarding the number of control keys and game goals was chosen to avoid a time-consuming training session and to reduce learning effects. At last, the games should offer a change in pace or perspective, since those are considered as important characteristics for the delay sensitivity as shown in [247]. Finally, the following games were used in this study:

- GTA 5: an action game offering a shooting range scenario with static (first scenario) and moving (second scenario) targets – game bricks: destroy/shoot
- Rayman Legends: a platformer game in which the player has to jump over static (first scenario) and oncoming (second scenario) obstacles – game bricks: avoid/move
- Project Cars: a racing simulator game offering a change from first-person (first scenario) to third-person (second scenario) camera – game bricks: avoid/match/move

While among the GTA and Rayman scenarios, the pace was strongly different but not the performed actions, in Projects Cars solely the perspective was changed. In order to investigate the impact of delay, a study using a within-subject design with simulated delay conditions of 0 ms (reference condition), 100 ms, 200 ms, 300 ms, and 400 ms was conducted. This resulted in a total of 30 conditions. While each game was tested one at a time, the delay conditions were randomized using a Latin square design for each scenario to reduce learning effects. All games were played in a resolution of 1080p and with a framerate of 30 fps. Twenty-nine participants (15 females, 14 males), aged between 19 and 46 (median age 28) took part in the study. Most of them were students and non-expert gamers, but familiar with controlling similar games and knowledge about network delay in gaming (23 of 29).

As the study was conducted in 2017, the knowledge about standardized test designs as presented in Section 3.3 and methods of the ITU-T Rec. P.809 were not available at that point in time. However, the test room conditions closely adhered to the ITU-T Rec. P.910 [184] and P.911 [249]. While already an electronic questionnaire using the 7-point extended continuous scales, the AutoHotkey script to set up the conditions, as well as a stimulus duration of 90 seconds were used, the study was different compared to the design presented in Section 3.2 in the following aspects.

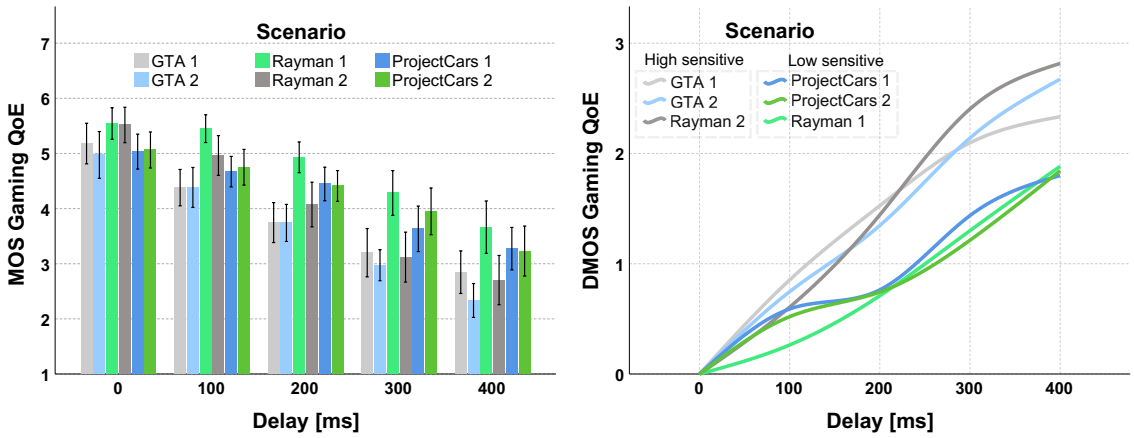
Firstly, there was no network emulator and cloud gaming system used to introduce the network delay, but the delay was artificially added using an Arduino microcontroller[1], which buffered the received commands of the input device for the delay duration. Consequently, the direct input delay simulates a network delay in a cloud gaming system. Secondly, the used measurement tools and instructions were strongly reduced. As no video streaming was performed, instructions regarding the video quality dimensions (cf. Section 3.3.1) were omitted. To judge the overall gaming QoE, participants were asked: "How do you rate the overall quality?". The perception of delay was measured by asking: "How frequently did you notice delayed reactions of the game?" (always to never). Using agreement labels ranging from "strongly disagree" to "strongly agree", the annoyance ("Delayed reactions of the game annoyed me."), the control ("I had control over the game."), the fairness ("I perceived the scenario under these conditions as fair."), the difficulty ("The scenario under these conditions appeared difficult to me."), and a self-judgement of playing performance ("How do you rate your own performance?") were measured. Lastly, the willingness to continue playing ("Would you play the game under these conditions again?" using the labels extremely unlikely to extremely likely) was assessed.

6.1.2 Results

Comparison of delay impact between games and scenarios

In the following, a comparison of the impact of the delay on the gaming QoE ratings between the games and between the scenarios will be presented. For the analysis, also the differential MOS (DMOS), calculated by the difference between the reference condition and each test condition, was used. This step, which also reduces the influence of other quality factors, was performed since already small differences between the games at the reference condition existed. To illustrate the impact of delay in more detail, the MOS and DMOS for all scenarios and delay conditions is shown in Fig. 6.1. The figure exhibits that the delay negatively affected the gaming QoE for every scenario. However, for Rayman both scenarios were differently affected by the delay even at low delays. Also, it can be seen that both Project Car scenarios, which are only different in their perspective, and the first scenario of Rayman, where obstacles were static, resulting in a slower pace and higher necessary temporal accuracy compared to the second scenario, form a group of weakly delay influenced scenarios. The other three scenarios, the fast-paced Rayman scenario, and both GTA scenarios, are strongly sensitive towards delay. Furthermore, it can be observed in Fig. 6.1b that at a delay of 200 ms, the separation of the previously described clustering of weakly and strongly delay sensitive scenarios is most distinct. This finding is in line with the study results in [80], where a significant delay influence was observed for delay higher than 180 ms. This knowledge could be used to reduce the number of delay conditions for future studies that aim at investigating the mediating effect of game characteristics on delay sensitivity (cf. Section 6.2).

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



(a) Bar plots of means and 95 % confidence interval.

(b) Means of DMOS of gaming QoE.

Figure 6.1: MOS and DMOS of gaming QoE for all game scenarios under the simulated delay conditions.

In order to investigate whether the game or the scenario is more important in regard to the delay sensitivity, a three-way repeated measure ANOVA using the *game* (N = 3), the *scenario* (N = 2), and the delay (N = 5) as independent variables, and the DMOS of the *gaming QoE* as the dependent variable, was calculated. The ANOVA revealed an interaction effect of game and scenario, $F(2,58) = 8.30$, $p < .001$, $\eta_p^2 = .22$. This shows that there is a difference between the scenarios for one game, but this does not apply to all games. While the difference between the first scenario of GTA ($M = 1.36$, $SD = 1.17$) and the second scenario of GTA ($M = 1.38$, $SD = 1.07$) was very small, the difference between the first Rayman scenario ($M = 0.83$, $SD = 0.88$) and the second Rayman scenario ($M = 1.45$, $SD = 1.23$) was high. The results reveal that it cannot be generalized that the game Rayman as a whole is weaker or stronger sensitive towards delay than the game GTA, since this depends on the chosen scenario. The same holds true for the game Project Cars, where the first scenario ($M = 0.92$, $SD = 1.01$) was similarly affected by delay than the second scenario ($M = 0.86$, $SD = 0.94$). For the main effect of game, $F(2,58) = 10.46$, $p < .001$, $\eta_p^2 = .27$, a pairwise comparison showed that the delay influence was lower for Project Cars compared to the other two games, $p < .001$. Not surprisingly, also a very strong main effect of delay was found, $F(1.57,45.63) = 114.25$, $p < .001$, $\eta_p^2 = .80$, $\varepsilon = .52$ (Greenhouse-Geisser correction was applied).

Influence of delay on assessed quality features

So far, it was shown that the change of pace (from static to oncoming obstacles) in Rayman caused strong differences in respect to the delay sensitivity based on the overall gaming QoE. However, this behavior was not observed due to the change of pace in GTA (static to moving targets). By including the assessed quality aspects to the analysis, the cause of this contradiction can be investigated. Therefore, a two-way repeated measure ANOVA with the independent variables *scenario* and *delay* was performed for each game separately. The results of the main effect of the game scenario are provided in Table 6.1 for every assessed quality aspect.

For GTA, the two-way ANOVA yielded only a tendency for a main effect of the scenario on the gaming QoE and willingness to continue playing. However, a very strong difference between both scenarios was observed for the quality aspects difficulty, control, and judgment of own performance. Furthermore, a moderately strong difference between both scenarios was revealed for fairness.

Table 6.1: Results of ANOVA for the main effect of the game scenario (significant differences are highlighted in bold). The column labeled as *start* indicates the lower delay value in ms for which an effect was found.

Quality Aspect	GTA				Rayman				Project Cars			
	F(1,28)	p	η_p^2	start	F(1,28)	p	η_p^2	start	F(1,28)	p	η_p^2	start
Gaming QoE	2.46	.12	.08	400	21.53	**<.001**	.44	100	0.53	.47	.02	-
Perception	0.14	.71	.01	-	21.89	**<.001**	.44	100	0.47	.50	.02	-
Annoyance	0.33	.57	.01	-	18.14	**<.001**	.39	100	0.34	.57	.01	-
Difficulty	21.19	**<.001**	.43	200	29.58	**<.001**	.58	100	1.68	.20	.06	-
Control	13.15	**<.001**	.32	300	37.98	**<.001**	.51	100	4.20	**<.05**	.13	400
Fairness	6.42	**<.05**	.19	400	38.05	**<.001**	.58	100	2.28	.14	.08	-
Performance	24.00	**<.001**	.46	0	21.01	**<.001**	.43	100	1.80	.19	.06	-
Continue	3.09	.09	.10	400	32.46	**<.001**	.54	100	2.44	.13	.08	-

Concerning Rayman, an interaction effect of scenario and delay for all quality aspects except for annoyance and judgment of own performance was determined. However, also for these two quality aspects, a significant main effect of scenario was observed. A paired comparison of the reference and 100 ms delay conditions showed that there was no difference for the first scenario, but for the second scenario. Additionally, when comparing the delay conditions 300 ms and 400 ms, there was no difference for the second scenario, but in the first scenario for a majority of quality aspects. This illustrates that the chosen conditions are in a good range since at the higher end saturation seems to be present, while a different behavior between the scenarios can already be observed for the lower end.

For the game Project Cars, the ANOVA showed a significant effect of scenario on the control ratings. The second scenario, using a first-person perspective, was rated higher. This is in line with reports of the participants in the interviews. Many participants stated that they have a clear preference regarding the camera perspective, which influences especially their control over the game.

Based on these results, it can be concluded that the control ratings were significantly impacted for all games, but only for Rayman, where also the annoyance and perception of delay were significantly affected, a strong change in the gaming QoE was observed. In fact, a simple multiple linear regression using all game scenarios, as shown in Table 6.2, revealed that 76.9 % (adjusted R-squared) of the variability in the gaming QoE can be explained by the perception of delay, perceived difficulty, and control. When comparing the regression weights between the different game scenarios, it can be observed that only for the fast-paced scenarios control was very impactful. For the slow-paced scenarios, apparently participants were still able to control the game, but their experience still suffered from the introduced delay.

6.1.3 Discussion

The obtained results revealed that for Rayman the higher pace, corresponding with a change in temporal accuracy and a higher number of necessary decisions, resulted in a strong difference between both scenarios of the same game. However, both GTA scenarios were similar in respect to their delay sensitivity, even though the pace changed as well. However, for the low paced (first) GTA scenario the delay influence was also strong. An explanation can be derived from the regression analysis, which showed that the perception of delay seems to explain the judgement of the gaming QoE for a delay

Table 6.2: Multiple linear regression predicting gaming QoE based on the most dominant quality aspects.

Scenario	Adaptation	Adj. R^2	Unstandardized Regression Weight			
			Intercept	Control	Perception	Difficulty
GTA1	slow	.78	6.48	0.13	-0.57	-0.21
GTA2	fast	.81	4.47	0.39	-0.40	-0.16
Rayman 1	slow	.64	6.48	0.14	-0.40	-0.31
Rayman 2	fast	.83	5.92	0.32	-0.36	-0.39
PCars1	1st person	.67	5.64	0.22	-0.46	-0.17
PCars2	3rd person	.78	5.98	0.20	-0.53	-0.19
All	-	.77	5.99	0.22	-0.48	-0.24

degradation the most. A possible reason why the delay was perceived strongly in both GTA scenarios might be that for GTA a mouse cursor had to be moved in order to point at targets, whereas in Rayman solely single keystrokes are required to move the character. The instantaneous permanent feedback of the mouse cursor could be a reason for a higher perception of delay. This indicates that for the development of a delay sensitivity classification, the game characteristics modified in this study could be valuable. A comparison of the Project Cars scenarios showed no differences in any quality aspect but control, which might have changed due to the preferences of players. However, it should not be generalized that the perspective is not an influencing factor for delay, since Claypool proved its influence in a shooting scenario [248]. A variety of games similar to Project Cars exist, e.g. Mario Cart, that typically possess a higher reaction time of the vehicle. Project Cars as a realistic racing simulation exhibits a certain inertia of the car. When comparing the results for Project Cars with the study conducted by Jarschel et al. [79], it can be observed that the delay influence of Project Cars was considerably weaker than in the racing game (Gran Tourismo HD Concept) used in the mentioned study. In addition to the reaction time of the car, also the width of the road or consequences in case of errors (leaving the road) could change the delay sensitivity in a racing scenario, as shown in [250].

6.2 Game Content Classification

As shown in the previous section and also described in Section 2.2.2, the influence of delay strongly depends on the game scenario and its underlying characteristics. The same applies to bitrate requirements regarding the video stream. Among other works, in [17] it was shown that the perceived video quality at the same encoding settings can vary strongly depending on the selected game scene, i.e., some scenes are more complex regarding their encoding than others. Finally, also the influence of the framerate, either due to a low encoding framerate or due to packet losses, depends on the game content as shown in [88]. These findings have severe consequences for planning and reporting subjective tests, but also imply that gaming QoE prediction models will most likely not reach a high accuracy if they do not incorporate content information. Hence, as also stated by Möller et al. in [190], a classifier of the degree of interactivity should be established and other relationships of game characteristics and gaming QoE should be further explored in order to derive a meaningful categorization of the games. Thus, in the remainder of this section, a first attempt to classify games with respect to their sensitivity towards delay, frame losses, and encoding complexity based on expert judgements will be described.

6.2.1 Method

The development of the classifications includes the following four steps:

- Creation of a dataset containing subjective ratings for various games with different characteristics played under conditions which influence the input quality and video quality
- Identification of possibly relevant game characteristics which may influence the impact of encoding bitrate, network delay, and frame losses on input quality or video quality
- Quantification of derived characteristics by experts
- Development of decision trees to predict the corresponding game class

As a dataset to derive the delay sensitivity classification, the dataset used for the development of the ITU-T Rec. G.1072 (cf. Section 5.5 and 7.1.2) was used in combination with 21 different game scenarios in a crowdgaming study following the methodology described in Section 4.1. For the latter, a total of 375 participants rated nine different open-source games used and modified with respect to their pace, number of objects to interact with, required accuracy, game rules when losing the game, as well as predictability of events. A full description of the used games can be found in [251]. As independent variable for the crowdgaming studies, a delay of 0 ms and 200 ms was used, which allowed the calculation of the DMOS between both conditions. The value of 200 ms was selected due to the findings presented in the previous section, to reduce the number of required conditions while still enabling a good separation of the game scenarios with respect to their delay sensitivity. To create the encoding complexity classification, the subjective data of 21 games gathered for the ITU-T Rec. G.1072 using the passive test paradigm (cf. Section 7.1.1) was used. Here, the DMOS of video quality was calculated for the reference condition using a bitrate of 50 Mbps compared to 1 Mbps (both at 1080p and 60 fps). The value of 1 Mbps was selected as a good separation of subjective ratings was shown in [88]. Finally, to obtain the frame loss sensitivity classification, the temporal video quality ratings, i.e., video discontinuity, of 16 games played or watched at a framerate of 60 fps and 20 fps (both at 1080p and 50 Mbps) from the complete ITU-T Rec. G.1072 dataset were used. The value of 20 fps, for which the DMOS was calculated compared to the 60 fps condition, was selected as results reported in [88] only showed a very low drop of temporal video quality at a framerate of 30 fps.

For the identification of the game characteristics, two separate focus group interviews were performed. To derive characteristics related to the delay sensitivity, the focus group methodology involved a total of nine participants divided into three groups. Based on participants' self-judgement, the groups included four highly, two medium, and three less-experienced gamers (cf. pre-test questionnaire described in Section 3.3.2). The group consisted of seven male and two female players with an age range of 19-27 years (M = 23 years). In the introductory of the focus group, the concept of cloud gaming, a live demonstration of the impact of delay, as well as the precision-deadline model by Claypool [247] as an example of a game characteristic were provided to the participants. Next, participants played 12 different game scenarios with 0ms, 150ms, and 300ms delay for 60 seconds each. After each scenario, participants reported their thoughts about the impact of the delay as well as about factors that may cause the varying delay sensitivities of the scenarios. In a second phase, all participants were given the liberty to openly discuss with the group about the identified game characteristics with the task to come up with a definition of those characteristics which are visually quantifiable by someone with reasonable gaming knowledge. For the identification of characteristics influencing the encoding complexity, also a two-step focus group interview with three experts was conducted. The procedure remained the same as

in the focus group described above, but also persons with a fair knowledge of video encoding were invited. Instead of showing different delay stimuli, participants were shown videos encoded with a bitrate of 1 Mbps to be able to perceive coding artifacts and thus, allowing them to judge which content characteristics could have cause them.

After the characteristics were identified, the descriptions were slightly modified during an additional expert interview, in which also examples and scale categories were discussed and applied. For the sensitivity towards frame losses, it was decided to use a mixture of the identified characteristics. This step allowed the quantification of derived characteristics, which was done in two additional studies involving 11 participants to quantify the characteristics potentially relevant for the encoding complexity, and 15 participants to quantify the characteristics potentially relevant for the delay and frame loss sensitivity. The participants were aged between 20 to 33 years (M = 25.7 years) and had an average gaming expertise of 3.5 according to the pre-test questionnaire described in Section 3.3.2. In both studies, participants were provided with representative video scenes for each game and were asked to rate the characteristics using a digital questionnaire. The full list of the identified characteristics as well as details about the studies, quantification method, consistency of quantification ratings, and data analysis can be found in [17], [18], [28].

In the following sections, the final remaining characteristics for the classifications as well as the decision trees and their performance will be described.

6.2.2 Encoding Complexity Classification

To derive the decision tree for the encoding complexity classification, the DMOS of the video quality was used for a K-means clustering. The analysis suggested an optimum of three clusters, referred to as low, medium, and high in the following. Next, a decision tree based on the mean values of the game characteristics derived from the quantification studies was created. The final decision tree is depicted in Fig. 6.2. The classification reached an overall accuracy of 96 % due to one miss-predictions for the low class (N = 21). The following five game characteristics are included in the decision tree and are thus regarded as important influencing factors on the encoding complexity of a game scene.

Movement Type: Movement type is defined as the total number of camera directions. When considering a video as a 2D representation, this characteristic refers to the directions in which the video is changing. The directions can be vertical or horizontal movements, as well as a mixture of movements (e.g., diagonal).

Length of Shapes: The characteristic Length of Shapes describes the summed length of contours (shapes) of moving objects averaged over the time of the game scene. Movements of objects or within objects as well as in the background or environment should be considered.

Degrees of Freedom: Degrees of Freedom (DoF) is defined as the freedom of camera movement. There can be up to six DoF due to three possible translations (back and forward, left and right, or up and down) as well as three rotations (vertical axis and height) of the camera.

Frequency of Object Movements: The amount of object movements, which also include elements that are not controlled by the player such as background objects, is defined as the percentage of the total duration of the time frame, in which game objects are moving, to the overall duration of the scene.

Texture Details: Texture details refer to the graphical details in the game and depend on the number of used quads (polygonal shapes made of triangles). The more polygons used, the higher are

the texture details. The game environment as well as other elements such as characters or obstacles should be considered.

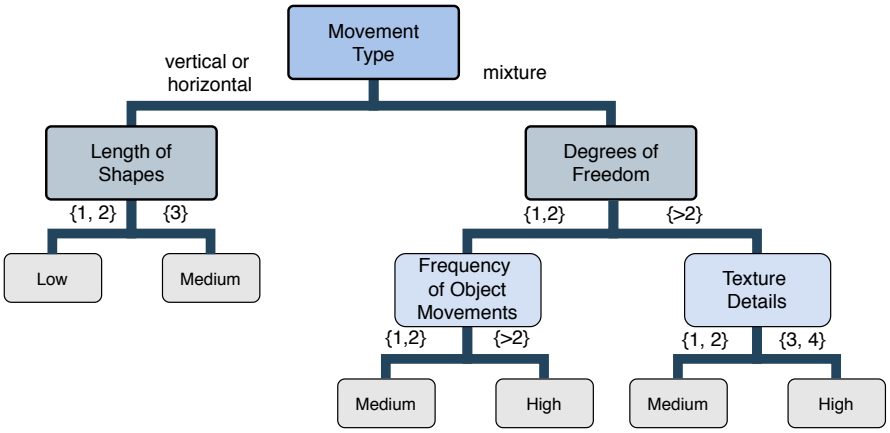


Figure 6.2: Decision tree determining the encoding complexity of a game scene (cf. [28]).

6.2.3 Delay Sensitivity Classification

To create the decision tree for the delay sensitivity classification, the DMOS of the input quality, which was calculated as the mean value of the GIPS items (cf. Chapter 5), was used for a K-means clustering. The analysis suggested the optimum number of clusters to be two (silhouette value = 0.77), referred to as low and high in the following. Game scenarios with a DMOS of approximately higher than 1.5 were grouped to the high delay sensitive class (cf. [18]). Afterward, to map the game characteristics to the clusters, a decision tree based on the mean values of game characteristic derived from the quantification studies was created. The final decision tree is illustrated in Fig. 6.3. The classification reached an overall accuracy of 86 % due to 3 and 1 miss-predictions for the low and high class, respectively (N = 30). The following four game characteristics are included in the decision tree and are thus regarded as important influencing factors on the delay sensitivity of a game scenario.

Type of Input: The type of input describes the temporal aspects of player inputs on a spectrum of discrete to continuous. In some games, players are continuously giving input, for example in a shooting game where players are always moving their mouse (or analog control sticks). Some games have discrete inputs meaning that players interact using pressing a button, for example, a jumping game where players must jump using pressing a key. In games with quasi–continuous inputs players interact with the game using holding a key or constantly pressing a key.

Number of Input Directions: The number of possible input directions, similar to the degree of freedom, consists of translations (back and forward, left and right, up and down) as well as rotations (vertical axis and height) for one or multiple input devices/elements.

Temporal Accuracy: Temporal accuracy describes the available time interval for a player to perform a desired interaction. The time interval is strongly dependent on the mechanics and pace of a game scenario. In other words, this game characteristic describes the available reaction time of a player.

Prediction Difficulty: Predictability describes if a player is able to estimate the upcoming events in the game. This can for example relate to positions of objects (spatial) or time points of events (temporal).

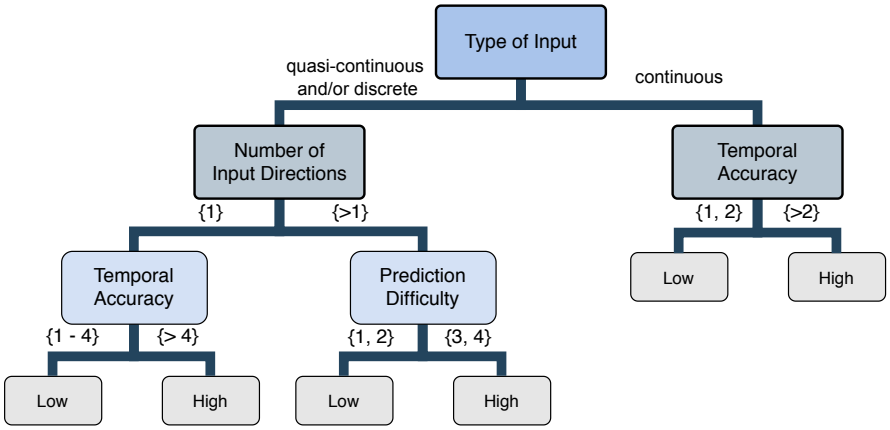


Figure 6.3: Decision tree determining the delay sensitivity of a game scenario (cf. [18]).

6.2.4 Frameloss Sensitivity Classification

To derive the decision tree for the frameloss sensitivity classification, the DMOS of the temporal input quality was used for a K-means clustering. The analysis suggested the optimum number of clusters to be two, referred to as low and high in the following. Next, to map the game characteristics to the clusters, a decision tree based on the mean values of the game characteristics derived from the quantification studies was created. The final decision tree is shown in Fig. 6.4.

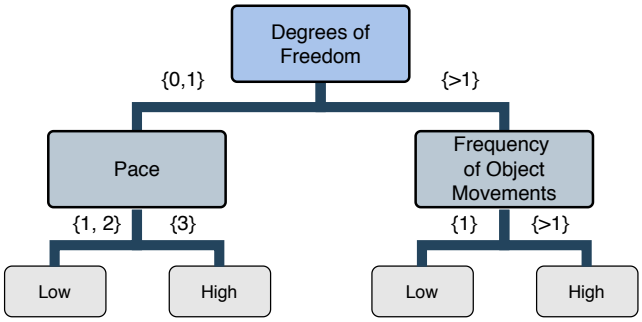


Figure 6.4: Decision tree determining the frameloss sensitivity of a game scenario (cf. [28]).

The classification reached an overall accuracy of 87 % due to 2 miss-predictions for the high class (N = 16). The decision tree comprises three characteristics that are regarded as important influencing factors on the frameloss sensitivity of a game scenario: the degrees of freedom, the frequency of object movements, and in addition the pace of the game scenario. The pace is defined as how fast the visible game elements (e.g., environment, characters, or obstacles) in the video are changed.

6.2.5 Discussion

As pointed out in the state of the art, a high impact of the game content for gaming research was shown in many studies. While single game characteristics were hypothesized and identified amongst other in [80], [193], [245], [252], a collective attempt to combine this knowledge based on empirical data was missing. However, the classification of games with respect to the impact of technical parameters is a highly challenging task due to the broad variability of games including their diverse rules and interaction possibilities. Even though the classification presented in this section is based on a high number of games and reached a very satisfying accuracy, there is still more data required for a validation. A limitation is also that the method still relies on expert judgments for the quantification of the characteristics. While

at least for the video encoding complexity, some success by using metrics such as Forward/backward or Intra-coded Macroblocks (PFIM) to describe motion in subsequent images, and Intra-coded Block Size (IBS) for spatial scene complexity [193], [253] combined with player actions intensity was achieved, an automated parametrisation of game characteristics with respect to interaction quality is still missing. Also, it must be considered that the findings of the study presented in Section 6.1 also apply to the classification of games. It cannot be concluded based on one scenario of a game, how each other scenario will be related to the impact of technical parameters such as delay on the gaming QoE. Thus, when applying the presented classification to a game as a whole, one must make sure to select a representative scenario to reduce a loss of accuracy, which is not possible for every game. However, the classification can be very helpful to explain research findings, to develop game adaptation strategies [211], and to develop opinion models like the one in ITU-T Rec. G.1072 (cf. Section 7.4).

6.3 Summary

In this chapter, research about the game scenario as an important quality factor was presented. At first, it was shown that for gaming research, not a game in its totality but rather the concrete game scenarios under test should be considered. It was shown that two different scenarios of the same game can differ in their sensitivity towards delay as much as completely different games themselves. In the presented study, the pace of the game scenario (or other related consequences) was very influential. Lastly, the difference between the weakly and strongly delay sensitive scenarios was most distinct at a delay of 200 ms.

This condition was also used for the game content classification presented in this chapter. The classification, on the contrary to the well-known genre classification, aims to group game scenarios based on their representative characteristics with respect to their delay sensitivity, frame loss sensitivity, and encoding complexity. A large list of potentially important characteristics that can explain different impacts of technical parameters such as delay or reduced bitrates on gaming QoE was created using the focus group method. Finally, the classification consists of decision trees using the results of a quantification method based on expert judgements. A high accuracy of the decision trees was shown for each content class. While the classification still relies on expert judgements to quantify the game characteristics, it promises many advantages for the planning of subjective tests and can furthermore be used for the development of gaming QoE models.

Chapter 7

Empirical Investigation of the Cloud Gaming Taxonomy

The aim of this chapter is two-fold: on the one hand, the reliability and validity of the measurement instrument used are to be examined, and on the other hand, the gaming taxonomy is to be empirically validated. The former allows judging whether the used instrument is accurate enough for the assessment of gaming QoE, or if it requires modification. The latter makes it possible to draw conclusions about the interrelationships of the quality aspects, as well as to assess the necessity of the individual elements of the taxonomy.

For the investigation of the measurement instrument and the empirical validation of the cloud gaming taxonomy, Structural Equation Modeling (SEM) will be used. SEM can be regarded as a CFA combined with multiple regression. While it is initially a confirmatory technique, it also can be used for exploratory purposes [236]. SEM, in comparison with CFA, extends the possibility of relationships among the latent variables. An SEM model consists of a measurement model, which links indicators to latent variables (cf. Section 5.4), and a structural model, which describes the relationship among the latent variables. Both of these models will be used in the following to reach the aim of the chapter.

To develop these models, a large database containing subjective ratings of all quality aspects, which were described in Section 3.2.3, was created. The dataset is comprised of over 500 participants and 30 video games. The test design, which follows ITU-T Rec. P.809 and was also explained in Section 3.2.3, was used to create them. As a result, the database includes data measured by a passive test paradigm as well as each data measured by an interactive test paradigm. However, since not all quality aspects are included in the first-mentioned, the following analyzes mostly consider the interactive database.

7.1 Cloud Gaming Datasets

7.1.1 Passive Viewing-and-listening Dataset

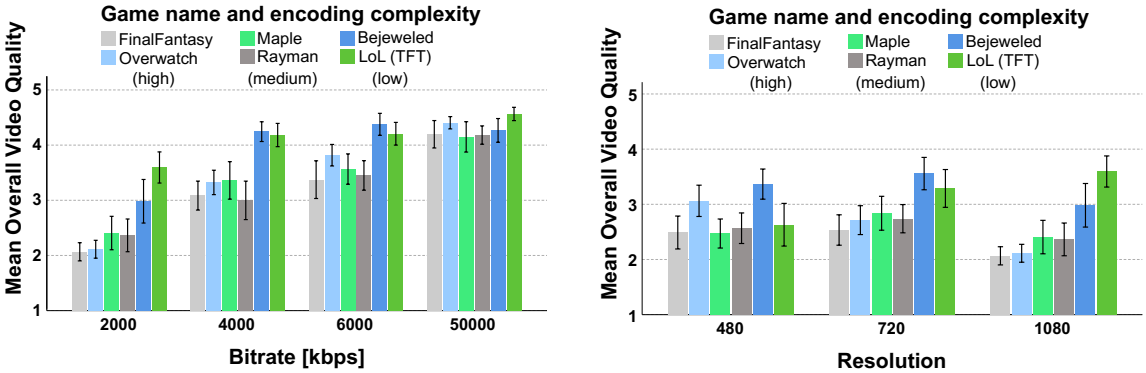
The passive dataset contains ratings of 266 participants who rated the overall video quality and the video quality dimensions described in Section 3.2.3 for a total of 21 game scenes encoded in various conditions. The dataset was created in seven study blocks, each containing three games. A large part of the dataset, including 15 games, was published in [254] based on studies conducted in 2019. In [254] also demographic information about the participants and more details about the encoding settings can be found. The remaining data, which also included some additional conditions examining packet

loss, was assessed in 2018 and also used for the development of the ITU-T Rec. G.1072. For brevity, these conditions are excluded from the following descriptions. As independent variables, the encoding parameters framerate, resolution, and bitrate were used in the combinations summarized in Table 7.1.

Table 7.1: Encoding conditions used for the passive G.1072 dataset.

Resolution	480p											
Framerate (fps)	60	60	60	60	30	30	30	30	-	-	-	-
Bitrate (Mbps)	0.3	1	2	50	0.3	1	2	50	-	-	-	-
Resolution	720p											
Framerate (fps)	60	60	60	60	30	30	30	30	20	20	20	20
Bitrate (Mbps)	1	2	4	50	1	2	4	50	0.3	1	2	50
Resolution	1080p											
Framerate (fps)	60	60	60	60	30	30	30	30	20	20	20	20
Bitrate (Mbps)	2	4	6	50	1	2	4	50	0.3	1	2	50

The values were chosen in such a way that they use the full range of the scale and contain some common values for comparison within the parameter resolution, i.e., 2 Mbps and 50 Mbps. In Fig. 7.1 the impact of the encoding bitrate (while keeping the resolution at 1080p and framerate at 60 fps constant) as well as the impact of the resolution (while keeping the bitrate at 2000 kbps and framerate at 60 fps constant) is shown. Please note that the ratings are converted to the typically used 5-point range for video quality according to [188]. It can be observed in Fig. 7.1a that the bitrate conditions span the range of the scale well and that strong differences exist among the games due to their encoding complexity variations. In Fig. 7.1b, it is illustrated that especially for encoding complex games, the highest resolution, i.e., in this case 1080p, does not result in the highest quality as the 720p condition was rated better. This shows the potential of encoding strategies based on available resources.



(a) Video Quality affected by encoding bitrate

(b) Video Quality affected by resolution

Figure 7.1: Bar plots of means and 95 % confidence interval showing the impact of encoding bitrate and resolution on video quality: a) bitrate conditions at 1080p and 60 fps, b) resolution conditions at 2000 kbps and 60 fps.

7.1.2 Interactive Dataset

The interactive dataset contains ratings of 180 participants who rated the gaming QoE and the quality aspects described in Section 3.3.2 for a total of 9 game scenarios under various network and encoding

conditions. The dataset was created in 9 study blocks, each containing one game. However, for one game, a second study using some additional conditions related to video encoding parameters were used to investigate the impact of spatial video quality on input quality (cf. Section 7.3). For brevity, these conditions are excluded from the following descriptions in this section.

In Table 7.2 the demographic information about the participants is summarized. The data shows that a 40 /60 gender balance was reached, which is at the acceptable level. The participants are between the age of 18 to 41 (Mdn = 27), mostly intermediate to experts based on the self-judgement of gaming expertise, and are mainly PC gamers. These statistics can be considered as representative of the target group of cloud gaming users.

Table 7.2: Demographic statistics of test participants in interactive G.1072 dataset.

Gender	female	male	transgender	others	
	73	107	0	0	
Age	18-25	26-30	30-35	35-41	
	32%	31%	35%	2%	
General gaming expertise (beginner – intermediate - expert) [%]					
15.6	11.1	41.7	23.9	7.8	
Experience with used game (beginner – intermediate - expert) [%]					
32.8	27.8	26.7	8.3	4.4	
Hours per week spent on playing video games					
0	0-1	1-5	5-10	10-20	>20
21.7	24.4	27.8	18.9	5.6	1.7
Likes playing (strongly disagree – undecided - strongly Agree) [%]					
1.7	2.2	15.6	52.2	28.3	
Device	PC	console	smartphone	others	
	52.2	24.4	22.8	0.6	

As independent variables, the following system parameters, resulting in a total of 17 conditions, were investigated:

- Round-trip time delay (*0*, 25, 50, 100, 200, 400 ms)
- Encoding framerate (10, 20, 30, *60* fps)
- Encoding bitrate (2, 4, 6, *50* Mbps)
- Frame loss rate caused by packet loss (*0*, 10, 20, 30, 40, 50 %)

It must be noted that only one parameter type was changed at a time (the values for the reference condition are italicized above). For the frame loss rate conditions, packet loss (uniform) combined with an additional network delay of 25 ms was simulated on the video bit-stream. The delay was used to avoid unrealistic error concealment which is only possible in a local network. Once more, the values were chosen in such a way that they use the full range of the scale and contain some common values with the bitrate-resolution pairs used in the passive tests, i.e., 2, 4, 6 Mbps, and 50 Mbps.

For a further analysis of the data, an initial data cleansing was performed with respect to missing data and unengaged answering of the questions. After the data cleansing, between 17 to 22 participants for each game remained. An analysis of the overall gaming QoE ratings for the reference condition revealed that 24 out of 266 participants rated the reference condition lower than 4, and nine lower than 3 on the 7-point scale. These ratings were not treated as outliers for the SEM analysis presented in Section 7.2 and 7.3, as they could still contain valuable information about the relationships between the assessed quality aspects. The quality aspects might explain the reason for the lower ratings of the

reference condition. However, they might not seem suitable for modeling the concrete impact of the network and encoding condition, and thus, were excluded for the ITU-T Rec. G.1072 development. The user factors assessed in the pre-test questionnaire were separately analyzed using a Mann-Whitney U test summarized in Table 7.3. The analysis revealed that the 24 unsatisfied users significantly less liked playing, significantly less spend time playing video games, and are significantly less experienced gamers than the other 242 participants. This finding can be interpreted as a confirmation that these user factors are important participant requirements and should be, if applicable, screened before inviting participants to subjective tests.

Table 7.3: Mann-Whitney U statistic for comparison of user satisfaction with the reference condition based on overall gaming QoE ratings.

User factor	Unsatisfied				Satisfied				Test statistics		
	N	M	SD	Mean Rank	N	M	SD	Mean Rank	U	z	p
Like Playing	24	3.54	0.88	100.27	242	3.95	0.87	136.80	2107	-2.44	.015
Video Game Time	24	1.58	2.54	89.94	242	4.35	5.07	137.82	1859	-3.01	.002
Gaming Expertise	24	2.21	1.02	89.27	242	2.95	1.12	137.89	1843	-3.12	.002

Not only among different user groups variations in the overall gaming QoE can be observed, but also between the various games tested. This is illustrated in Fig. 7.2 showing the ratings of the game-related quality aspects.

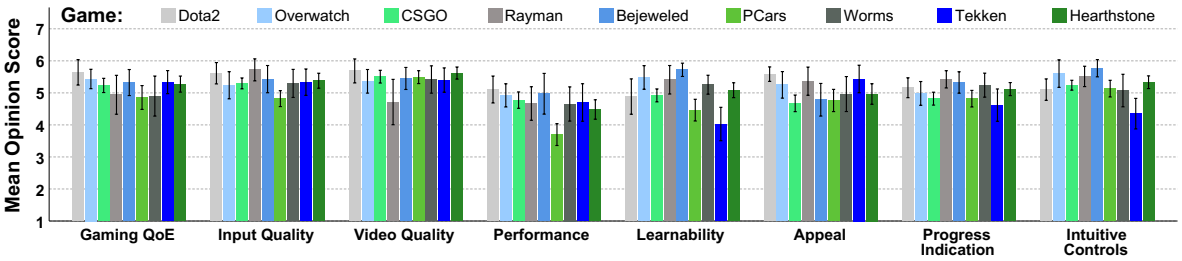


Figure 7.2: Bar plots of means and 95 % confidence interval of game-related quality aspects of the reference condition during the interactive tests.

It can be observed that the mean of the overall gaming QoE is lower for the games Rayman (M = 4.91, SD = 1.37), Worms (M = 4.9, SD = 1.22), and Project Cars (M = 4.86, SD = 1.21) compared to the other games, which also show a smaller standard deviation of in average 0.81. However, the differences are not statistically significant according to an independent one-way ANOVA, $F(8,257) = 1.79$, $p = .08$, $\eta^2 = .053$. The remaining dependent variables assessed can possibly explain these differences. For Rayman some participants noticed some discontinuities in the video (smoothness), and for Project Cars, as reported in Section 3.2, the inertia of the car was sometimes considered as a responsiveness issue leading to statistically significant reduced input quality, $F(8,257) = 2.60$, $p < .01$, $\eta^2 = .075$, playing performance, $F(8,257) = 5.30$, $p < .001$, $\eta^2 = .142$, positive affect, $F(8,257) = 3.24$, $p < .001$, $\eta^2 = .105$, as well as competence, $F(8,257) = 3.37$, $p < .01$, $\eta^2 = .087$. However, for Worms, apart from the slightly lower flow and immersion ratings (only trends, no statistical significance), no conspicuous features can be observed. To complete the report about the gathered ratings, the results of the PX aspects are visualized in Fig. 7.3. In general, it must be stated that the differences between games with respect to the PX aspects measured by the iGEQ as well as the post-game variables, i.e., appeal, progress indication, intuitive controls, and learnability, are rather low. Though, the latter

showed the highest variation.

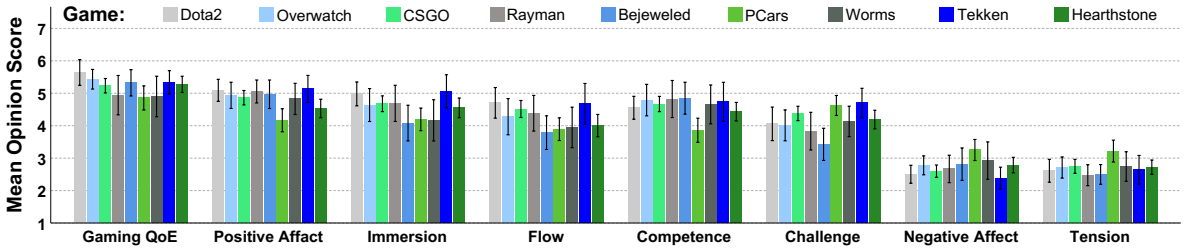


Figure 7.3: Bar plots of means and 95 % confidence interval of player experience aspects of the reference condition during the interactive tests.

In closing of this section, and to understand the variability of the assessed data of the dependent variables, the impact of the test conditions on the gaming QoE and its underlying aspects will be examined. In preparation, the outlier sampling method described in [220] was performed for each condition and dependent variable separately. In the case of more than two outliers on a dependent variable per condition (N = 84), the full sample of this condition was deleted. The same applies to all samples of participants who rated the reference condition lower than 4.0 (N = 228). Consequently, from the initial 2960 ratings, 2648 remained. This data was also used for the development of the ITU-T Rec. G.1072. While for brevity a detailed hypothesis analysis is not presented, for three selected games, Overwatch, Tekken 7, and Bejeweled 3, the gaming QoE of all 17 conditions is illustrated in Fig. 7.4. Based on the classification presented in Section 6.2 and Table D.1 in Appendix D, Bejeweled is considered to be low in terms of delay sensitivity, frameloss sensitivity, as well as encoding complexity. Overwatch is categorized as high in all three classes. Finally, Tekken is classified as low with respect to delay sensitivity, but as high regarding the frameloss sensitivity and encoding complexity.

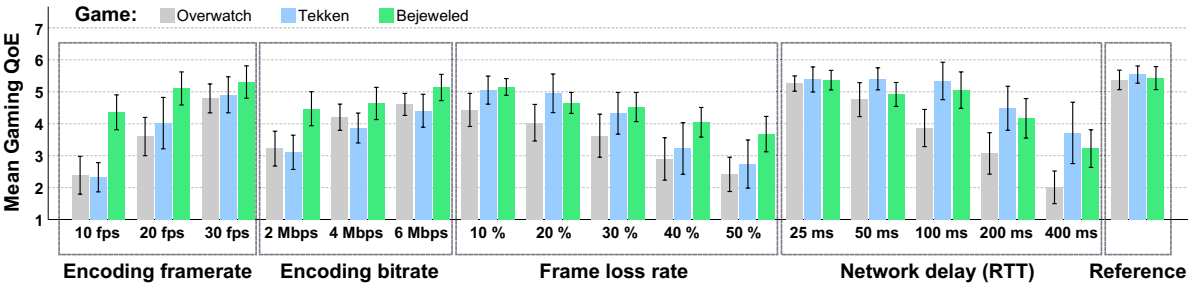


Figure 7.4: Bar plots of means and 95 % confidence interval of gaming QoE for all conditions of the interactive tests for three different games.

The subjective ratings reveal several interesting findings. First, it can be observed that the range of the scale is fully used and that all expected impacts of the parameters are visible, which is in line with the crowd gaming studies presented in Chapter 4. Next, it can be seen that Overwatch is strongly affected by all types of network and encoding conditions, whereas Tekken is not strongly impacted by a network delay. This is a reasonable finding as players frequently and repeatably pressed buttons to interact with the game (in the gamer scene this may be called "button smashing"), which resulted in a lower perception of the delay. This is not possible for Overwatch as it requires high temporal and spatial accuracy. The puzzle game Bejeweled, on the other hand, is not strongly impacted by any type of network and encoding parameter investigated in comparison to a sensitive game such as Overwatch. Consequently, these findings are in line with the classification of the three games. Lastly, the data

reveals that a variable loss of frames in case of frame losses due to packet loss compared to a constantly reduced encoding framerate, even though the average framerate is similar, has a much stronger negative impact on the gaming QoE (as the encoding framerate of the reference condition is 60 fps, the average framerate of the 30 fps and 50 % frame loss condition are similar).

Finally, in Fig. 7.5 the majority of assessed quality aspects are shown for four different conditions (one of each network and encoding factor manipulated) for the game Dota 2, which is highly delay sensitive and encoding complex, but only moderately sensitive towards frame losses due to its comparably low video scene motion. The data reveals that the positive PX aspects are high, and the negative PX aspects, i.e., tension and negative affect, are low for the reference condition, whereas the challenge rating is moderate. For a reduced encoding bitrate of 2 Mbps, the gaming QoE, (spatial) video quality, and PX aspects are negatively influenced whereas the input quality and video discontinuity are not. A very strong impact of the gaming QoE and input quality can be observed for the 200 ms network delay condition, which has no impact on the video quality aspects but even stronger consequences for the PX compared to the reduced bitrate condition. For the reduced encoding framerate to 10 fps and the frame loss rate condition of 40 %, which are rated similarly, all quality aspects are impacted by about 2 mean opinion scores. Overall, it can be observed that the challenge aspect was not strongly affected in any condition.

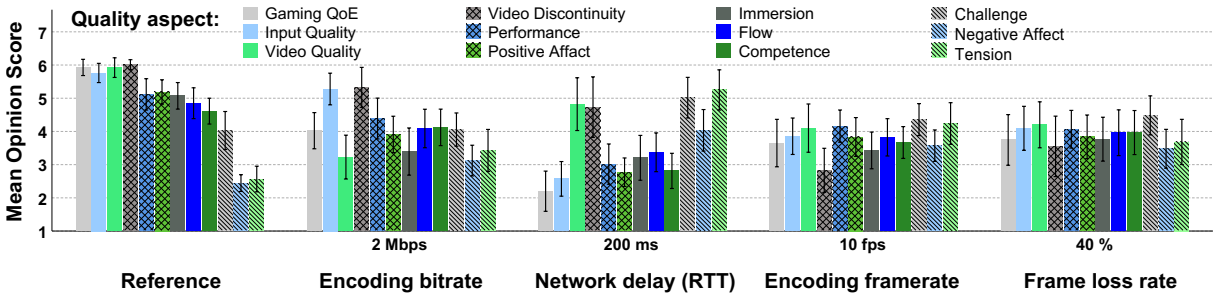


Figure 7.5: Bar plots of means and 95 % confidence interval of gaming QoE and underlying quality aspects for four exemplary network and encoding conditions for the game DotA2.

7.2 Measurement Model of Cloud Gaming QoE

In the previous section, it was shown that the assessed data fulfills the requirements of sufficient variability in the data and expected trends with respect to the impact of investigated network and encoding parameters on the full range of the rating scales. In this section, the measurement instruments, i.e., the questionnaires used in the subjective tests, should now be examined with regards to their reliability and validity. Therefore, a similar procedure as in the model development phase presented in Section 5.4 will be considered. Again, as part of a structural equation model, a measurement model will be created and analyzed. This step does not only investigate whether the measurement instruments are psychometrically sound, it is also a prerequisite for a structural model describing the relationships of quality aspects of the cloud gaming taxonomy in the next section. As the input quality and its underlying factors were already investigated in great detail in Chapter 5, the focus of the following analysis is put on the PX aspects measured by the iGEQ as well as the remaining game-related quality aspects, i.e., appeal and playing quality, measured in the post-game questionnaire.

Player Experience

As reported in Section 3.1.2, the GEQ was subject to some criticism with respect to its reliability and factor structure. However, in most cases, the validity and reliability of the GEQ were not investigated for the purpose of assessing the PX during cloud gaming *per se* nor was the iGEQ investigated. The fact that the iGEQ consists of only two items per factor limits a factor analysis to some extend. By an initial analysis, the seven PX factors and their associated items are investigated by means of a CFA in AMOS using the hypothesized factor structure, i.e., the validity of the indicator variables are tested, for the interactive dataset described in the previous section. However, neither an acceptable model fit nor discriminant validity could be confirmed. Even though the CFI of 0.965 (excellent) and SRMR of 0.091 (acceptable) were satisfying, the RMSEA of 0.083 did not fulfill the criteria for good model fit. In addition, the Fornell-Larcker-criterion was violated for two factors: a) the square root of the AVE for competence (0.885) was less than its correlation with positive affect (0.898), and b) the square root of the AVE for negative affect (0.817) was less than its correlation with tension (0.823). The latter confirms the structural issues identified by Law et al. in [154]. Thus, a Principal Axis Factor analysis, i.e., an EFA, was performed for all 14 items of the iGEQ. Thereby, it was observed that the best fitting factor structure varied among the tested games. For all games with exception of Rayman and Bejeweled, negative affect and tension did not form separate factors. The same applies to positive affect and competence with the exception of Dota 2. For the latter as well as for Worms and Hearthstone, also flow and immersion did not form own factors. Consequently, instead of the proposed 7-factor solution, mostly a 5- or even 4-factor solution revealed.

For the whole dataset, a 5-factor solution was revealed. The original factors positive affect and competence fall into a single factor referred to as Positive player experience (PPX) in the following, and the factors negative affect and tension emerged into a single factor called Negative player experience (NPX) in the remainder of the work. The latter confirms the findings reported in [157] suggesting a single negativity factor. Bartlett's test of sphericity was significant, $\chi^2(91) = 31644$, $p < .001$, and the KMO exceeded the minimum value of 0.50 and at 0.912 was regarded as great, which indicates adequately correlation among the variables. Also, there are only 3 % non-redundant residuals, which means that the factoring solution is determined substantially more by the variance and covariance of the included variables than by error. In Table 7.4 the pattern matrix and Cronbach's α values are reported. The table shows high loadings of each item on the corresponding factors, i.e., each loading is higher than 0.5 and the average of item loadings per factor is higher than 0.7, indicating good convergent validity. Additionally, no cross-loading differs less than 0.2 for any item, suggesting a good discriminant validity. It must be noted that the item CO2 represents a Heywood case, which should be investigated further in the upcoming analysis. Lastly, the factors show very good reliability as indicated by the Cronbach's α values close to .9, with the exception of the challenge factor. The analysis also revealed that removing NE1 would improve the reliability slightly to .88.

Next, a CFA was performed in AMOS in a similar way as presented in Chapter 5 for the global measurement model of input quality. The initial CFA revealed strong convergent validity issues for the challenge factor, as the loading of CH1 was at 0.32. For 728 ratings the difference between the subjective ratings of CH1 and CH2 was larger than 1.0, and for 344 ratings the difference was larger than 2.0. This issue occurred more frequently for the 200 ms and 400 ms delay conditions. As the number of items, i.e., only 2 items for the factor, leaves no possibilities of changes, it was decided to remove the factor challenge from the player experience model. Additionally, as indicated by the

Table 7.4: Pattern matrix and Cronbach's α of player experience factors (cross loadings lower .3 are omitted). Interpretation of factors: 1) immersion, 2) flow, 3) negative player experience, 4) positive player experience, 5) challenge.

Factor	1	2	3	4	5
Cronbach's α	.89	.89	.87	.92	.72
IM1	.90				
IM2	.72				
FL1		.86			
FL2		.85			
NE1			.76		
NE2			.94		
TE1			.72		
TE2			.73		
CO1				.83	
CO2				1.04	
PO1				.64	
PO2				.63	
CH1					.80
CH2					.77

Cronbach's α values discussed above, the item NE1 negatively impacted the model fit, RMSEA = 0.087, and thus, was also removed from the model. The revised measurement model of player experience, which is illustrated in Fig. 7.6, showed a very good model fit as summarized in Table 7.5. Whereas the CMIN/DF of 14.46 (DF = 37) is far greater than 3, this issue should not be taken too critical as the value strongly depends on the sample size [255], which was very high for the used dataset (N = 2916). However, the CFI of 0.982 and SRMR of 0.029 are excellent and the RMSEA of 0.068 acceptable.

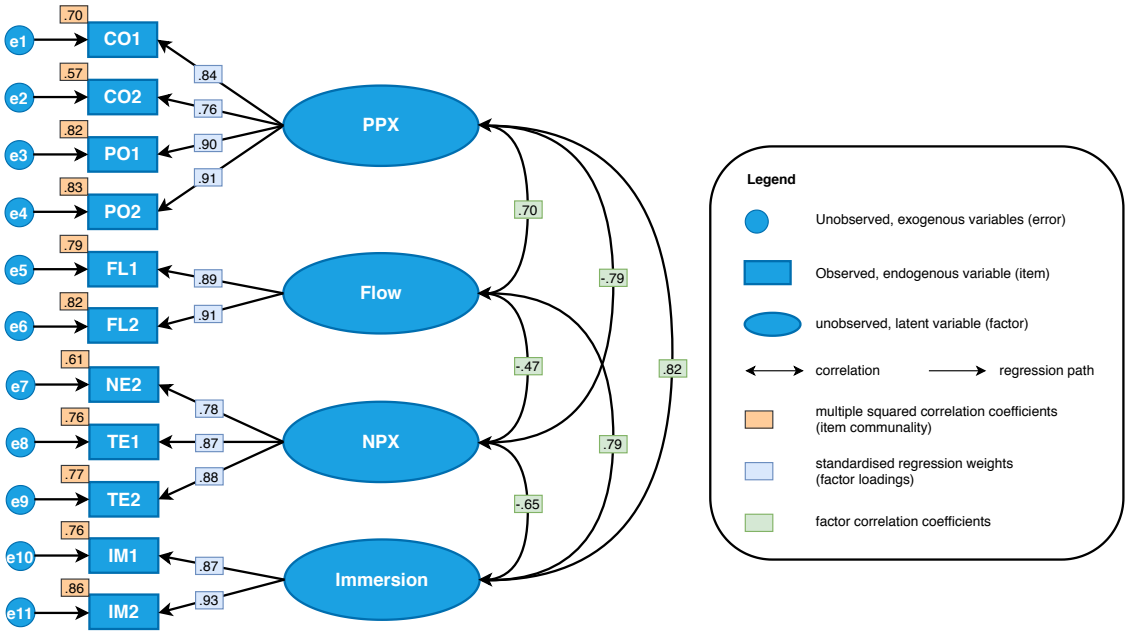


Figure 7.6: Measurement model of player experience.

The reliability and validity measures of the player experience model, which are presented in Table 7.5, also fulfilled the required criteria: CR > 0.6, AVE > 0.5, MaxR(H) > 0.7, and the Fornell-Larcker-criterion was met for every factor.

Table 7.5: Model reliability and validity measures of player experience model. The square root of AVE is shown on diagonal in bold faces.

					Inter-construct correlations			
	CR	AVE	MSV	MaxR(H)	PPX	NPX	Flow	Immersion
PPX	0.915	0.729	0.679	0.928	**0.854**			
NPX	0.881	0.713	0.625	0.890	-0.791	**0.844**		
Flow	0.893	0.806	0.626	0.894	0.698	-0.474	**0.898**	
Immersion	0.895	0.810	0.679	0.902	0.824	-0.653	0.791	**0.900**

Appeal and Playing Quality

In the following, the post-game variables, i.e., appeal as well as intuitive controls, learnability, and performance indication, will be analyzed with respect to their reliability and validity. Conceptually, the latter three belong in the cloud gaming taxonomy to the quality aspects called playing quality. It must be mentioned that appeal actually consists of items measuring aesthetics-related aspects, and does not include system personality nor novelty as considered in the cloud gaming taxonomy. This would be a subject of future work, in case of strong effects for the current model of appeal can be shown. In an attempt to find a well-fitting factor structure, all factors were used for an EFA. Bartlett's test of sphericity was significant, $\chi^2(28) = 15715$, $p < .001$, and the KMO exceeded the minimum value of 0.50 and at 0.812 was regarded as good, which indicates adequacy correlation among the variables. In all cases, intuitive controls and learnability were not indistinguishable and formed a single factor. Even though the performance indication factor showed a good fit presented in Section 5.5. based on the crowdgaming test dataset, it was problematic for the interactive G.1072 dataset. The suggested three-factor solution revealed average loadings of performance indication of 0.630, which is below the threshold of 0.7. Additionally, low Cronbach's α of 0.67 resulted for this factor and thus, no suitable solution was found when including performance indication. It might be that the very well developed games did not cause clear signs of issues as compared to crowdgaming dataset. Consequently, the factor was drop from further analysis, and the following two factors shown in Table 7.6 were derived.

The revised measurement model of the post-game factors, which is illustrated in Fig. 7.7, showed acceptable model fit as indicated by the following fit measures: CMIN/DF = 2.555 (DF = 16), CFI = 0.975, SRMR of 0.063, RMSEA of 0.093, and PClose = .024. The reliability and validity measures of the post-game factors model, which are presented in Table 7.7, also fulfilled the required criteria: CR > 0.6, AVE > 0.5, MaxR(H) > 0.7, and the Fornell-Larcker-criterion was meet for every factor.

Measurement Model of Cloud Gaming Taxonomy

As now all quality aspects of the cloud gaming taxonomy considered in the present work are investigated separately, in the next step, they will be combined in a global measurement model. The measurement model is illustrated in Fig. 7.8. It must be noted that for visualization reasons, player experience is

Table 7.6: Pattern matrix and Cronbach's α of player experience factors (cross-loadings lower .3 are omitted). Interpretation of factors: 1) playing quality, 2) appeal.

Factor	1	2
Cronbach's α	0.89	0.92
LE1	0.89	
LE2	0.74	
LE3	0.84	
IC1	0.85	
IC3	0.63	
AP1		0.90
AP2		0.93
AP3		0.83

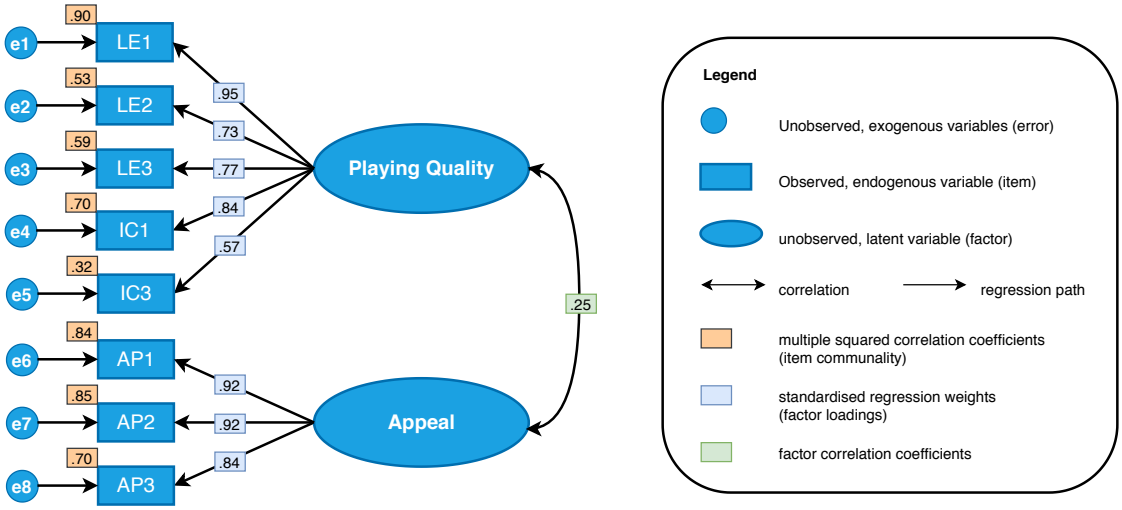


Figure 7.7: Measurement model of post-game factors.

Table 7.7: Model reliability and validity measures of post-game factors model. The square root of AVE is shown on diagonal in bold faces.

	CR	**AVE**	**MSV**	**MaxR(H)**	**Playing Quality**	**Appeal**
Playing Quality	0.885	0.611	0.06	0.937	**0.782**	
Appeal	0.921	0.796	0.06	0.929	0.245	**0.892**

used as a 2nd-order factor including the quality aspects immersion, flow, NPX, as well as PPX. The fit statistic CMIN/DF = 12.43 (DF = 298) was once more violated due to the very high sample size, and the RMSEA of 0.063 was acceptable. However, the overall model fit with respect to a CFI of 0.96, and SRMR of 0.037 can be regarded as excellent. The reliability and validity measures of the global measurement model, which are presented in Table 7.8, also fulfilled the required criteria: CR > 0.6, AVE > 0.5, MaxR(H) > 0.7, and the Fornell-Larcker-criterion was meet for every factor.

In order to investigate whether the model is biased due to the selected data, i.e., whether it is over-fitted, the dataset was split into a training (N = 1715) and test dataset (N = 1201). Next, configural invariance, metric invariance, and scalar invariance were tested.

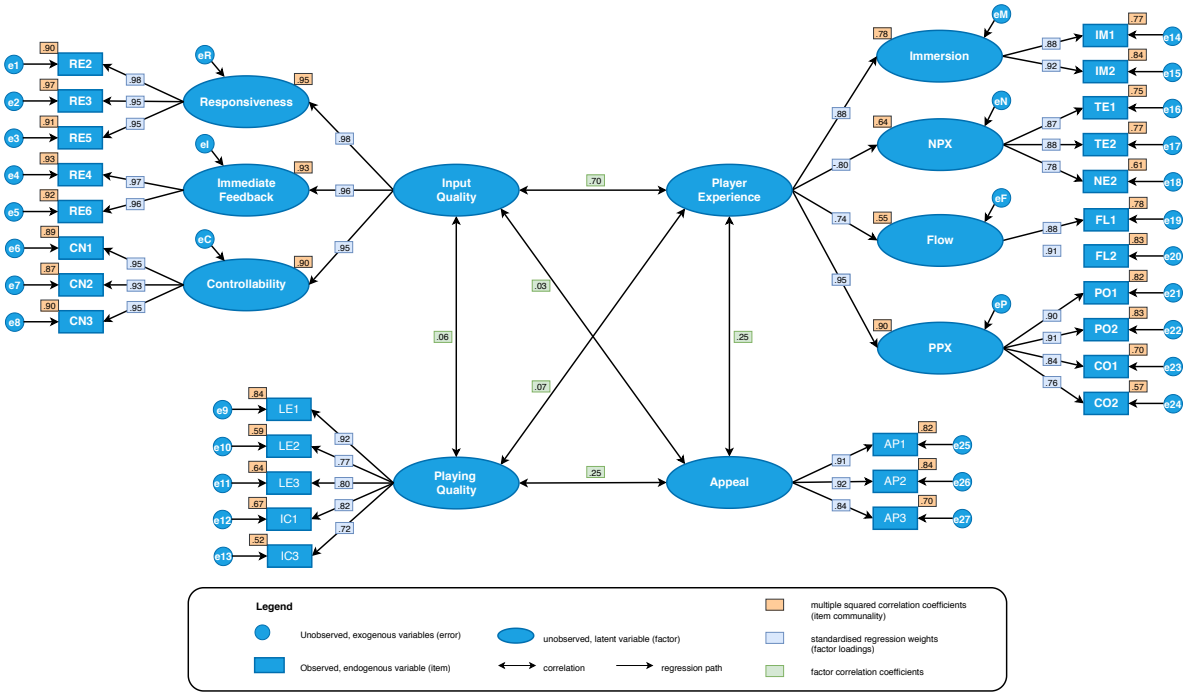


Figure 7.8: Global measurement model of cloud gaming QoE.

Table 7.8: Model reliability and validity measures of global cloud gaming QoE measurement model including immersion (IM), appeal, positive PX (PPX), negative PX (NPX), flow, playing quality (PQ), and input quality (IPQ). The square root of AVE is shown on diagonal in bold faces.

	CR	AVE	MSV	MaxR(H)	IM	Appeal	PPX	NPX	Flow	PQ	IPQ
Immersion	0.89	0.81	0.68	0.90	**0.90**						
Appeal	0.92	0.79	0.08	0.93	0.28	**0.89**					
PPX	0.92	0.73	0.68	0.93	0.83	0.23	**0.86**				
NPX	0.88	0.71	0.62	0.89	-0.65	-0.11	-0.79	**0.84**			
Flow	0.89	0.81	0.63	0.89	0.79	0.25	0.70	-0.47	**0.90**		
Playing Q.	0.90	0.63	0.06	0.91	-0.06	0.24	0.13	-0.15	-0.04	**0.80**	
Input Q.	0.98	0.93	0.49	0.98	0.58	0.03	0.68	-0.70	0.42	0.07	**0.96**

Configural invariance, also known as pattern invariance, describes whether the same items measure the targeted construct across different groups [256], in this case the two different datasets. Therefore, the overall model fit is calculated for both groups simultaneously and freely (i.e., without any cross-group path constraints) [257]. The analysis revealed again a very good fit indicated by CFI of 0.949, SRMR of 0.058, RMSEA of 0.049, and PClose of 0.795, providing evidence for configural invariance.

Metric invariance examines whether, in addition to the constructs being measured by the same items, the factor loadings of those items are equivalent across the groups. Attaining invariance of factor loadings suggests that the construct has the same meaning to participants across groups [256]. To judge this, the regression weights of the model are constrained to be equal across groups. Next, a chi-square difference test between the unconstrained and constrained models is conducted [257]. The analysis showed no significant difference, $\chi^2(20) = 22.43$, $p = .318$, providing evidence to suggest that the factor loadings are invariant across groups, and thus, allowing the use of the measured factors also in a structural model.

Scalar invariance describes whether mean comparisons across the groups can be justified, i.e., that the item intercepts and structural covariances are also equivalent across groups. For testing scalar invariance, the intercepts and covariances are constrained to be equal across groups, and the fit of the scalar model is compared with the fit of the metric model [256]. The analysis showed no significant differences for the measurement intercepts, $\chi^2(27) = 25.70$, $p = .535$, nor for the structural covariances, $\chi^2(42) = 41.56$, $p = .490$. Thus, scalar invariance can be assumed indicating that there is no measurement bias, i.e., there are no causes leading to changes in the way that participants are responding to items across the groups.

7.3 Structural Model of Cloud Gaming QoE

In this section, the RQ5 about the relevance of the individual quality aspects for the overall gaming QoE will be addressed. As it was shown in the previous section that the measurement model is operating adequately, one can have confidence in findings related to the assessment of a hypothesized structural model, which will be presented in this section.

Structural model

A structural model displays the interrelations among hypothesized latent constructs in the proposed model as a succession of structural equations similar to running several regression equations [236].

The hypothesized model can then be tested statistically in a simultaneous analysis for all variables to assess the degree to which they are consistent with the data. If the goodness-of-fit is appropriate, the model suggests the plausibility of postulated relations among variables. Whereas traditional multivariate procedures are incapable of either assessing or correcting for measurement error, SEM provides explicit estimates of these error variance parameters [258], and are correcting for the effects of measurement errors. If a hypothesized theory cannot be confirmed, the model can be revised using post-hoc modeling. However, the changes in the model should have a theoretical justification. To compensate for the risk of chance factors by deriving the model from particular sample data, a cross-validation strategy whereby the final model is tested on an independent test dataset should be performed [258]

Within the context of structural modeling, a variable that exerts an influence on other constructs is called an exogenous variable, typically referred to as an independent variable, and is not influenced by other factors in the quantitative model. On the other hand, there are so-called endogenous variables, typically referred to as dependent variables, which are affected by exogenous and other endogenous variables in the model [258]. While the cloud gaming taxonomy serves as the theoretical foundation for the present thesis, it does not represent a fully developed casual theory as it does not explicitly explain the relationships among the quality aspects nor which of the aspects have no relation to each other. However, the two-layer structure of the QoE aspects theorizes that the game-related aspects should be considered as exogenous variables, whereas the PX aspects and overall gaming QoE are the endogenous variables. Thus, as an initial theory, it was postulated that the game-related aspects are linked to all aspects of playing experience, which again is linked to the overall quality. This theory is similar to the work about the PXI presented in [97] which reports a relationship between functional consequences and game enjoyment mediated by psychosocial consequences.

Before starting to build the structural model, two requirements with respect to the used data must be investigated. First, the data should not violate multicollinearity, i.e., a strong correlation between two or more predictors in a regression model, which can be tested using the Variance inflation factor (VIF) test. According to Myers, a VIF value of greater than 10 is an indicator of multicollinearity concerns [259]. Fortunately, this was not the case for the used dataset as the lowest value of VIF is 1.18 for spatial video quality and 8.83 for positive PX. Second, the data should be free of multivariate influentials, i.e., outliers that exhibit an excessive influence over the estimates in a model. This requirement was tested using Cook's distance, which should not exceed a value of 1 [257]. For the used dataset, the Cook's distance had a maximum value of 0.03 and thus, not multivariate influentials were identified.

To create the structural model, first, the measurement model was used to calculate composite scores for each latent construct, i.e., the factor scores calculated as the weighted mean based on the items corresponding to each factor. This step was performed to reduce the complexity of reporting results and to ease the investigation of interaction effects. Next, also the SVQ, expressed as the mean of video fragmentation and video unclearness, the TVQ, and overall gaming QoE were included in the model. After the initial structural model, i.e., all game-related aspects connected to all PX aspects, was created and investigated, the following key findings were made: 1) playing quality did neither contribute to the gaming QoE directly nor served as a mediating variable in combination with other quality aspects (only a small effect on immersion was observed but no model fit improvement resulted), 2) in contrast to the other playing experience aspects, Flow did not contribute to the gaming QoE nor did an indirect effect together with other aspects improve the model, 3) the TVQ has a strong correlation with input quality, 4) the SVQ has only a very low impact on immersion but the path improved the model fit, and 5) appeal and TVQ have a small but statistically significant impact on gaming QoE. Consequently, the aspects flow and playing quality were discarded entirely in order to derive a concise model. Lastly, as expected, there is a strong correlation between the remaining PX aspects. A decision had to be made to either error terms of PX aspects covary or to propose a relationship among them. Thus, for the latter option, different paths were tested. NPX showed a weaker influence on gaming QoE than PPX and immersion, and the indirect effect of PPX over immersion to gaming QoE was very strong. Finally, the solution producing the highest R-squared, i.e., the proportion of variability explained, for gaming QoE was selected as this theory was considered to be sound.

The final structural model is shown in Fig. 7.9. The model showed an excellent fit as indicated by CMIN/df = 8.404, df = 6, CFI = .997, SRMR = 0.014, RMSEA = .050, and PClose = .45. All remaining paths are significant at the .001 level (two-tailed). It must be noted that the correlations of exogenous variables are omitted for better visibility or reduced complexity. Overall, the model reached an R-squared of .71, which can be regarded as great considering individual subjective ratings are considered and not MOS values of test conditions. If the same model is applied for the MOS values of the dataset, an R-squared of .94 is reached.

The structural model depicted in Fig. 7.9 can also be called a path model, i.e., a SEM using only single-indicator measurements as features that are assumed to be measured without error [260]. This assumption is typically wrong and can, in case of high measurement errors, bias the parameter estimates. However, due to the high reliability of the factors and the high sample size, also a full latent model, i.e., a structural model which calculates the latent variables directly based on the questionnaire items using the measurement model, showed an excellent model fit: CMIN/df = 8.574, df = 204, CFI = .980, SRMR = 0.028, RMSEA = .051, and PClose = .23. Furthermore, no major changes in the regression

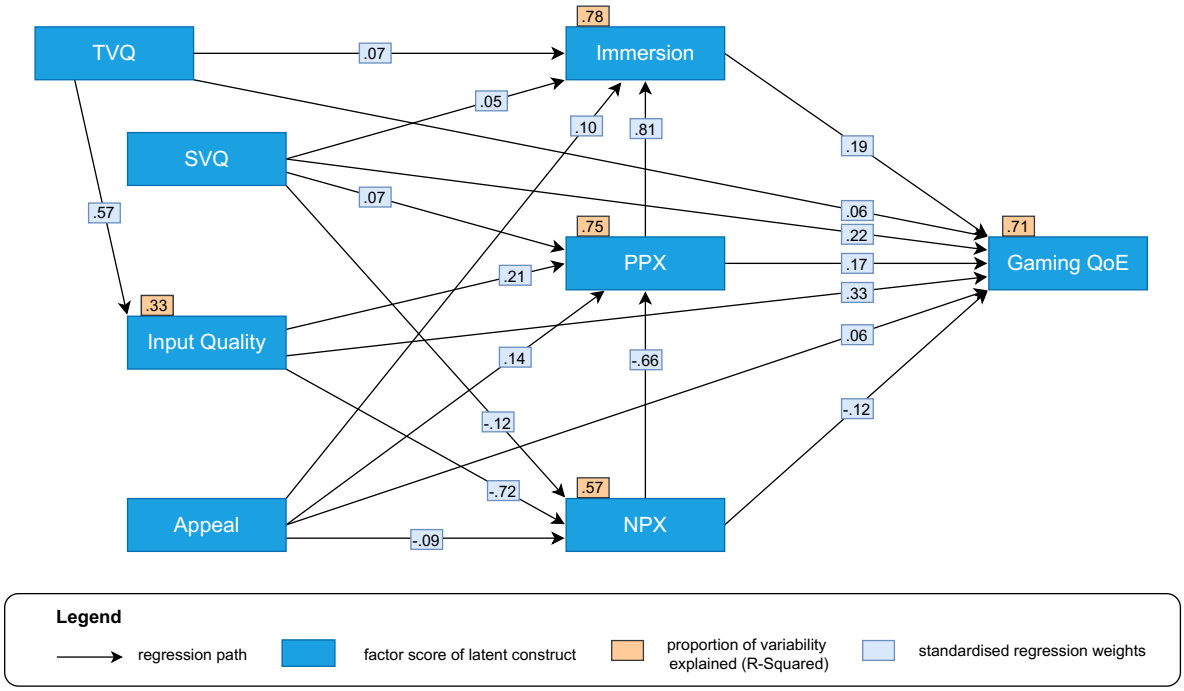


Figure 7.9: Structural model of cloud gaming QoE using standardized regression weights.

weights were observed and the resulting R-squared of 0.72 for gaming QoE was very similar to the solution presented in Fig. 7.9. Therefore, also a full latent casual model is supported. To be very precise, since for the video quality aspects only single-predictor variables (also called unique factors) are available, the model should be called a partially latent structural regression model [260].

As a post hoc model fitting was performed, cross-validation should be employed as suggested by Byrne [258]. The model derived based on the training dataset will be tested on the test dataset which represents independent samples from the same population. Using the full latent model, a Chi-square difference tests between the configural model, i.e., all parameters are estimated for the training and test groups simultaneously without constraining parameters to be equal across groups, and the constrained measurement weights model, showed no significant difference, $\Delta\chi^2\ (13) = 13.15$, $p = .436$.

Also constraining the regression weights of the structural model (scalar invariance) to be equivalent in the two groups showed no significant difference between the groups, $\Delta\chi^2\ (6) = 9.20$, $p = .163$. Lastly, the ΔCFI never exceeded a value of .001., which is below the .01 cutoff point proposed by Cheung and Rensvold [258]. Consequently, the postulated SEM can be regarded as equivalent across the training and test data.

Regarding the impact of the quality aspects on the overall gaming QoE, it can be summarized that the game-related aspects exert a stronger influence compared to the PX aspects. Indeed, in line with the PXI model, the PX aspects mediate the relationship between game-related aspects and gaming QoE. Amongst them, the input quality had the strongest direct and overall influence on gaming QoE. In Table 7.9 all indirect effects, i.e., mediation effects, are summarized (but limited to three factors). To be highlighted are that PPX mediates the positive relationship between input quality and immersion, that input quality mediates the positive relationship between TVQ and QoE, that immersion mediates the influence of PPX on QoE, and that the positive relation between appeal and immersion is mediated by PPX. While the indirect effects of SVQ and appeal are rather small, it must be noted that multi-path effects usually are not very high as two variables are multiplied with each other.

Table 7.9: Indirect Standardized effect estimates (SE) for three-paths of gaming QoE (QoE), input quality (IPQ), spatial video quality (SVQ), temporal video quality (TVQ), immersion (IM), postitive player experience (PPX), negative player experience (NPX), and appeal (AP).

Indirect path	p-value	SE	Indirect path	p-value	SE
TVQ » IPQ » NPX	< .01	-0.42	SVQ » NPX » PPX	< .01	0.08
TVQ » IPQ » PPX	< .01	0.12	SVQ » PPX » IM	< .01	0.06
TVQ » IPQ » QoE	< .01	0.19	SVQ » NPX » QoE	< .001	0.01
TVQ » IM » QoE	< .001	0.01	SVQ » PPX » QoE	< .001	0.01
IPQ » NPX » PPX	< .001	0.48	SVQ » IM » QoE	< .001	0.01
IPQ » PPX » IM	< .01	0.17	AP » NPX » PPX	< .01	0.06
IPQ » NPX » QoE	< .001	0.09	AP » PPX » IM	< .01	0.12
IPQ » PPX » QoE	< .001	0.04	AP » NPX » QoE	< .001	0.01
NPX » PPX » IM	< .01	-0.54	AP » PPX » QoE	< .001	0.03
NPX » PPX » QoE	< .001	-0.11	AP » IM » QoE	< .001	0.02
PPX » IM » QoE	< .001	0.16			

Next, it should also be discussed which quality aspects are not presented in the structural model, namely playing quality, flow, and also the self-judged playing performance. Regarding the latter, it was found that the effect of TVQ on performance is mediated by input quality. Additionally, performance is a statistically significant ($p < .001$) mediator of the relationships between input quality and NPX and PPX resulting in standardized effect estimates of -.14, and .16, respectively. However, including performance to the model did no show benefits with respect to the explained variability in the gaming QoE, as there was no direct effect ($p = .74$) of performance and gaming QoE, nor by any of the indirect paths such as input quality over performance to gaming QoE ($p = .77$). Consequently, performance was not added to the model and must not be added to the taxonomy. A reason for this finding could be that many players can still perform well even under bad network and encoding conditions, but rate the gaming QoE low. Concerning playing quality and flow, there are two limitations that should be considered: 1) only well-designed games were selected for the test, as the purpose was not the evaluate the games but the cloud gaming service, and 2) the duration of the stimuli was rather short, which might not be enough for players to reach a state of flow. Thus, more research would be required to make final conclusions about their importance.

Lastly, it must be added that game scenario also has a significant impact on the relationships between the quality factors. For a concise overview, the standardized direct effect estimates for each game are summarized in Table 7.10. It can be observed that for some games, direct effects become not-significant. While the input quality is rather invariant with respect to the significance (but not effect size), the TVQ and SVQ are less dominant for several games. This is in line with the findings presented in the previous sections and literature. The analysis also shows that the importance of PX aspects such as flow and immersion strongly depends on the games used. Finally, the R-squared for the specific game models showed a maximum of .87 for Dota2, compared to .71 for the overall model. If applicable, a game-specific model would thus increase also the prediction accuracy of models.

Table 7.10: Direct standardized effect estimates (SE) summarized per column for each game including gaming QoE (QoE), input quality (IPQ), spatial video quality (SVQ), temporal video quality (TVQ), immersion (IM), postitive player experience (PPX), and negative player experience (NPX). Non-significant effects are highlighted in bold.

Direct effect		TVQ IPQ	IPQ NPX	IPQ PPX	IPQ QoE	SVQ NPX	SVQ PPX	SVQ IM	SVQ QoE	TVQ IM	TVQ QoE	NPX QoE	PPX QoE	IM QoE
Game 1	SE	.65	-.84	.17	.53	-.06	.10	.09	.12	.05	**-.01**	-.18	.25	**-.01**
	p	<.001	<.001	<.001	<.001	.034	<.001	<.001	<.001	.026	**.614**	<.001	<.001	**.863**
Game 2	SE	.58	-.73	.18	.25	-.13	.09	.13	.19	.13	.14	**-.07**	.18	.31
	p	<.001	<.001	<.001	<.001	<.001	.002	<.001	<.001	<.001	<.001	**.199**	.012	<.001
Game 3	SE	.70	-.79	.21	.46	**-.02**	**.00**	**-.04**	.14	.10	**-.09**	-.10	.25	.12
	p	<.001	<.001	<.001	<.001	**.541**	**.947**	**.310**	<.001	.025	**.076**	.117	<.001	.022
Game 4	SE	.65	-.76	**.07**	.32	-.18	**-.02**	**-.02**	.18	**.04**	.17	**.03**	**.12**	.38
	p	<.001	<.001	**.151**	<.001	<.001	**.618**	**.401**	<.001	**.229**	<.001	**.635**	**.141**	<.001
Game 5	SE	.45	-.78	.28	.31	-.17	.15	.05	.13	**.02**	.09	-.19	.25	**.14**
	p	<.001	<.001	<.001	<.001	<.001	<.001	.021	<.001	**.464**	.001	.002	.011	**.060**
Game 6	SE	.60	-.64	.34	.24	-.10	**.05**	-.16	**.05**	**.03**	**.05**	**-.03**	.18	.35
	p	<.001	<.001	<.001	<.001	.017	**.168**	<.001	**.338**	**.383**	**.354**	**.661**	.048	<.001
Game 7	SE	.49	-.68	.33	.38	**.01**	**.04**	**-.01**	.09	.10	**.08**	-.21	.41	-.21
	p	<.001	<.001	<.001	<.001	**.790**	**.285**	**.686**	.049	.008	**.151**	.008	<.001	.029
Game 8	SE	.56	-.57	.22	.30	-.21	.13	.13	.37	.08	**-.02**	-.18	**.12**	.14
	p	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	**.443**	<.001	**.060**	.010
Game 9	SE	.48	-.59	.26	.67	**.02**	**-.07**	**.01**	.09	**.06**	**.04**	**-.04**	**-.01**	.21
	p	<.001	<.001	<.001	<.001	**.765**	**.116**	**.755**	.026	**.082**	**.314**	**.489**	**.928**	.007

7.4 Opinion Model predicting Gaming QoE

An interesting finding of the model is that the game-related aspects seem to be very dominant. Thus, a strongly reduced model was created which only consists of TVQ, SVQ, and input quality. The model is shown in Fig. 7.10. It can be observed that the regression weights did not change substantially and that despite the absence of the PX aspects, 62 percent of the variability in the gaming QoE is explained. If using the MOS values of the test conditions, even an R-squared of .92 was reached. This model was largely followed by the ITU-T Rec. G.1072, which will be shortly described in the following.

The ITU-T Rec. G.1072 contains an opinion model predicting the gaming QoE of cloud gaming services. The recommendation follows the scope of the present thesis with respect to stakeholder perspectives and systems considered. It uses network and encoding parameters, as well as a classification of game content described in Chapter 6, to predict gaming QoE using various impairment factors. Thus, as a parametric model, no subjective ratings are required for the prediction, but only the system factors encoding bitrate, encoding framerate, resolution, network delay, and packet loss rate. As a dataset, the passive and interactive datasets presented in Section 7.1 was used. The impairment factors are derived from subjective ratings of the corresponding quality aspects, e.g., spatial video quality, and modeled by non-linear curve fitting. For the prediction of the overall score, a linear regression is used. To create the impairment factors and regression, a data transformation from the MOS values of each test condition to the R-scale was performed, similar to the well-known E-model [261]. The R-scale, which results from an s-shaped conversion of the MOS scale, promises benefits regarding the additivity of the impairments and compensation for the fact that participants tend to avoid using the extremes of rating scales. To calculate the predicted MOS, the data conversion using the R-scale must be reversed as explained in [203]. The highest assessed MOS value of gaming QoE was 4.64.

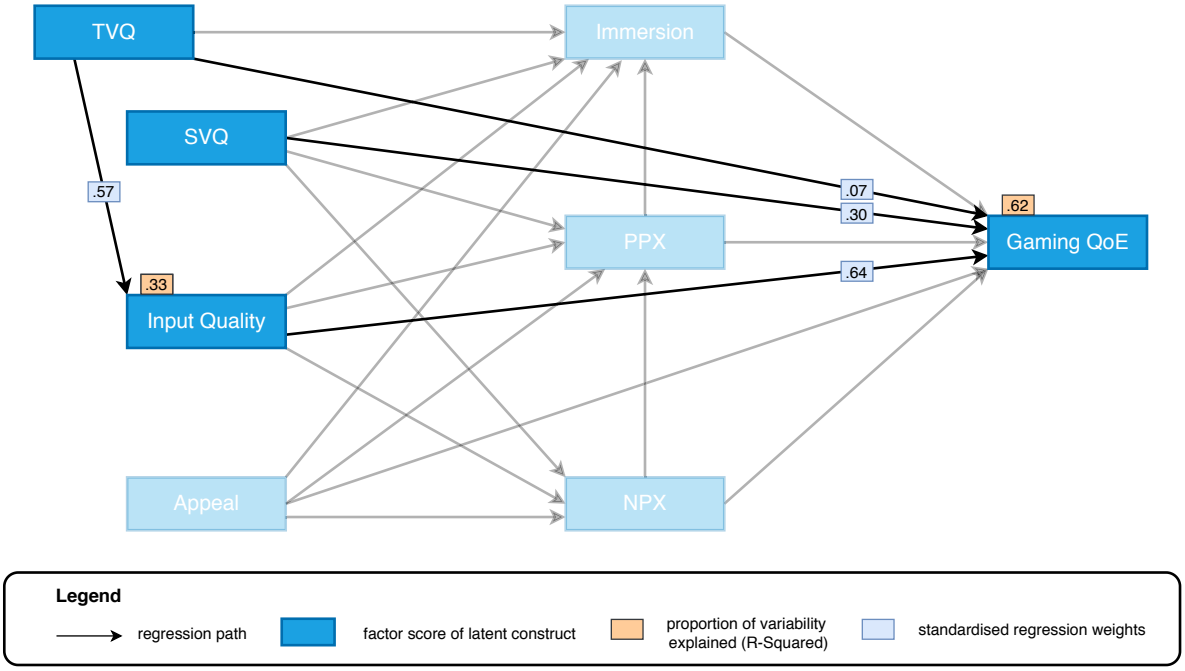


Figure 7.10: Structural model of cloud gaming QoE similar to core model of ITU-T Rec. G.1072 using standardized regression weights.

The core model and the included impairment factors are defined as follows:

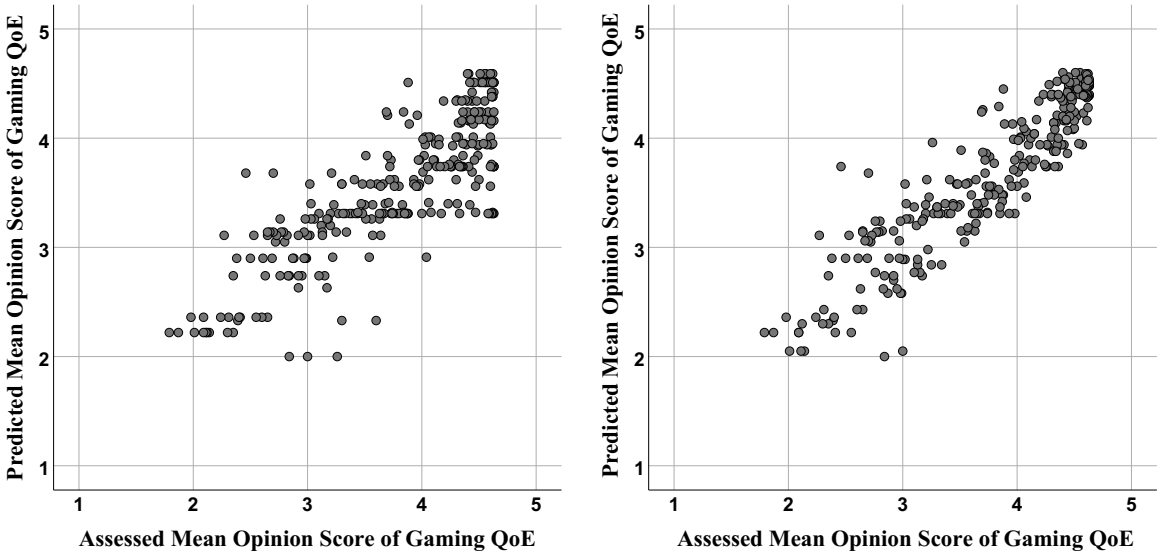
$$R_{QoE} = R_{max} - 0.79 \cdot I_{VQ_{cod}} - 0.90 \cdot I_{VQ_{trans}} - 0.23 \cdot I_{TVQ} - 0.63 \cdot I_{IPQ_{frames}} - 0.85 \cdot I_{IPQ_{delay}} \quad (7.1)$$

where, R_{QoE} is the overall estimated gaming QoE ranging between 0 (worst) and 100 (best);
R_{max} is the reference value indicating the best possible gaming QoE (i.e., a value of 100);
$I_{VQ_{trans}}$ is the estimated spatial video quality impairment for video compression artifacts;
$I_{VQ_{cod}}$ is the estimated spatial video quality impairment for video transmission errors;
I_{TVQ} is the estimated temporal video quality impairment for frame rate reductions;
$I_{IPQ_{frames}}$ is the input quality impairment for frame rate reductions; and
$I_{IPQ_{delay}}$ is the input quality impairment for network delay degradations [203].

In line with the SEM presented in Fig. 7.10, the input quality and spatial video quality have the strongest impact on the overall gaming QoE. For more details about the development process, influence of the specific parameters on the impairments, modeling approach, and system configurations, the reader is referred to the recommendation text presented in [203] as well as to the corresponding ITU-T contributions [24], [25], [26], [27], [29], and [30].

As a concluding remark about the G.1072, its performance to predict the gaming QoE on the test dataset will be discussed. As means to evaluate the performance, the predicted MOS of gaming QoE is compared to the subjective (assessed) MOS using the RMSE and PLCC summarized in Table 7.11 as well as the scatter plot visualized in Fig. 7.11. The performance is reported separately for using the game content classification presented in Chapter 6 and without it.

While some predictions in the MOS-range of 3 are particularly scattered, overall it can be observed that the model performs very well. Furthermore, it can be observed that the model using the game classification shows a noticeably higher accuracy (R-squared of 0.81) than the model not using the



(a) Model without classification.

(b) Model with classification.

Figure 7.11: Scatter plot of the predicted and assessed MOS ratings on the test dataset based on Figure G.1072(20)_F02 in ITU-T Rec. G.1072 [203].

Table 7.11: Performance evaluation of the ITU-T Rec. G.1072 model.

	Considering classification		Without considering classification	
	R-scale	MOS-scale	R-scale	MOS-scale
RMSE	8.03	0.33	12.19	0.47
PLCC	0.89	0.9	0.8	0.82

classification (R-squared of 0.67). This once more illustrates the importance of the game as an influencing factor. While the variability of the gaming QoE data explained is much lower than the 92 percent using the SEM model shown in Fig. 7.10, these results are still very satisfying as for the G.1072 model the quality is predicted only based on network and encoding parameters and not directly based on subjective ratings of quality aspects.

7.5 Discussion

With respect to the interactive dataset, an analysis of the collected data showed that the selected parameters caused the subjective ratings to span the full range of the used scales. This is an important finding regarding the suitability of the data for the development of quality prediction models. The impact of the independent variables on the quality aspects fulfilled the expectations of reports in the literature (cf. Section 2.2.2) and also those shown in the crowdgaming studies (cf. Section 4.2.4). Based on the results of PX aspects shown for the game Dota2, one can conclude that participants reported a higher challenge only for the delay conditions, but not for the other degradation types. Additionally, the impact of the network and encoding parameters on the input quality and video quality was dependent on the evaluated game scenario, which can be explained by the classification presented in the previous chapter. Interestingly, according to Fig. 7.4, a very low framerate of 10 fps, which corresponds to distance between the frames of 100 ms, has a higher negative impact on gaming QoE than a delay of

100 ms or even 200 ms. However, for the game Dota2, which was impacted by a delay the most among all games, an exception to this finding was revealed. In Dota2, a delayed movement of the mouse causes a strongly visual delay of the whole camera scene movement. It appears that participants generally criticize visual issues that affect their interaction with the cloud gaming system very strongly. The importance of taking such game information into account was also proven in the scope of the ITU-T Rec. G.1072 results presented. Using the proposed game content classification improved the prediction performance significantly.

Regarding the developed SEM of gaming QoE, it can be concluded that, generally, the PX aspects mediate the relationships between the game-related aspects and the overall gaming QoE. This finding is similar to the research results presented by Abeele et al. in [97]. A mediating variable can be considered as a reason for an effect. For the concrete example of the SEM presented in this chapter, a high input quality leads to a high gaming QoE. The reason for this effect is because a high input quality leads to high immersion, which leads to a high gaming QoE. Thus, immersion is one reason that input quality leads to gaming QoE. However, in sum, it was shown that the game-related aspects are equally or even more dominant for the judgement of the overall gaming QoE. By using only game-related aspects one can explain 62 percent of the variability in the gaming QoE on the subjective level and 92 percent on the MOS level based on the collected data. This performance is even higher than a model that would only consider the PX aspects (R-squared of .60 on the subjective level, and .84 on the MOS level). These results indicate that the cloud gaming taxonomy indeed covers the majority of important quality aspects which players consider during their judgement process.

In terms of empirical validity of the cloud gaming taxonomy, the collected data confirms to a large extent the theory of the cloud gaming taxonomy. However, it was shown that not all components of the taxonomy contributed significantly in their relationships with the gaming QoE. Thus, an updated version of the QoE layer of the cloud gaming taxonomy is presented in Figure 7.12. The modifications are performed based on the SEM analysis presented in this thesis. Quality features and aspects highlighted in black font were confirmed by the SEM analysis. Others highlighted in gray could not be confirmed based on the findings presented in this thesis. Lastly, the features in brackets were not investigated (system personality, interactive behavior, involvement, and presence) directly.

The conceptually most visible change in the taxonomy is the rearrangement of the PX aspects. In particular, tension and negative affect are considered as one concept called negative player experience, which is in line with the findings presented in Section 3.1.2. Furthermore, the quality feature competence forms together with positive affect the concept of positive player experience. While in the previous version of the cloud gaming taxonomy competence was considered to be related with flow, in the updated version a balanced challenge of the game scenario is considered to influence flow. However, the low variance in the challenge aspect results in the exclusion of this features from the measurement model presented above, and also the aspects flow and playing quality did not show a significant impact. The fact that flow and challenge, which by theory are also related to each other, did not show a significant impact on gaming QoE might be caused by the rather short stimulus duration. Additionally, while it was assumed that also the network degradation would strongly increase the perceived challenge, it could be that the selected game scenarios were not challenging enough for many participants. Also, it would be interesting to confirm this finding in a more ecologically valid environment in which participants are not clearly part of a research study but rather play in their natural home environment. Furthermore, questionnaires assessing more facets of the flow concept should be considered for future investigations.

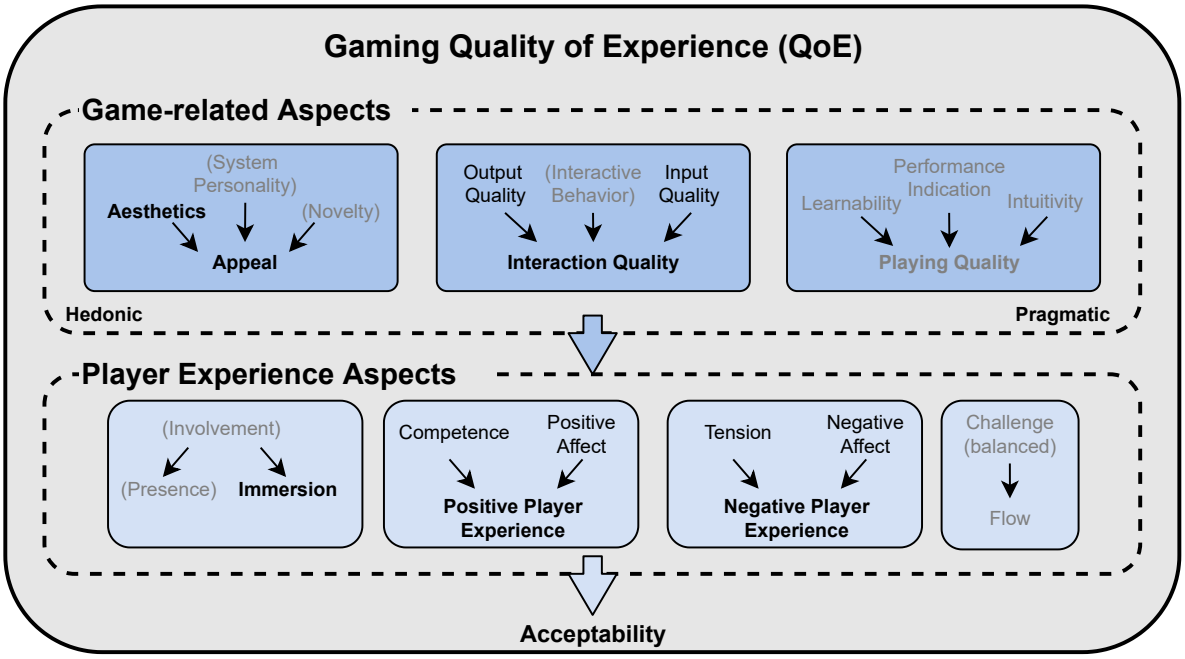


Figure 7.12: Updated version of the cloud gaming taxonomy based on the SEM analysis with respect to quality features and aspects.

The low impact of the aspects appeal and playing quality could be caused by the selection of the games used for the dataset. If games with severe design issues would be used, these aspects may still represent valuable variables to understand the judgement process of players. It was also shown that the importance of quality aspects could be game-dependent (cf. Table 7.10). Thus, even though it would require a high effort, cloud gaming providers may consider game-specific quality models.

7.6 Summary

In this chapter, at first, a large database containing subjective ratings for a total of 2648 stimuli of various network and encoding conditions was presented. The dataset was created based on the test design and measurement tools presented in Section 3.3.2 and was also used for the development of the ITU-T Rec. G.1072. While for the latter, also the data derived by the passive test paradigm was used, the focus of Chapter 7 is on the interactive dataset which covers the majority of quality aspects of the cloud gaming taxonomy. The impact of the network and encoding parameters on the quality aspects (cf. Section 2.2.2), as well as differences between the game scenarios (cf. Section 6.1) was in line with expectations, although with the exception of a very low variance in the challenge aspect. The differences between the games with respect to the game-related quality aspects appeal and playing quality were rather low.

Next, a measurement model for the PX was presented. For the seven factors of the iGEQ, an EFA suggested a five-factor solution. The aspects negative affect and tension merged into one factor named negative PX, whereas positive affect and competency merged into one factor named positive PX. This is mostly in line with the finding from Law et al. in [154]. A follow-up CFA revealed that the aspect challenge leads to strong convergent validity issues, and was thus discarded from the model. Regarding the playing quality, the aspects learnability and intuitive control resulted in a one-factor solution that was distinct from the appeal aspect. The overall measurement model reached an excellent

fit, suggesting that the used measurement tools are reasonably consistent with the data and do not require respecification. Furthermore, good discriminant validity shows that the assessed factors are distinct from each other. These findings are essential for testing any causal assumptions, i.e., they allow having confidence in the findings generated from a structural model [257].

Using the cloud gaming taxonomy as a theoretical foundation, the interrelations among the quality aspects were investigated. The structural model fulfilled the requirements of multicollinearity, absence of multivariate influentials, and was shown to be independent of the used sample data demonstrated by a split into test and training dataset. The analysis revealed that many aspects of the taxonomy are strongly related to each other but that some aspects do not significantly contribute in explaining variability in the overall gaming QoE. Concretely, the aspects of flow, performance, and playing quality were removed. Furthermore, it was shown that TVQ strongly mediates the positive relationship between input quality and gaming QoE. The SEM reached a very high R-squared of .71 on the subjective level, and .94 on the MOS level. Also, it was shown that input quality and SVQ have the strongest direct effect on gaming QoE. Furthermore, the assumption was confirmed that the PX aspects mediated the relationships between the game-related quality aspects and gaming QoE.

Therefore, due to their dominant impact, for the development of the opinion model ITU-T Rec. G.1072, only the game-related quality aspects SVQ, TVQ, and input quality are used. It was shown in the summary of the opinion model G.1072 that the game-related quality aspects can be predicted well based on the network parameters delay, and packet loss rate, and the encoding parameters resolution, framerate, and bitrate. Considering that the model represents an opinion model, it reached a high performance indicated by an RMSE below 0.5 and PLCC above 0.8. The performance of the model can be even further improved if the game content classification proposed in Section 6.2 is applied.

Chapter 8

Conclusion and Outlook

8.1 Summary

The thesis at hand is focused on creating and evaluating a comprehensive assessment method of gaming Quality of Experience (QoE) for cloud gaming services. The basis for this work was the cloud gaming taxonomy developed in 2013 that targets to cover the majority of relevant influencing factors and quality aspects of gaming QoE for cloud gaming services. The dissertation started with an introduction to components of the cloud gaming taxonomy as well as the quality judgement process (Chapter 2). In a joint effort of the gaming research community and studies presented in this thesis, a great number of influencing factors grouped into user, system, and context factors was identified. This work led to the ITU-T Rec. G.1032. Within the presented work, the most dominant network and encoding parameters were investigated.
Next, the *creation of the assessment method*, composed of the following steps, was targeted:

1. Comprehensive collection of information on available types and methods to assess gaming QoE (Section 3.1)
2. Decision that quantitative subjective methods will be in the focus of the thesis as they can serve as a ground-truth for an empirical investigation of the taxonomy (Section 3.1.3)
3. Selection of suitable methods to cover the majority of quality aspects included in the cloud gaming taxonomy (Section 3.3)
4. Identification of quality aspects for which no assessment tool was available (Section 3.3.2)

The knowledge gained from the previously described steps contributed strongly to the ITU-T Rec. P.809 about subjective assessment methods for gaming QoE. As in the last step, input quality was identified as an aspect for which no measurement tool was available, it was decided to create a new assessment instrument for this purpose.

As for the development of a new questionnaire, a large number of subjective ratings from a diverse target group is required, which would result in a very high effort by using traditional laboratory tests, a new framework using the method of crowdsourcing for gaming QoE assessment was developed (Chapter 4). The framework combines knowledge of the ITU-T Rec. P.808 on the use of crowdsourcing for subjective evaluation of speech quality and the ITU-T Rec. P.809 about the assessment of gaming QoE. In a series of six studies, the impact of artificially added network and encoding degradations was investigated. Here, the item pool for the development of the new questionnaire assessing the input quality was used. The results gathered through the new crowdsourcing method were finally compared to results obtained from lab studies, which revealed that both methods can provide comparable results.

Next, the development of the Gaming Input Quality Scale (GIPS) was presented (Chapter 5), which aims at a psychometrically validated instrument for assessing the input quality of cloud gaming services. The GIPS covers the aspects controllability, responsiveness, and immediate feedback using only 8 items. In addition, a factor named performance indication was derived which may serve as a moderator variable between parameters under test and the input quality. For the development of the tool, a measurement model using an EFA and CFA was created. The model showed excellent model fit as well as reliability and validity measures based on various datasets. However, discriminant issues were found in the case that no packet loss conditions on the control stream are contained in the data.

Following, as the developed method should be applied to create subjective datasets for the development of gaming QoE prediction models, a *test design* was created by the following steps:

1. Conclusion based on an empirical study that a passive test paradigm can be used to some extent to replace time consuming interactive test (Section 3.2)
2. Adaptation and contribution to the standardization activities ITU-T Rec. P.910, P.809, and P.918 (Section 3.3.1)
3. Selection of most important quality factors as dependent variables
4. Distribution of parts of the concise assessment method into post-condition and post-game questionnaires (Section 3.3.2)
5. Deriving requirement specifications on the basis of the ITU-T Rec. G.1032, related work, own empirical studies about the stimulus material selection as well as device properties ([126], [262])

During the last step, in particular, a conducted study revealed that the impact of a network delay can vary strongly even for scenarios within the same game (Section 6.1). This finding has considerable consequences for study designs and quality modeling. Building up on this finding, a classification of game content with respect to the delay sensitivity, frame loss sensitivity, and encoding complexity was developed (Section 6.2). It was shown that the classification reached a satisfying accuracy, and that many study results can be explained by the derived game characteristics.

The developed methods and test design were then used to create a large dataset of subjective ratings for various network and encoding parameters (Section 7.1). Based on this dataset, an *evaluation of the developed assessment method* was performed (Section 7.2) using Structural Equation Modeling (SEM). In the first step, the reliability and validity of the used assessment method based on a measurement model were confirmed. However, similar to reports in the recent literature, only a five-factor solution of the PX aspects emerged. Next, the relationships between the quality aspects covered by the taxonomy were examined through a structural model. It turned out that for the collected data, the aspects playing quality, as well as flow and challenge did not contribute significantly to the judgement of gaming QoE nor on related quality aspects. Additionally, it was shown that the PX aspects mediate the relationships between the game-related aspects and the overall gaming QoE. Also, the game-related aspects by themselves already cover the variability in the gaming QoE ratings very well. Thus, in the last step, it was shown that these aspects and the used test design can be successfully used to develop an opinion model predicting gaming QoE, the ITU-T Rec. G.1072, solely based on network and encoding parameters. Finally, it was shown that the developed game content classification can significantly improve the performance of the ITU-T Rec. G.1072 model.

8.2 Answers to Research Questions

In the remainder of this section, the answers to the research questions that were identified at the start of this dissertation will be discussed. The research questions will highlight how the overarching aim of the thesis – to provide and evaluate a comprehensive assessment method of gaming QoE for cloud gaming services – was achieved.

RQ1 *Is the cloud gaming taxonomy representing all relevant quality aspects?*
It can be concluded that the cloud gaming taxonomy, given the scope of the thesis, indeed covers the majority of relevant quality aspects a player considers for the judgement of gaming QoE. This claim is supported based on the following findings:

- During the series of studies described in this thesis, participants did not report on missing aspects that are conceptually different to those covered by the used assessment methods.
- The literature reviews (cf. Section 3.1.2 and Table 3.5) did not reveal any other fundamentally important aspects that are of relevance for the scope of the thesis (only more player and game design-related concepts were identified).
- The variance of the gaming QoE explained by the structural model in Section 7.3 of 70% and 94% (on the individual level and MOS level, respectively) was very high.

RQ2 *How can the broad range of quality aspects of gaming QoE be measured?*
To solve this task, the following strategy was applied. At first, it was concluded that behavioural and psycho-physical methods on their own cannot be used to evaluate a cloud gaming system accurately and generalizable. Thus, it was decided to focus on questionnaire-based assessments, for which the thesis presents detailed information about available tools. However, the majority of those is composed of an extensive number of items. A combination of such questionnaires (to cover all quality aspects) would only be possible in multiple consecutive studies as this would cause a high fatigue of test participants. Therefore, it was decided to only use questionnaires with a limited number of items per quality aspect (ideally 2-3 items). The iGEQ was identified as the only questionnaire that measures a broad range of player experience (PX) aspects in a concise way. Whereas the PXI seems also to be a suitable candidate, the questionnaire was validated only in 2020. Consequently, the iGEQ using 14 items to assess 7 PX aspects was selected. The iGEQ was investigated using a factor analysis and adjusted to produce reliable results for five distinct quality aspects. As for the input quality no measurement tool was available, the concise GIPS (8 items) was developed and validated based on multiple datasets. Finally, the game-related aspects playing quality and appeal are measured in the post-game questionnaire and not during the post-condition questionnaire to further reduce the number of items required for each test condition. The resulting assessment method allows the measurement of the broad range of quality aspects in a similar time frame as the stimulus duration of interactive tests, and was shown to provide highly reliable results.

RQ3 *How should a subjective test for assessing gaming QoE be designed?*
The dimensionality of gaming QoE and the number of quality factors is complex to such an extent that certain aspects must be limited. While targeting a test design allowing to build a quality prediction model, this was achieved due to the following steps:

- Based on the overview of quality factors provided by the ITU-T Rec. G.1032 and literature review, it was decided which factors will be used as dependent variables, whereas the remaining factors should be kept as constant as possible.
- It was concluded that from the perspective of network and services providers, the following parameters must be investigated: network delay, down-link packet loss, encoding framerate, encoding bitrate, and video resolution.
- The thesis provides guidance to select the network and encoding parameters to span the full range of the proposed rating scales (decisions can also be made based on ITU-T Rec. G.1072, or due to video quality metrics such as VMAF [263] or NDNetGaming [264]).
- Based on an empirical study, it was shown that it is possible to perform passive tests for a detailed evaluation of encoding parameters.
- Interactive tests are used to investigate the most dominant influencing factors on the interaction quality, i.e., network impairments and encoding framerates.
- A block design in which all conditions are tested for one game at a time was used, and randomization of conditions within each block was performed to reduce order effects.
- Finally, the selection of game scenarios was reduced by choosing representative scenarios for each content class presented in Section 6.2.

On the example of the test design presented in Section 3.3.2, it was shown that for interactive tests, 17 different conditions for one game, and for the passive tests, 72 conditions for 3 games can be tested within one subjective test session.

RQ4 *Is there an alternative to traditional laboratory studies for gaming QoE assessment?*
In Chapter 4, a new approach using the method of crowdsourcing was presented. The method was developed based on the ITU-T Rec. P.808 and P.809 and it was shown that results obtained from traditional laboratory studies are comparable to those collected using the crowdgaming framework. Furthermore, the results were in line with the expected relationships of investigated system parameters and assessed quality aspects. The crowdgaming method was successfully used to develop the GIPS and for a series of studies related to game characteristics [18], player adaptations [210], and delay compensation methods [82], [211]. While the method can still be improved with respect to the setup, e.g., using web-based streaming services such as Google Stadia to test high-quality games, the underlying method has proven to be sound and useful.

RQ5 *How relevant are the individual quality aspects for the overall gaming QoE?*
Based on the structural model presented in Chapter 7, all relationships among the relevant quality aspects are described quantitatively. The SEM analysis has shown that game-related quality aspects are equally important if not dominant compared to PX aspects with respect to their impact on the overall gaming QoE. The PX aspects mediated the relationships between the game-related aspects and the overall gaming QoE, similar to the results from Abelle et al. presented in [97]. Furthermore, they only explained a small additional portion of the variability of the gaming QoE ratings. The aspects flow,

challenge, and playing quality did not show any significant impact on the gaming QoE nor did they serve as mediators. However, it must be noted that these findings only apply to the collected data. If game scenarios that are not well designed are in the focus of research, it appears plausible that the importance of these aspects increases.

8.3 Contribution of Thesis

In summary, the main contributions of the thesis are

1. An exhaustive and *empirical validation of a taxonomy* of quality aspects of cloud gaming services as well as applied assessment methods.
2. A *crowdgaming framework* to conduct quality assessment studies in a crowdsourcing environment which potentially increases the ecological validity of study results.
3. The GIPS, a psychometrically validated, and reliable instrument to measure the *input quality*, was developed on the basis of a large dataset. GIPS closes a critical gap to assess the full spectrum of relevant quality aspects and has proven to be a highly useful tool for gaming QoE research.
4. Insights on how the *quality judgement process* of players might be composed of based on a structural model of the cloud gaming taxonomy.
5. Complementary findings related to various *quality factors* such as network and encoding parameters, test paradigm, and content factors.
6. Guidelines to create *test designs* which cover all relevant steps to collect data for the development of quality prediction models.
7. An expert-based *classification* of game content with respect to the impact of network and encoding impairments.
8. Contributions *ITU-T Recommendations* G.1032, P.809, and G.1072, which can lead to more reliable, valid, and comparable research results in the future.

8.4 Limitations and Future Work

In the following last section, some topics covered by the presented research or related to the topic will be discussed with respect to limitations and possible future work.

1 – Generalizability of findings

Users: Even though it was argued that the participants in the conducted studies, especially those in the laboratory environments, resemble the target group of cloud gaming service users well, it must be noted that they were predominantly students of the Technische Universität Berlin. Thus, they do not ideally cover the population of players. While it was shown that very unexperienced users rated test conditions without any impairments lower than experienced participants, they may also not perceive degradations as much as expert gamers. The user factors in the thesis do not sufficiently cover parameters that describe the sensitivity of users towards degradations. In the future, pre-screening methods such as assessing reaction times, or detection rates of missing frames might be considered. Lastly, no social interactions between the players are considered in the presented research.
Games: While the games used in the crowdgaming tests were of an abstract nature, the games selected for the lab studies were all very well designed. The importance of the aspects appeal and playing

quality might increase if games with certain design flaws would be used. Also, whereas for the G.1072 dataset participants were offered to select a game of their liking, this did not apply to the other studies.
Systems: Neither the test environment, nor the screen size as investigated in [71] and [126], nor the input device as shown in [262] were shown to be influential quality factors. Thus, there is a high likelihood that the methods and findings of the thesis also apply to other gaming systems such as traditional online gaming and mobile cloud gaming to some extent. Further studies in this direction would be required, which could be interesting future work. Also, it would be interesting to investigate whether the developed models can be applied to virtual reality gaming.
Data: While a large effort was made to collect a large amount of data for various games and test conditions, the availability of additional data would have helped to improve the validation of results and findings reported in this thesis, especially for the work reported in Section 3.2.

2 – Stability of findings

There is a high chance that the relationships of quality aspects found based on the SEM also hold true in several years, but it is likely that the influence of technical parameters on them will change in the future. Cloud gaming is a rapidly developing technology, and hence the expectations, i.e., also the desired quality features, of players will likely change over time. Some of these technical advancements may include compensation or concealment techniques for delays, advanced network protocols that handle packet losses more efficiently, but also higher resolutions, framerates, and better video codecs such as AV1. As it was shown that the influence of low framerate was unexpectedly low in various studies presented, such parameters may require long-term studies to reach higher validity.

3 – Ecological validity

When interpreting the results of all conducted studies, it must be considered that the participants did not play the games as they would in their daily life. First, it was clear to participants that they are part of a subjective user study and they potentially devoted more attention to possible degradations than usual. Second, they were not solely intrinsically motivated as a monetary compensation for their effort was paid. The bonus payment for participants, who had the highest performance during subjective tests, most likely only reduced this impact to some extent. Also, participants might be more critical if they would pay for a cloud gaming service. A possibility to reach a higher ecological validity might be to integrate surveys directly within game scenarios played by users in their natural environment and motivate them to participate due to in-game rewards. The presented crowdgaming framework, which can be improved by using web-based cloud gaming solutions such as Google Stadia or Parsec, could serve as a good starting point.

4 – Test design

While it was shown that reliable results can be collected by the used methods and test designs, there might still be some ways to improve them further. First, the selection of the stimulus duration should be invested further. The suggestions of the ITU-T Rec. P.809 regarding the stimulus duration were of a rather practical nature and should be confirmed empirically. This is especially important as the low impact of flow might have been caused by an inappropriate stimulus duration. Second, as advertised in the thesis, a multi-method approach combining behavioural, psycho-physical, and

subjective assessment methods promises several advantages especially when targeting changes of QoE over time, the perception of degradations, or advanced data cleansing approaches. Third, the use of trapping questions in laboratory studies might be considered and a more advanced participant screening could be developed. Lastly, other scaling methods such as DCR could be investigated which may allow more insights into perception thresholds linked to quality factors.

5 – GIPS development

While every effort was taken to make sure that most of the important factors for construct input quality were identified during the expert interviews and by the feedback of participants in studies, there could be some factors which were not considered. Further, even though the dataset used for the GIPS development was very large, additional iterations might be helpful to confirm the claim that the revealed discriminant issues are caused by the absence of control stream packet losses in some validation datasets. Finally, the scale could be validated further across various user groups as well as other types of games.

6 – Game content classification

While the content classification presented in Section 6.2 can be considered as a significant step for decrypting the influence of the game content towards network and encoding impairments, it still relies on human judgement of the identified characteristics. An automatization to determine the content classes would be highly beneficial for cloud gaming providers and researchers. A potential approach was recently published in [265].

7 – Quality modeling

One research goal was enabling the development of quality prediction models for cloud gaming services. While this was already achieved in form of an opinion model, the ITU-T Rec. G.1072, more advanced models can be created. The following considerations could be made

- Improve the parametrization of systems, e.g., an objective description of the interplay of delay, data throughput, and packet loss leading a frame loss rate as described in [203].
- Use more realistic simulations of network parameters, e.g., Pareto distribution for delay or recordings of real network traffic.
- Consider new challenges evoked by 5G networks, edge cloud architectures, and shared resources with respect to VMs and GPUs.
- Conduct more research targeting a parametric description of user characteristics leading to models such as the proposed ARCU model in [92].
- Investigate different rating scales (similar to the work presented in [188]) to allow combining MOS-scale and EC-ACR results for quality modeling or to compare research findings.

Furthermore, building upon the presented research, two highly interesting projects were already started at the ITU-T SG12:

- Development of gaming QoE models for (mobile) online gaming (cf. ITU-T work item G.OMMOG [266])
- Development of gaming QoE monitoring models for cloud gaming (cf. ITU-T work item P.BBQCG [267])

References

[1] M. J. Wolf, *The Video Game Explosion: A History From Pong to Playstation and Beyond*. ABC-CLIO, 2008, ISBN: 978-0-313-08243-6.

[2] T. Wijman, *The World's 2.7 Billion Gamers Will Spend $159.3 Billion on Games in 2020; The Market Will Surpass $200 Billion by 2023*, Available at: https://newzoo.com/insights/articles/newzoo-games-market-numbers-revenues-and-audience-2020-2023/, 2020.

[3] S. Stewart, *Video Game Industry Silently Taking Over Entertainment World*, Available at: https://www.ejinsight.com/eji/article/id/2280405/20191022-video-game-industry-silently-taking-over-entertainment-world, 2019.

[4] Entertainment Software Association, *2020 Essential Facts About the Video Game Industry*, Available at: https://www.theesa.com/esa-research/2020-essential-facts-about-the-video-game-industry/, 2020.

[5] G. Fernandes, *Half a Billion Dollars in 2020: The Cloud Gaming Market Evolves as Consumer Engagement and Spending Soar*, Available at: https://newzoo.com/insights/articles/global-cloud-gaming-market-report-consumer-engagement-spending-revenues-2020-2023/, 2020.

[6] S. Rosenbaum, *How does cloud gaming work? - A picture book explanation of the latest gaming trend*, Available at: https://www.polygon.com/2020/10/14/21410181/cloud-gaming-101-explainer-internet, 2020.

[7] S. Möller, D. Pommer, J. Beyer, and J. Rake-Revelant, "Factors Influencing Gaming QoE: Lessons Learned From the Evaluation of Cloud Gaming Services", in *Proceedings of the 4th International Workshop on Perceptual Quality of Systems (PQS 2013)*, 2013, pp. 1–5. DOI: 10.21437/PQS.2013-31.

[8] R. Shea, J. Liu, E. Ngai, and Y. Cui, "Cloud gaming: Architecture and performance", *Network, IEEE*, vol. 27, pp. 16–21, Jul. 2013. DOI: 10.1109/MNET.2013.6574660.

[9] Qualinet White Paper on Definitions of Quality of Experience, *COST Action IC 1003*, P. Le Callet, S. Möller, and A. Perkis, Eds., 2013.

[10] R. Bernhaupt, *Evaluating User Experience in Games: Concepts and Methods*. 2010, p. 287, ISBN: 9781848829626. DOI: 10.1007/978-1-84882-963-3.

[11] S. Möller, S. Schmidt, and S. Zadtootaghaj, "New ITU-T Standards for Gaming QoE Evaluation and Management", in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2018, pp. 1–6. DOI: 10.1109/QoMEX.2018.8463404.

[12] S. Möller, S. Schmidt, and J. Beyer, "Gaming Taxonomy: An Overview of Concepts and Evaluation Methods for Computer Gaming QoE", in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, 2013, pp. 236–241. DOI: 10.1109/QoMEX.2013.6603243.

[13] S. Schmidt, *Messung der Quality of Experience von Computerspielen*, unpublished, B.Sc. thesis, Berlin, Jan. 2013.

[14] S. Schmidt, S. Zadtootaghaj, and S. Möller, "Towards the Delay Sensitivity of Games: There Is More Than Genres", in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–6. DOI: 10.1109/QoMEX.2017.7965676.

[15] S. Schmidt, S. Möller, and S. Zadtootaghaj, "A Comparison of Interactive and Passive Quality Assessment for Gaming Research", in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2018, pp. 1–6. DOI: 10.1109/QoMEX.2018.8463417.

[16] S. Schmidt, B. Naderi, S. S. Sabet, S. Zadtootaghaj, and S. Möller, "Assessing Interactive Gaming Quality of Experience Using a Crowdsourcing Approach", in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2020, pp. 1–6. DOI: 10.1109/QoMEX48832.2020.9123122.

[17] S. Zadtootaghaj, S. Schmidt, N. Barman, S. Möller, and M. G. Martini, "A Classification of Video Games Based on Game Characteristics Linked to Video Coding Complexity", in *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*, IEEE, 2018, pp. 1–6. DOI: 10.1109/NetGames.2018.8463434.

[18] S. S. Sabet, S. Schmidt, S. Zadtootaghaj, C. Griwodz, and S. Möller, "Delay Sensitivity Classification of Cloud Gaming Content", in *Proceedings of the 12th ACM International Workshop on Immersive Mixed and Virtual Environment Systems*, ser. MMVE '20, Istanbul, Turkey: Association for Computing Machinery, 2020, pp. 25–30, ISBN: 9781450379472. DOI: 10.1145/3386293.3397116.

[19] S. Schmidt, S. Zadtootaghaj, and S. Möller, "Updates on the first draft of Influence Factors in Gaming Quality of Experience (QoE)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.41, 2017.

[20] S. Zadtootaghaj, S. Schmidt, and S. Möller, "Influence Factors on Gaming Quality of Experience (QoE)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.104, 2017.

[21] S. Schmidt, S. Zadtootaghaj, and S. Möller, "Update on the Proposal for a Draft New Recommendation on Subjective Evaluation Methods for Gaming Quality (P.GAME)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.17, 2017.

[22] S. Schmidt, S. Zadtootaghaj, S. Möller, F. Metzger, M. Hirth, and M. Sužnjević, "Update on the Proposal for a Draft New Recommendation on Subjective Evaluation Methods for Gaming Quality (P.GAME)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.98, 2017.

[23] S. Schmidt, S. Zadtootaghaj, S. Möller, F. Metzger, M. Hirth, and M. Sužnjević, "Subjective Evaluation Methods for Gaming Quality (P.GAME)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.205, 2018.

[24] S. Schmidt, S. Zadtootaghaj, S. Möller, F. Metzger, M. Hirth, M. Sužnjević, N. Barman, and M. G. Martini, "Requirement Specification and Possible Structure for an Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.200, 2018.

[25] S. Schmidt, S. Zadtootaghaj, F. Schiffner, S. Möller, S. S. Sabet, C. Griwodz, N. Barman, and M. G. Martini, "Data Assessment for an Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.293, 2018.

[26] S. Schmidt, S. Zadtootaghaj, M. Utke, S. Möller, N. Barman, M. G. Martini, S. S. Sabet, and C. Griwodz, "First Draft for an Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.387, 2019.

[27] S. Schmidt, S. Zadtootaghaj, S. Möller, and S. S. Sabet, "Proposal for an Opinion Model Predicting Gaming QoE for Mobile Online Gaming", ITU-T Study Group 12, Geneva, ITU-T Contribution C.441, 2019.

[28] S. Schmidt, S. S. Sabet, S. Zadtootaghaj, S. Möller, C. Griwodz, N. Barman, and M. G. Martini, "Proposal of a Content Classification for Cloud Gaming Services", ITU-T Study Group 12, Geneva, ITU-T Contribution C.444, 2019.

[29] S. Schmidt, S. Zadtootaghaj, S. Möller, N. Barman, M. G. Martini, S. S. Sabet, and C. Griwodz, "Performance Evaluation of the Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.445, 2019.

[30] S. Schmidt, S. Zadtootaghaj, S. Möller, B. Nabajeet, M. G. Martini, S. S. Sabet, and C. Griwodz, "Opinion Model Predicting Gaming QoE (G.OMG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.446, 2019.

[31] S. Schmidt, S. Zadtootaghaj, and S. Möller, "Corrigendum for ITU-T Recommendation G.1072: Opinion Model Predicting Gaming QoE ", ITU-T Study Group 12, Geneva, ITU-T Contribution C.511, 2020.

[32] S. Schmidt, B. Naderi, S. Zadtootaghaj, and S. Möller, "Guidelines for the Assessment of Gaming QoE Using Crowdsourcing", ITU-T Study Group 12, Geneva, ITU-T Contribution C.376, 2019.

[33] B. Naderi, S. Schmidt, S. Zadtootaghaj, and S. Möller, "Draft text for P.CROWDG Recommendation Subjective Evaluation of Gaming Quality with a Crowdsourcing Approach", ITU-T Study Group 12, Geneva, ITU-T Contribution C.437, 2019.

[34] S. Schmidt, S. S. Sabet, B. Naderi, S. Zadtootaghaj, C. Griwodz, and S. Möller, "Evaluation of Interactive Test Paradigm for P.CROWDG Recommendation", ITU-T Study Group 12, Geneva, ITU-T Contribution C.485, 2020.

[35] S. Möller and A. Raake, *Quality of Experience: Advanced Concepts, Applications and Methods*. Jan. 2014, ISBN: 978-3-319-02680-0. DOI: 10.1007/978-3-319-02681-7.

[36] J. Juran and A. B. Godfrey, *Quality Handbook*, 5th ed. 1999, ISBN: 9780070340039.

[37] ITU-T Recommendation E.800, *Definitions of Terms Related to Quality of Service*. Geneva: International Telecommunication Union, 2008.

[38] U. Jekosch, *Voice and Speech Quality Perception: Assessment and Evaluation*. Springer-Verlag Berlin Heidelberg, 2005. DOI: 10.1007/3-540-28860-0.

[39] A. Raake, *Speech Quality of VoIP*. Wiley Online Library, 2006, ISBN: 978-0-470-03299-2. DOI: 10.1002/9780470033005.

[40] M. Wältermann, *Dimension-Based Quality Modeling of Transmitted Speech*. Springer Science & Business Media, 2013, ISBN: 978-3-642-35018-4. DOI: 10.1007/978-3-642-35019-1.

[41] F. Köster, *Multidimensional Analysis of Conversational Telephone Speech*. Springer, 2018, ISBN: 978-981-10-5223-1. DOI: 10.1007/978-981-10-5224-8.

[42] A. Hines and J. D. Kelleher, "A Framework for Post-Stroke Quality of Life Prediction Using Structured Prediction", in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–6. DOI: 10.1109/QoMEX.2017.7965672.

[43] N. Côté, *Integral and Diagnostic Intrusive Prediction of Speech Quality*. Springer Science & Business Media, 2011, ISBN: 978-3-642-18462-8. DOI: 10.1007/978-3-642-18463-5.

[44] ITU-T Recommendation P.800, *Methods for Subjective Determination of Transmission Quality*. Geneva: International Telecommunication Union, 1996.

[45] ITU-T Recommendation P.10/G.100, *Vocabulary for Performance, Quality of Service and Quality of Experience*. Geneva: International Telecommunication Union, 2017.

[46] F. Hammer, S. Egger-Lampl, and S. Möller, "Quality-of-User-Experience: a Position Paper", *Quality and User Experience*, vol. 3, no. 1, p. 9, 2018. DOI: 10.1007/s41233-018-0022-0.

[47] J. Nielsen, *Usability Engineering*. Morgan Kaufmann, 1994, ISBN: 978-0-08-052029-2.

[48] J. Preece, Y. Rogers, H. Sharp, D. Benyon, S. Holland, and T. Carey, *Human-computer Interaction*. Addison-Wesley Longman Ltd., 1994, ISBN: 978-0-201-62769-5.

[49] M. Hassenzahl and N. Tractinsky, "User Experience-a Research Agenda", *Behaviour & information technology*, vol. 25, no. 2, pp. 91–97, 2006. DOI: 10.1080/01449290500330331.

[50] M. Hassenzahl, S. Diefenbach, and A. Göritz, "Needs, Affect, and Interactive Products–Facets of User Experience", *Interacting with computers*, vol. 22, no. 5, pp. 353–362, 2010. DOI: 10.1016/j.intcom.2010.04.002.

[51] D. Kahneman, E. Diener, and N. Schwarz, *Well-Being: Foundations of Hedonic Psychology*. Russell Sage Foundation, 1999, ISBN: 9780871544247.

[52] S. Diefenbach, N. Kolb, and M. Hassenzahl, "The Hedonic in Human-Computer Interaction: History, Contributions, and Future Research Directions", in *Proceedings of the 2014 conference on Designing interactive systems*, 2014, pp. 305–314. DOI: 10.1145/2598510.2598549.

[53] E. D. Mekler and K. Hornbæk, "Momentary Pleasure or Lasting Meaning? Distinguishing Eudaimonic and Hedonic User Experiences", in *Proceedings of the 2016 chi conference on human factors in computing systems*, 2016, pp. 4509–4520. DOI: 10.1145/2858036.2858225.

[54] E. L.-C. Law, "The Measurability and Predictability of User Experience", New York, NY, USA: Association for Computing Machinery, 2011, ISBN: 9781450306706. DOI: 10.1145/1996461.1996485.

[55] J. A. Bargas-Avila and K. Hornbæk, "Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience", New York, NY, USA: Association for Computing Machinery, 2011, ISBN: 9781450302289. DOI: 10.1145/1978942.1979336.

[56] E. L.-C. Law, P. van Schaik, and V. Roto, "Attitudes Towards User Experience (UX) Measurement", *International Journal of Human-Computer Studies*, vol. 72, no. 6, pp. 526–541, 2014. DOI: 10.1016/j.ijhcs.2013.09.006.

[57] J. Juul, *Half-Real - Video Games Between Real Rules and Fictional Worlds*. MIT Press, 2005, ISBN: 9780262101103.

[58] N. Lazzaro and K. Keeker, "What's My Method? A Game Show on Games", CHI EA '04, pp. 1093–1094, 2004. DOI: 10.1145/985921.985922.

[59] W. Cai, R. Shea, C.-Y. Huang, K.-T. Chen, J. Liu, V. C. Leung, and C.-H. Hsu, "A Survey on Cloud Gaming: Future of Computer Games", *IEEE Access*, vol. 4, pp. 7605–7620, 2016. DOI: 10.1109/ACCESS.2016.2590500.

[60] ITU-T Recommendation G.1032, *Influence Factors on Gaming Quality of Experience*. Geneva: International Telecommunication Union, 2017.

[61] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hoßfeld, "An Evaluation of QoE in Cloud Gaming Based on Subjective Tests", June, 2011, pp. 330–335, ISBN: 9780769543727. DOI: 10.1109/IMIS.2011.92.

[62] C.-Y. Huang, K.-T. Chen, D.-Y. Chen, H.-J. Hsu, and C.-H. Hsu, "GamingAnywhere: The first Open Source Cloud Gaming System", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 10, no. 1s, pp. 1–25, 2014. DOI: 10.1145/2537855.

[63] M. Manzano, J. A. Hernandez, M. Uruena, and E. Calle, "An Empirical Study of Cloud Gaming", in *2012 11th Annual Workshop on Network and Systems Support for Games (NetGames)*, IEEE, 2012, pp. 1–2. DOI: 10.1109/NetGames.2012.6404021.

[64] N. Barman and M. G. Martini, "H.264/MPEG-AVC, H.265/MPEG-HEVC and VP9 Codec Comparison for Live Gaming Video Streaming", in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–6. DOI: 10.1109/QoMEX.2017.7965686.

[65] M. Carrascosa and B. Bellalta, "Cloud-Gaming: Analysis of Google Stadia Traffic", *arXiv preprint arXiv:2009.09786*, 2020.

[66] C.-Y. Huang, C.-H. Hsu, Y.-C. Chang, and K.-T. Chen, "GamingAnywhere: An Open Cloud Gaming System", *Proceedings of the 4th ACM Multimedia Systems Conference on - MMSys '13*, vol. 2, no. 3, pp. 36–47, 2013. DOI: 10.1145/2483977.2483981.

[67] S. Choy, B. Wong, G. Simon, and C. Rosenberg, "A Hybrid Edge-Cloud Architecture for Reducing on-Demand Gaming Latency", *Multimedia systems*, vol. 20, no. 5, pp. 503–519, 2014. DOI: 10.1007/s00530-014-0367-z.

[68] R. Shea and J. Liu, "On GPU Pass-Through Performance for Cloud Gaming: Experiments and Analysis", in *2013 12th Annual Workshop on Network and Systems Support for Games (NetGames)*, IEEE, 2013, pp. 1–6. DOI: 10.1109/NetGames.2013.6820614.

[69] K.-T. Chen, C.-Y. Huang, and C.-H. Hsu, "Cloud Gaming Onward: Research Opportunities and Outlook", in *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, IEEE, 2014, pp. 1–4. DOI: 10.1109/ICMEW.2014.6890683.

[70] J. Beyer, L. Skorin-Kapov, and J. Van Looy, "Influence Factors in Gaming Quality of Experience (QoE)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.340, 2016.

[71] J. Beyer, "Quality-Influencing Factors in Mobile Gaming", Ph.D. dissertation, Technische Universität Berlin, 2016. DOI: 10.14279/depositonce-6406.

[72] I. Slivar, L. Skorin-Kapov, and M. Sužnjević, "Cloud Gaming QoE Models for Deriving Video Encoding Adaptation Strategies", in *Proceedings of the 7th international conference on multimedia systems*, 2016, pp. 1–12. DOI: 10.1145/2910017.2910602.

[73] M. Sužnjević, L. Skorin-Kapov, and M. Matijasevic, "The Impact of User, System, and Context Factors on Gaming QoE: A Case Study Involving MMORPGs", in *2013 12th Annual Workshop on Network and Systems Support for Games (NetGames)*, IEEE, 2013, pp. 1–6. DOI: 10.1109/NetGames.2013.6820606.

[74] C. I. Jennett, "Is Game Immersion Just Another Form of Selective Attention? an Empirical Investigation of Real World Dissociation in Computer Game Immersion", Ph.D. dissertation, University College London, 2010.

[75] L. Vermeulen, J. Van Looy, F. De Grove, and C. Courtois, "You Are What You Play?: A Quantitative Study Into Game Design Preferences Across Gender and Their Interaction With Gaming Habits", in *DiGRA 2011: Think, design, play*, Digital Games Research Association (DiGRA), 2011.

[76] L. Vermeulen, E. Núñez Castellar, and J. Van Looy, "Challenging the Other: Exploring the Role of Opponent Gender in Digital Game Competition for Female Players", *Cyberpsychology, Behavior, and Social Networking*, vol. 17, no. 5, pp. 303–309, 2014. DOI: 10.1089/cyber.2013.0331.

[77] J. Beyer, V. Miruchna, and S. Möller, "Assessing the Impact of Display Size, Game Type, and Usage Context on Mobile Gaming QOE", *2014 6th International Workshop on Quality of Multimedia Experience, QoMEX 2014*, pp. 69–70, 2014. DOI: 10.1109/QoMEX.2014.6982297.

[78] M. Abdallah, C. Griwodz, K.-T. Chen, G. Simon, P.-C. Wang, and C.-H. Hsu, "Delay-Sensitive Video Computing in the Cloud: A Survey", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 3s, pp. 1–29, 2018. DOI: 10.1145/3212804.

[79] M. Jarschel, D. Schlosser, S. Scheuring, and T. Hoßfeld, "Gaming in the clouds: QoE and the users' perspective", *Mathematical and Computer Modelling*, vol. 57, no. 11-12, pp. 2883–2894, 2013, ISSN: 08957177. DOI: 10.1016/j.mcm.2011.12.014.

[80] A. Sackl, R. Schatz, T. Hossfeld, F. Metzger, D. Lister, and R. Irmer, "QoE Management made uneasy: The case of Cloud Gaming", pp. 492–497, 2016. DOI: 10.1109/ICCW.2016.7503835.

[81] K. Raaen, R. Eg, and C. Griwodz, "Can Gamers Detect Cloud Delay?", in *2014 13th Annual Workshop on Network and Systems Support for Games*, IEEE, 2014, pp. 1–3. DOI: 10.1109/NetGames.2014.7008962.

[82] S. S. Sabet, C. Griwodz, S. Schmidt, S. Zadtootaghaj, and S. Möller, "Towards Applying Game Adaptation to Decrease the Impact of Delay on Quality of Experience", in *IEEE ISM 2018 – The 20th IEEE International Symposium on Multimedia*, electronic, Piscataway, NJ: IEEE, Dec. 2018, pp. 1–6. DOI: 10.1109/ISM.2018.00028.

[83] V. Clincy and B. Wilgor, "Subjective Evaluation of Latency and Packet Loss in a Cloud-Based Game", in *2013 10th International Conference on Information Technology: New Generations*, IEEE, 2013, pp. 473–476. DOI: 10.1109/ITNG.2013.79.

[84] I. Slivar, M. Sužnjević, L. Skorin-Kapov, and M. Matijasevic, "Empirical QoE Study of in-Home Streaming of Online Games", in *2014 13th Annual Workshop on Network and Systems Support for Games*, IEEE, 2014, pp. 1–6. DOI: 10.1109/NetGames.2014.7010133.

[85] H.-J. Hong, C.-F. Hsu, T.-H. Tsai, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu, "Enabling Adaptive Cloud Gaming in an Open-Source Cloud Gaming Platform", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 12, pp. 2078–2091, 2015. DOI: 10.1109/TCSVT.2015.2450173.

[86] I. Slivar, M. Sužnjević, 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 *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, 2015, pp. 1–6. DOI: 10.1109/QoMEX.2015.7148144.

[87] K. T. Claypool and M. Claypool, "On Frame Rate and Player Performance in First Person Shooter Games", *Multimedia systems*, vol. 13, no. 1, pp. 3–17, 2007. DOI: 10.1007/s00530-007-0081-1.

[88] S. Zadtootaghaj, S. Schmidt, and S. Möller, "Modeling Gaming QoE: Towards the Impact of Frame Rate and Bit Rate on Cloud Gaming", in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2018, pp. 1–6. DOI: 10.1109/QoMEX.2018.8463416.

[89] 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", *2015 7th International Workshop on Quality of Multimedia Experience, QoMEX 2015*, 2015. DOI: 10.1109/QoMEX.2015.7148120.

[90] C.-Y. Huang, C.-H. Hsu, D.-Y. Chen, and K.-T. Chen, "Quantifying User Satisfaction in Mobile Cloud Games", in *Proceedings of Workshop on Mobile Video Delivery*, 2014, pp. 1–6. DOI: 10.1145/2579465.2579468.

[91] P. Quax, F. Hautekeete, and K. V. Lier, "A Taxonomy for and Analysis of the MMOG Landscape", 2008. DOI: 10.1145/2577387.2577392.

[92] L. Skorin-Kapov and M. Varela, "A Multi-dimensional View of QoE: the ARCU Model", in *2012 Proceedings of the 35th International Convention MIPRO*, IEEE, 2012, pp. 662–666, ISBN: 9781467325776.

[93] F. Metzger, A. Rafetseder, and C. Schwartz, "A Comprehensive End-to-End Lag Model for Online and Cloud Video Gaming", *5th ISCA/DEGA Work. Percept. Qual. Syst.(PQS 2016)*, pp. 15–19, 2016. DOI: 10.21437/PQS.2016-5.

[94] S. Möller, K.-P. Engelbrecht, C. Kühnel, A. Naumann, I. Wechsung, and B. Weiss, "Evaluation of Multimodal Interfaces for Ambient Intelligence", in *Human-Centric Interfaces for Ambient Intelligence*, Elsevier, 2010, pp. 347–370. DOI: 10.1016/B978-0-12-374708-2.00014-0.

[95] S. Möller, *Quality Engineering: Qualität kommunikationstechnischer Systeme*. Springer-Verlag, 2017, ISBN: 978-3-642-11548-6. DOI: 10.1007/978-3-642-11548-6.

[96] ITU-T Recommendation P.809, *Subjective Evaluation Methods for Gaming Quality*. Geneva: International Telecommunication Union, 2018.

[97] V. V. Abeele, K. Spiel, L. E. Nacke, D. Johnson, and K. Gerling, "Development and Validation of the Player Experience Inventory: A Scale to Measure Player Experiences at the Level of Functional and Psychosocial Consequences", *International Journal of Human-Computer Studies*, vol. 135, p. 102 370, 2020. DOI: 10.1016/j.ijhcs.2019.102370.

[98] J. Gutman, "A Means-End Chain Model Based on Consumer Categorization Processes", *Journal of marketing*, vol. 46, no. 2, pp. 60–72, 1982. DOI: 10.1177/002224298204600207.

[99] J. L. G. Sánchez, F. L. Gutiérrez, M. Cabrera, and N. P. Zea, "Design of Adaptative Video Game Interfaces: A Practical Case of Use in Special Education", in *Computer-aided design of user interfaces VI*, Springer, 2009, pp. 71–76. DOI: 10.1007/978-1-84882-206-1_7.

[100] S. Engl, "Mobile Gaming - Eine empirische Studie zum Spielverhalten und Nutzungserlebnis in mobilen Kontexten", 2010, Magister thesis, Universität Regensburg.

[101] M. Rajanen and J. Tapani, "A Survey of Game Usability Practices in North American Game Companies", 2018.

[102] D. Pinelle, N. Wong, and T. Stach, "Heuristic Evaluation for Games: Usability Principles for Video Game Design", *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, pp. 1453–1462, 2008, ISSN: 1605580112. DOI: 10.1145/1357054.1357282.

[103] I. Vilnai-Yavetz, A. Rafaeli, and C. S. Yaacov, "Instrumentality, Aesthetics, and Symbolism of Office Design", *Environment and Behavior*, vol. 37, no. 4, pp. 533–551, 2005. DOI: 10.1177/0013916504270695.

[104] M. Hassenzahl, A. Platz, M. Burmester, and K. Lehner, "Hedonic and Ergonomic Quality Aspects Determine a Software's Appeal", in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2000, pp. 201–208. DOI: 10.1145/332040.332432.

[105] H. Stelmaszewska, B. Fields, and A. Blandford, "Conceptualising User Hedonic Experience", in *Proceedings of ECCE-12, the 12th European Conference on Cognitive Ergonomics*, 2004, pp. 83–89.

[106] K. Poels, Y. De Kort, and W. Ijsselsteijn, "It Is Always a Lot of Fun! Exploring Dimensions of Digital Game Experience Using Focus Group Methodology", in *Proceedings of the 2007 conference on Future Play*, 2007, pp. 83–89. DOI: 10.1145/1328202.1328218.

[107] C. Murphy, "Why Games Work and the Science of Learning", in *Interservice, Interagency Training, Simulations, and Education Conference*, Citeseer, vol. 21, 2011.

[108] N. Lazzaro, "Why We Play Games: Four Keys to More Emotion Without Story", *Game Developer Conference (GDC)*, pp. 1–8, 2004.

[109] N. Schaffer, "Verifying an Integrated Model of Usability in Games", Ph.D. dissertation, Rensselaer Polytechnic Institute, 2009.

[110] M. Csikszentmihalyi, "Das Flow-Erlebnis. Jenseits von Angst und Langeweile: Im Tun Aufgehen (The Flow Experience; Beyond Fear and Boredom: Opening in Doing)", *Stuttgart: Klett-Cotta*, 1985.

[111] J. Chen, "Flow in Games", *Jenovachen.Com*, p. 20, 2006, ISSN: 00010782. DOI: 10.1145/1232743.1232769.

[112] M. Hassenzahl, "User experience (UX): Towards an experiential perspective on product quality", *Proceedings of the 20th International Conference of the Association Francophone d'Interaction Homme-Machine on - IHM '08*, pp. 11–15, 2008, ISSN: 1605582859. DOI: 10.1145/1512714.1512717.

[113] J. Chen, "Flow in Games (and Everything Else)", *Communications of the ACM*, vol. 50, no. 4, pp. 31–34, 2007. DOI: 10.1145/1232743.1232769.

[114] U. Hugentobler, "Messen von Flow mit EEG in Computerspielen", Ph.D. dissertation, University of Zurich, 2011, p. 196. DOI: 10.5167/uzh-61078.

[115] R. Agarwal and E. Karahanna, "Time Flies When You're Having Fun: Cognitive Absorption and Beliefs About Information Technology Usage", *MIS quarterly*, pp. 665–694, 2000. DOI: 10.2307/3250951.

[116] S. Weniger and C. Loebbecke, "Cognitive Absorption: Literature Review and Suitability in the Context of Hedonic Is Usage", *Department of Business, Media and Technology Management, University of Cologne, Germany*, 2011.

[117] J. Webster and H. Ho, "Audience Engagement in Multimedia Presentations", *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, vol. 28, no. 2, pp. 63–77, 1997. DOI: 10.1145/264701.264706.

[118] B. G. Witmer, C. J. Jerome, and M. J. Singer, "The Factor Structure of the Presence Questionnaire", *Presence: Teleoperators & Virtual Environments*, vol. 14, no. 3, pp. 298–312, 2005. DOI: 10.1162/105474605323384654.

[119] N. C. Nilsson, R. Nordahl, and S. Serafin, "Immersion Revisited: A Review of Existing Definitions of Immersion and Their Relation to Different Theories of Presence", *Human Technology*, vol. 12, no. 2, 2016. DOI: 10.17011/ht/urn.201611174652.

[120] QUALINET White Paper on Definitions of Immersive Media Experience (IMEx), *European Network on Quality of Experience in Multimedia Systems and Services, 14th QUALINET meeting*, A. Perkis and C. Timmerer, Eds., 2020, arxiv.org/abs/2007.07032.

[121] L. Ermi and F. Mäyrä, "Fundamental Components of the Gameplay Experience: Analysing Immersion", *Changing Views: Worlds in Play*, no. January 2011, pp. 15–27, 2005, ISSN: 2342-9666.

[122] M. Slater, "Place Illusion and Plausibility Can Lead to Realistic Behaviour in Immersive Virtual Environments", en, *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, 2009, ISSN: 0962-8436, 1471-2970. DOI: 10.1098/rstb.2009.0138.

[123] M. Lombard and M. T. Jones, "Defining Presence", en, in *Immersed in Media*, M. Lombard, F. Biocca, J. Freeman, W. IJsselsteijn, and R. J. Schaevitz, Eds., Cham: Springer International Publishing, 2015, ISBN: 978-3-319-10189-7. DOI: 10.1007/978-3-319-10190-3_2.

[124] M. Slater, "Measuring Presence: A Response to the Witmer and Singer Presence Questionnaire", en, *Presence: Teleoperators and Virtual Environments*, vol. 8, no. 5, pp. 560–565, 1999, ISSN: 1054-7460, 1531-3263. DOI: 10.1162/105474699566477.

[125] 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. DOI: https://doi.org/10.1145/985921.986048.

[126] S. Schmidt, S. Uhrig, and D. Reuschel, "Investigating the Relationship of Mental Immersion and Physiological Measures During Cloud Gaming", in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6. DOI: 10.1109/QoMEX48832.2020.9123133.

[127] H. L. O'Brien and E. G. Toms, "The Development and Evaluation of a Survey to Measure User Engagement", *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, pp. 50–69, 2010. DOI: https://doi.org/10.1002/asi.21229.

[128] H. L. O'Brien, P. Cairns, and M. Hall, "A Practical Approach to Measuring User Engagement With the Refined User Engagement Scale (UES) and New UES Short Form", *International Journal of Human-Computer Studies*, vol. 112, pp. 28–39, 2018. DOI: https://doi.org/10.1016/j.ijhcs.2018.01.004.

[129] 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. DOI: 10.1016/j.ijhcs.2008.04.004.

[130] B. Vankeirsbilck, T. Verbelen, D. Verslype, N. Staelens, F. De Turck, P. Demeester, and B. Dhoedt, "Quality of Experience Driven Control of Interactive Media Stream Parameters", in *2013 IFIP/IEEE International Symposium on Integrated Network Management (IM 2013)*, IEEE, 2013, pp. 1282–1287.

[131] Y. Gao, N. Bianchi-Berthouze, and H. Meng, "What Does Touch Tell Us About Emotions in Touchscreen-Based Gameplay?", *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 19, no. 4, pp. 1–30, 2012. DOI: 10.1145/2395131.2395138.

[132] A. Martınez-Rodrigo, B. Garcıa-Martınez, R. Alcaraz, P. González, and A. Fernández-Caballero, "Multiscale Entropy Analysis for Recognition of Visually Elicited Negative Stress From Eeg Recordings", *International journal of neural systems*, vol. 29, no. 02, p. 1 850 038, 2019. DOI: 10.1142/s0129065718500387.

[133] Y. S. Ju, J. S. Hwang, S. Kim, and H. J. Suk, "Study of Eye Gaze and Presence Effect in Virtual Reality", in *International Conference on Human-Computer Interaction*, Springer, 2019, pp. 446–449. DOI: 10.1007/978-3-030-23528-4_60.

[134] F. Dehais, A. Dupres, G. Di Flumeri, K. Verdiere, G. Borghini, F. Babiloni, and R. Roy, "Monitoring Pilot's Cognitive Fatigue With Engagement Features in Simulated and Actual Flight Conditions Using an Hybrid fNIRS-EEG passive BCI", in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, IEEE, 2018, pp. 544–549. DOI: 10.1109/smc.2018.00102.

[135] A. Clerico, A. Tiwari, R. Gupta, S. Jayaraman, and T. H. Falk, "Electroencephalography Amplitude Modulation Analysis for Automated Affective Tagging of Music Video Clips", *Frontiers in Computational Neuroscience*, vol. 11, p. 115, 2018. DOI: 10.3389/fncom.2017.00115.

[136] C. G. Burns and S. H. Fairclough, "Use of Auditory Event-Related Potentials to Measure Immersion During a Computer Game", *International Journal of Human-Computer Studies*, vol. 73, pp. 107–114, 2015. DOI: 10.1016/j.ijhcs.2014.09.002.

[137] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, "An Evaluation of Heart Rate and Electrodermal Activity as an Objective QoE Evaluation Method for Immersive Virtual Reality Environments", in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Lisbon, Portugal: IEEE, 2016, pp. 1–6, ISBN: 978-1-5090-0354-9. DOI: 10.1109/QoMEX.2016.7498964.

[138] A. Drachen, L. E. Nacke, G. Yannakakis, and A. L. Pedersen, "Correlation Between Heart Rate, Electrodermal Activity and Player Experience in First-Person Shooter Games", in *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, 2010, pp. 49–54. DOI: 10.1145/1836135.1836143.

[139] G. Du, S. Long, and H. Yuan, "Non-Contact Emotion Recognition Combining Heart Rate and Facial Expression for Interactive Gaming Environments", *IEEE Access*, vol. 8, pp. 11 896–11 906, 2020. DOI: 10.1109/access.2020.2964794.

[140] F. Bevilacqua, H. Engström, and P. Backlund, "Changes in Heart Rate and Facial Actions During a Gaming Session With Provoked Boredom and Stress", *Entertainment Computing*, vol. 24, pp. 10–20, 2018. DOI: 10.1016/j.entcom.2017.10.004.

[141] A. Denisova, "Adaptive Technologies in Digital Games: The Influence of Perception of Adaptivity on Immersion", Ph.D. dissertation, University of York, 2016.

[142] M. Lombard, T. B. Ditton, and L. Weinstein, "Measuring Presence: The Temple Presence Inventory", in *Proceedings of the 12th annual international workshop on presence*, 2009, pp. 1–15.

[143] U. Jekosch, *Sprache Hören und Beurteilen: Ein Ansatz zur Grundlegung der Sprachqualitätsbeurteilung*, habilitation, Universität/Gesamthochschule, Essen, 2000.

[144] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J.-N. Antons, K. Y. Chan, N. Ramzan, and K. Brunnström, "Psychophysiology-Based QoE Assessment: A Survey", *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 6–21, 2016. DOI: 10.1109/JSTSP.2016.2609843.

[145] R. J. Pagulayan, K. Keeker, D. Wixon, R. L. Romero, and T. Fuller, "User-Centered Design in Games", in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, USA: L. Erlbaum Associates Inc., 2002, pp. 883–906, ISBN: 978-0-8058-3838-1.

[146] A. I. Nordin, A. Denisova, and P. Cairns, "Too Many Questionnaires: Measuring Player Experience Whilst Playing Digital Games", *Seventh York Doctoral Symposium on Computer Science & Electronics*, no. October, pp. 69–75, 2014.

[147] M. H. Phan, J. R. Keebler, and B. S. Chaparro, "The Development and Validation of the Game User Experience Satisfaction Scale (GUESS)", *Human factors*, vol. 58, no. 8, pp. 1217–1247, 2016. DOI: 10.1177/0018720816669646.

[148] M. A. Federoff, "Heuristics and Usability Guidelines for the Creation and Evaluation of Fun in Video Games", Ph.D. dissertation, Citeseer, 2002.

[149] V. V. Abeele, L. E. Nacke, E. D. Mekler, and D. Johnson, "Design and Preliminary Validation of the Player Experience Inventory", in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, 2016, pp. 335–341. DOI: 10.1145/2968120.2987744.

[150] J. Wiemeyer, L. Nacke, C. Moser, and F. Floyd Müller, "Player Experience", in *Serious Games: Foundations, Concepts and Practice*, R. Dörner, S. Göbel, W. Effelsberg, and J. Wiemeyer, Eds. Springer International Publishing, 2016, pp. 243–271, ISBN: 978-3-319-40612-1. DOI: 10.1007/978-3-319-40612-1_9.

[151] K. Emmerich, N. Bogacheva, M. Bockholt, and V. Wendel, "Operationalization and Measurement of Evaluation Constructs", in *Entertainment Computing and Serious Games: International GI-Dagstuhl Seminar 15283, Dagstuhl Castle, Germany, July 5-10, 2015, Revised Selected Papers*. Springer International Publishing, 2016, pp. 306–331, ISBN: 978-3-319-46152-6. DOI: 10.1007/978-3-319-46152-6_13.

[152] K. Poels, Y. A. de Kort, and W. A. IJsselsteijn, "D3. 3: Game Experience Questionnaire: Development of a Self-Report Measure to Assess the Psychological Impact of Digital Games", 2007.

[153] W. A. IJsselsteijn, Y. A. de Kort, and K. Poels, "The Game Experience Questionnaire", *Eindhoven: Technische Universiteit Eindhoven*, pp. 3–9, 2013.

[154] E. L.-C. Law, F. Brühlmann, and E. D. Mekler, "Systematic Review and Validation of the Game Experience Questionnaire (GEQ)-Implications for Citation and Reporting Practice", in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 2018, pp. 257–270. DOI: 10.31234/osf.io/u94qt.

[155] F. De Grove, J. Van Looy, and C. Courtois, "Towards a Serious Game Experience Model: Validation, Extension and Adaptation of the GEQ for Use in an Educational Context", *Playability and Player Experience*, vol. 10, pp. 47–61, 2010.

[156] D. Johnson, M. J. Gardner, and R. Perry, "Validation of Two Game Experience Scales: The Player Experience of Need Satisfaction (PENS) and Game Experience Questionnaire (GEQ)", *International Journal of Human-Computer Studies*, vol. 118, pp. 38–46, 2018. DOI: 10.1016/j.ijhcs.2018.05.003.

[157] F. Brühlmann and G.-M. Schmid, "How to Measure the Game Experience? Analysis of the Factor Structure of Two Questionnaires", in *Proceedings of the 33rd annual acm conference extended abstracts on human factors in computing systems*, 2015, pp. 1181–1186. DOI: 10.1145/2702613.2732831.

[158] K. L. Norman, "GEQ (Game Engagement/Experience Questionnaire): A Review of Two Papers", *Interacting with computers*, vol. 25, no. 4, pp. 278–283, 2013. DOI: 10.1093/iwc/iwt009.

[159] R. M. Ryan, C. S. Rigby, and A. Przybylski, "The motivational pull of video games: A self-determination theory approach", *Motivation and Emotion*, vol. 30, no. 4, pp. 347–363, 2006, ISSN: 01467239. DOI: 10.1007/s11031-006-9051-8.

[160] K. M. Gerling, M. Miller, R. L. Mandryk, M. V. Birk, and J. D. Smeddinck, "Effects of Balancing for Physical Abilities on Player Performance, Experience and Self-Esteem in Exergames", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 2201–2210. DOI: 10.1145/2556288.2556963.

[161] K. Emmerich and M. Masuch, "The Impact of Game Patterns on Player Experience and Social Interaction in Co-Located Multiplayer Games", in *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 2017, pp. 411–422. DOI: 10.1145/3116595.3116606.

[162] M. Birk and R. L. Mandryk, "Control Your Game-Self: Effects of Controller Type on Enjoyment, Motivation, and Personality in Game", in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 685–694. DOI: 10.1145/2470654.2470752.

[163] E. H. Calvillo Gamez, "On the Core Elements of the Experience of Playing Video Games", Ph.D. dissertation, University College London (UCL), 2009, pp. 1–208.

[164] M. Boberg, E. Karapanos, J. Holopainen, and A. Lucero, "Plexq: Towards a Playful Experiences Questionnaire", *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play - CHI PLAY '15*, pp. 381–391, 2015. DOI: 10.1145/2793107.2793124.

[165] F.-L. Fu, R.-C. Su, and S.-C. Yu, "EGameFlow: A Scale to Measure Learners' Enjoyment of E-Learning Games", *Computers & Education*, vol. 52, no. 1, pp. 101–112, 2009. DOI: 10.1016/j.compedu.2008.07.004.

[166] D. K. Mayes and J. E. Cotton, "Measuring Engagement in Video Games: A Questionnaire", in *Proceedings of the human factors and ergonomics society annual meeting*, SAGE Publications Sage CA: Los Angeles, CA, vol. 45, 2001, pp. 692–696. DOI: 10.1177/154193120104500704.

[167] J. Takatalo, "Psychologically-based and Content-oriented Experience in Entertainment Virtual Environments", Ph.D. dissertation, University of Helsinki, 2011, pp. 1–98, ISBN: 9789521070471. DOI: 10.13140/2.1.4773.5687.

[168] S. A. Jackson and H. W. Marsh, "Development and validation of a scale to measure optimal experience: The Flow State Scale", *Journal of Sport & Exercise Psychology*, vol. 18, pp. 17–35, 1996. DOI: 10.1123/jsep.18.1.17.

[169] F. Rheinberg, R. Vollmeyer, and S. Engeser, "Die Erfassung des Flow-Erlebens (Capturing the Flow Experience)", in *Diagnostik von Motivation und Selbstkonzept, Tests und Trends in der pädagogisch-psychologischen Diagnostik - Band 2*, J. Stiensmeier-Pelster and F. Rheinberg, Eds., Göttingen: Hogrefe, 2003, ISBN: 9783840916748.

[170] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. Mcbroom, K. M. Burkhart, and J. N. Pidruzny, "Journal of Experimental Social Psychology The development of the Game Engagement Questionnaire : A measure of engagement in video game-playing", *Journal of Experimental Social Psychology*, vol. 45, no. 4, pp. 624–634, 2009, ISSN: 0022-1031. DOI: 10.1016/j.jesp.2009.02.016.

[171] M. J. Parnell, "Playing with scales: Creating a measurement scale to assess the experience of video games", *University College London, London, UK*, pp. 1–90, 2009.

[172] M. Chen, B. E. Kolko, E. Cuddihy, and E. Medina, "Modeling and Measuring Engagement in Computer Games", in *DiGRA Conference*, 2005.

[173] M. Chen, B. E. Kolko, E. Cuddihy, and E. Medina, "Modeling but NOT Measuring Engagement in Computer Games", *Proceedings of the 7th international conference on Games + Learning + Society Conference*, GLS'11, pp. 55–63, 2011.

[174] Igroup Presence Questionnaire (IPQ) Overview, *A Multi-Disciplinary Project Consortium Addressing New Interfaces Between Humans and the Real and Virtual Environment*, Available at: http://www.igroup.org/pq/ipq/index.php.

[175] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff, "Development of a New Cross-Media Presence Questionnaire: The ITC-Sense of Presence Inventory", *Proceedings of PRESENCE*, 2000.

[176] J. Keller and F. Blomann, "Locus of Control and the Flow Experience: An Experimental Analysis", *European Journal of Personality: Published for the European Association of Personality Psychology*, vol. 22, no. 7, pp. 589–607, 2008. DOI: 10.1002/per.692.

[177] D. Pavlas, F. Jentsch, E. Salas, S. M. Fiore, and V. Sims, "The Play Experience Scale: Development and Validation of a Measure of Play", *Human factors*, vol. 54, no. 2, pp. 214–225, 2012. DOI: 10.1177/0018720811434513.

[178] B. G. Witmer and M. J. Singer, "Measuring Presence in Virtual Environments: A Presence Questionnaire", *Presence*, vol. 7, no. 3, pp. 225–240, 1998. DOI: 10.1162/105474698565686.

[179] E. N. Wiebe, A. Lamb, M. Hardy, and D. Sharek, "Measuring engagement in video game-based environments: Investigation of the User Engagement Scale", *Computers in Human Behavior*, vol. 32, pp. 123–132, 2014, ISSN: 07475632. DOI: 10.1016/j.chb.2013.12.001.

[180] G. Barry, P. Schaik, A. Macsween, J. Dixon, and D. Martin, "Exergaming (XBOX Kinect™) Versus Traditional Gym-Based Exercise for Postural Control, Flow and Technology Acceptance in Healthy Adults: A Randomised Controlled Trial", *BMC Sports Science, Medicine and Rehabilitation*, vol. 8, Dec. 2016. DOI: 10.1186/s13102-016-0050-0.

[181] S. A. Jackson and R. C. Eklund, "Assessing Flow in Physical Activity: The Flow State Scale-2 and Dispositional Flow Scale-2", *Journal of sport & exercise psychology*, vol. 24, no. 2, 2002. DOI: 10.1123/jsep.24.2.133.

[182] M. H. Phan, "The development and validation of the game user experience satisfaction scale (GUESS)", Ph.D. dissertation, Wichita State Universit, May 2015.

[183] J. Briard and C. Quinquis, "QoE and Perceptive Quality of Video Game in Passive Mode", ITU-T Study Group 12, Geneva, ITU-T Contribution C.166, 2014.

[184] ITU-T Recommendation P.910, *Subjective Video Quality Assessment Methods for Multimedia Applications*. Geneva: International Telecommunication Union, 2008.

[185] J. Beyer, "Comparison of Interactive and Passive Test Methodologies to Measure Gaming Quality of Experience (QoE)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.390, 2016.

[186] J. Mullin, L. Smallwood, A. Watson, and G. Wilson, "New Techniques for Assessing Audio and Video Quality in Real-Time Interactive Communications", *IHM-HCI Tutorial*, pp. 1–63, 2001.

[187] ITU-T Recommendation P.851, *Subjective quality evaluation of telephone services based on spoken dialogue systems*. Geneva: International Telecommunication Union, 2003.

[188] F. Köster, D. Guse, M. Wältermann, and S. Möller, "Comparison Between the Discrete ACR Scale and an Extended Continuous Scale for the Quality Assessment of Transmitted Speech", *Fortschritte der Akustik, DAGA*, 2015.

[189] J. Beyer and S. Möller, "Proposal for a Draft New Recommendation on a Subjective Evaluation Methods for Gaming Quality (P.GAME)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.203, 2014.

[190] S. Möller, J.-N. Antons, J. Beyer, S. Egger, E. N. Castellar, L. Skorin-Kapov, and M. Sužnjević, "Towards a new ITU-T Recommendation for Subjective Methods Evaluating Gaming QoE", in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, 2015, pp. 1–6, ISBN: 9781479989584. DOI: 10.1109/QoMEX.2015.7148155.

[191] S. Möller, J. Antons, J. Beyer, S. Egger, E. N. Castellar, L. Skorin-Kapov and M. Sužnjević, "Comments on Draft P.GAME", ITU-T Study Group 12, Geneva, ITU-T Contribution C.270, 2015.

[192] A.-F. Perrin, T. Ebrahimi, S. Zadtootaghaj, S. Schmidt, and S. Möller, "Towards the Need Satisfaction in Gaming: A Comparison of Different Gaming Platforms", in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–3. DOI: 10.1109/qomex.2017.7965641.

[193] M. Claypool, "Motion and Scene Complexity for Streaming Video Games", in *Proceedings of the 4th International Conference on Foundations of Digital Games*, 2009, pp. 34–41. DOI: 10.1145/1536513.1536529.

[194] F. Schiffner and S. Möller, "Defining the Relevant Perceptual Quality Space for Video and Video-Telephony", in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2017, pp. 1–3. DOI: 10.1109/qomex.2017.7965662.

[195] ITU-T Recommendation P.918, *Dimension-based Subjective Quality Evaluation for Video Content*. Geneva: International Telecommunication Union, 2020.

[196] F. Schiffner and S. Möller, "Direct Scaling & Quality Prediction for perceptual Video Quality Dimensions", in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2018, pp. 1–3. DOI: 10.1109/qomex.2018.8463431.

[197] J. A. Krosnick and L. R. Fabrigar, "Designing Rating Scales for Effective Measurement in Surveys", *Survey measurement and process quality*, pp. 141–164, 1997. DOI: 10.1002/9781118490013.ch6.

[198] B. Weijters, E. Cabooter, and N. Schillewaert, "The Effect of Rating Scale Format on Response Styles: The Number of Response Categories and Response Category Labels", *International Journal of Research in Marketing*, vol. 27, no. 3, pp. 236–247, 2010. DOI: 10.1016/j.ijresmar.2010.02.004.

[199] A. R. Wildt and M. B. Mazis, "Determinants of Scale Response: Label Versus Position", *Journal of Marketing Research*, vol. 15, no. 2, pp. 261–267, 1978. DOI: 10.2307/3151256.

[200] R. Schatz, S. Egger, and K. Masuch, "The Impact of Test Duration on User Fatigue and Reliability of Subjective Quality Ratings", *Journal of the Audio Engineering Society*, vol. 60, no. 1/2, pp. 63–73, 2012.

[201] S. Hemminger *et al.*, "Network Emulation with NetEm", in *Linux conf au*, 2005, pp. 18–23.

[202] D. Guse, H. R. Orefice, G. Reimers, and O. Hohlfeld, "TheFragebogen: A Web Browser-based Questionnaire Framework for Scientific Research", in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2019, pp. 1–3. DOI: 10.1109/qomex.2019.8743231.

[203] ITU-T Recommendation G.1072, *Opinion Model Predicting Gaming Quality of Experience for Cloud Gaming Services*. Geneva: International Telecommunication Union, 2020.

[204] S. Rigby and R. Ryan, "The Player Experience of Need Satisfaction (PENS)", *Immersive*, 2007.

[205] B. Naderi, *Motivation of Workers on Microtask Crowdsourcing Platforms*. Springer, 2018. DOI: 10.1007/978-3-319-72700-4.

[206] T. Polzehl, B. Naderi, F. Köster, and S. Möller, "Robustness in Speech Quality Assessment and Temporal Training Expiry in Mobile Crowdsourcing Environments", in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[207] B. Naderi, T. Polzehl, I. Wechsung, F. Köster, and S. Möller, "Effect of Trapping Questions on the Reliability of Speech Quality Judgments in a Crowdsourcing Paradigm", in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[208] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best Practices for QoE Crowdtesting: QoE Assessment With Crowdsourcing", *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, Feb. 2014, ISSN: 1520-9210. DOI: 10.1109/TMM.2013.2291663.

[209] D. Saupe, F. Hahn, V. Hosu, I. Zingman, M. Rana, and S. Li, "Crowd Workers Proven Useful: A Comparative Study of Subjective Video Quality Assessment", in *QoMEX 2016: 8th International Conference on Quality of Multimedia Experience*, 2016.

[210] S. S. Sabet, S. Schmidt, C. Griwodz, and S. Möller, "Towards the Impact of Gamers' Adaptation to Delay Variation on Gaming Quality of Experience", in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, IEEE, 2019, pp. 1–6. DOI: 10.1109/qomex.2019.8743239.

[211] S. S. Sabet, S. Schmidt, S. Zadtootaghaj, B. Naderi, C. Griwodz, and S. Möller, "A Latency Compensation Technique Based on Game Characteristics to Mitigate the Influence of Delay on Cloud Gaming Quality of Experience", in *Proceedings of the 11th ACM Multimedia Systems Conference*, ser. MMSys '20, Istanbul, Turkey: Association for Computing Machinery, 2020, pp. 15–25, ISBN: 9781450368452. DOI: 10.1145/3339825.3391855.

[212] ITU-T Recommendation P.808, *Subjective evaluation of speech quality with a crowdsourcing approach*. Geneva: International Telecommunication Union, 2018.

[213] ITU-T Technical Report PSTR-CROWDS, *Subjective Evaluation of Media Quality Using a Crowdsourcing Approach*. Geneva: International Telecommunication Union, 2019.

[214] R. Zequeira Jiménez, L. Fernández Gallardo, and S. Möller, "Outliers Detection vs. Control Questions to Ensure Reliable Results in Crowdsourcing. A Speech Quality Assessment Case Study", in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 1127–1130. DOI: 10.1145/3184558.3191545.

[215] P. K. Paranthaman and S. Cooper, "ARAPID: Towards Integrating Crowdsourced Playtesting into the Game Development Environment", in *Proceedings of the Annual Symposium on CHI in Play*, 2019. DOI: 10.1145/3311350.3347163.

[216] F. Brühlmann, G.-M. Schmid, and E. Mekler, "Online Playtesting With Crowdsourcing: Advantages and Challenges", CHI 2016 Workshop: Lightweight Games User Research for Indies and Non-Profit Organizations, Jan. 2016, p. 4.

[217] I. Guy, A. Hashavit, and Y. Corem, "Games for Crowds: A Crowdsourcing Game Platform for the Enterprise", in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 2015, pp. 1860–1871. DOI: 10.1145/2675133.2675189.

[218] M. Hirth, K. Borchert, F. Allendorf, F. Metzger, and T. Hoßfeld, "Crowd-Based Study of Gameplay Impairments and Player Performance in DotA 2", in *Proceedings of the 4th Internet-QoE Workshop on QoE-based Analysis and Management of Data Communication Networks*, 2019, pp. 19–24. DOI: 10.1145/3349611.3355545.

[219] K. A. Pituch and J. P. Stevens, *Applied Multivariate Statistics for the Social Sciences: Analyses With SAS and IMB's SPSS*. Routledge, 2015, ISBN: 978-0415836654.

[220] D. C. Hoaglin and B. Iglewicz, "Fine-Tuning Some Resistant Rules for Outlier Labeling", *Journal of the American statistical Association*, vol. 82, no. 400, pp. 1147–1149, 1987. DOI: 10.1080/01621459.1987.10478551.

[221] S.-Y. Chen, Z. Feng, and X. Yi, "A General Introduction to Adjustment for Multiple Comparisons", *Journal of thoracic disease*, vol. 9, no. 6, p. 1725, 2017. DOI: 10.21037/jtd.2017.05.34.

[222] B. Naderi, R. Z. Jiménez, M. Hirth, S. Möller, F. Metzger, and T. Hoßfeld, "Towards Speech Quality Assessment Using a Crowdsourcing Approach: Evaluation of Standardized Methods", *Quality and User Experience*, vol. 6, no. 1, pp. 1–21, 2020. DOI: 10.1007/s41233-020-00042-1.

[223] L. Tay and A. Jebb, "Scale Development", *The SAGE encyclopedia of industrial and organizational psychology. Thousand Oaks, CA: Sage*, 2017.

[224] T. R. Hinkin, J. B. Tracey, and C. A. Enz, "Scale Construction: Developing Reliable and Valid Measurement Instruments", *Journal of Hospitality & Tourism Research*, vol. 21, no. 1, pp. 100–120, 1997. DOI: 10.1177/109634809702100108.

[225] N. Schmitt and D. M. Stuits, "Factors Defined by Negatively Keyed Items: The Result of Careless Respondents?", *Applied Psychological Measurement*, vol. 9, no. 4, pp. 367–373, 1985. DOI: 10.1177/014662168500900405.

[226] D. A. Harrison and M. E. McLaughlin, "Exploring the Cognitive Processes Underlying Responses to Self-Report Instruments: Effects of Item Context on Work Attitude Measures", in *Academy of Management Proceedings*, Academy of Management Briarcliff Manor, NY 10510, vol. 1991, 1991, pp. 310–314. DOI: 10.5465/ambpp.1991.4977169.

[227] R. L. Worthington and T. A. Whittaker, "Scale Development Research: A Content Analysis and Recommendations for Best Practices", *The counseling psychologist*, vol. 34, no. 6, pp. 806–838, 2006. DOI: 10.1177/0011000006288127.

[228] D. Djaouti, J. Alvarez, J.-P. Jessel, G. Methel, and P. Molinier, "A Gameplay Definition through Videogame Classification", *International Journal of Computer Games Technology*, vol. 2008, pp. 1–7, 2008, ISSN: 1687-7047. DOI: 10.1155/2008/470350.

[229] K. M. Gerling, M. Birk, R. L. Mandryk, and A. Doucette, "The Effects of Graphical Fidelity on Player Experience", in *Proceedings of international conference on Making Sense of Converging Media*, 2013, pp. 229–236. DOI: 10.1145/2523429.2523473.

[230] I. Wechsung, "An Evaluation Framework for Multimodal Interaction", *Springer International. doi*, vol. 10, pp. 978–3, 2014. DOI: 10.1007/978-3-319-03810-0.

[231] C. Homburg and A. Giering, "Konzeptualisierung und Operationalisierung komplexer Konstrukte. Ein Leitfaden für die Marketingforschung", *Marketing: Zeitschrift für Forschung und Praxis*, vol. 18, no. 1, pp. 5–24, 1996, ISSN: 0344-1369.

[232] P. Cabrera-Nguyen, "Author Guidelines for Reporting Scale Development and Validation Results in the Journal of the Society for Social Work and Research", *Journal of the Society for Social Work and Research*, vol. 1, no. 2, pp. 99–103, 2010. DOI: 10.5243/jsswr.2010.8.

[233] A. Field, *Discovering Statistics Using SPSS*. 2009, vol. 58, p. 821, ISBN: 1847879071.

[234] J. K. Ford, R. C. MacCallum, and M. Tait, "The Application of Exploratory Factor Analysis in Applied Psychology: A Critical Review and Analysis", *Personnel psychology*, vol. 39, no. 2, pp. 291–314, 1986. DOI: 10.1111/j.1744-6570.1986.tb00583.x.

[235] M. Matsunaga, "How to Factor-Analyze Your Data Right: Do's, Don'ts, and How-To's", *International journal of psychological research*, vol. 3, no. 1, pp. 97–110, 2010. DOI: 10.21500/20112084.854.

[236] J. B. Schreiber, A. Nora, F. K. Stage, E. A. Barlow, and J. King, "Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review", *The Journal of educational research*, vol. 99, no. 6, pp. 323–338, 2006. DOI: 10.3200/JOER.99.6.323-338.

[237] L. Hu and P. M. Bentler, "Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives", *Structural equation modeling: a multidisciplinary journal*, vol. 6, no. 1, pp. 1–55, 1999. DOI: 10.1080/10705519909540118.

[238] J. F. Hair, W. C. Black, B. J. Babin, and Anderson, *Multivariate Data Analysis*. Prentice-Hall Upper Saddle River, NJ, 2010, vol. 7, ISBN: 9780135153093.

[239] N. K. Malhotra and S. Dash, *Marketing Research: An Applied Orientation*. Prentice Hall, 2010, vol. Sixth Edition, ISBN: 978-0-13-608543-0.

[240] C. Fornell and D. F. Larcker, "Evaluating Structural Equation Models With Unobservable Variables and Measurement Error", *Journal of marketing research*, vol. 18, no. 1, pp. 39–50, 1981. DOI: 10.1177/002224378101800104.

[241] J. Gaskin J. Lim. (2016). "Model Fit Measures, AMOS Plugin. Gaskination's StatWiki", [Online]. Available: http://statwiki.kolobkreations.com (visited on 10/05/2020).

[242] C. Shu-Hui, W. Wann-Yih, and J. Dennison, "Validation of EGameFlow: A Self-Report Scale for Measuring User Experience in Video Game Play", *Computers in Entertainment (CIE)*, vol. 16, no. 3, pp. 1–15, 2018. DOI: 10.1145/3238249.

[243] Z. Chen, M. S. El-nasr, A. Canossa, J. Badler, S. Tignor, and R. Colvin, "Modeling Individual Differences through Frequent Pattern Mining on Role-Playing Game Actions", no. November, pp. 2–7, 2015.

[244] X. Fang, S. Chan, J. Brzezinski, and C. Nair, "Development of an Instrument to Measure Enjoyment of Computer Game Play", *INTL. Journal of human–computer interaction*, vol. 26, no. 9, pp. 868–886, 2010. DOI: 10.1080/10447318.2010.496337.

[245] S. Heintz and E. L.-c. Law, "The Game Genre Map : A Revised Game Classification", *Chi Play '15*, no. October, pp. 175–184, 2015. DOI: 10.1145/2793107.2793123.

[246] M. J. Wolf, "The Medium of the Video Game", in, University of Texas Press, 2001, ch. Genre and the Video Game, ISBN: 0-292-79150-X.

[247] M. Claypool and K. Claypool, "Latency Can Kill: Precision and Deadline in Online Games", MMSys '10, pp. 215–222, 2010. DOI: 10.1145/1730836.1730863.

[248] M. Claypool and K. T. Claypool, "Latency and Player Actions in Online Games", *Communications of the ACM*, vol. 49, no. 11, p. 40, 2006. DOI: 10.1145/1167838.1167860.

[249] ITU-T Recommendation P.911, *Subjective Audiovisual Quality Assessment Methods for Multimedia Applications*. Geneva: International Telecommunication Union, 1998.

[250] P. Singh, "Impact of Game Characteristics on the Delay Sensitivity of Video Games", unpublished, M.S. thesis, Technische Universität Berlin, Berlin, May 2018.

[251] A. Umme, "Towards a Classification of Video Games with Respect to Their Delay Sensitivity", unpublished, M.S. thesis, Technische Universität Berlin, Berlin, 2019.

[252] E. Aarseth, S. M. Smedstad, and L. Sunnanå, "A Multi-Dimensional Typology of Games", *Proceedings of the 2003 DiGRA International Conference: Level Up*, 2003, ISSN: 2342-9666.

[253] I. Slivar, M. Sužnjević, and L. Skorin-Kapov, "Game Categorization for Deriving QoE-Driven Video Encoding Configuration Strategies for Cloud Gaming", *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 3s, pp. 1–24, 2018. DOI: 10.1145/3132041.

[254] S. Zadtootaghaj, S. Schmidt, S. S. Sabet, S. Möller, and C. Griwodz, "Quality Estimation Models for Gaming Video Streaming Services Using Perceptual Video Quality Dimensions", in *Proceedings of the 11th ACM Multimedia Systems Conference*, ser. MMSys '20, Istanbul, Turkey: Association for Computing Machinery, 2020, pp. 213–224. DOI: 10.1145/3339825.3391872.

[255] J. S. Tanaka, "How big is big enough?: Sample size and goodness of fit in structural equation models with latent variables", *Child development*, pp. 134–146, 1987. DOI: 10.2307/1130296.

[256] S. Bialosiewicz, K. Murphy, and T. Berry, "An Introduction to Measurement Invariance Testing: Resource Packet for Participants", *American Evaluation Association*, pp. 1–37, 2013.

[257] J. E. Gaskin, *Structural Equation Modeling*. MyEducator, 2020.

[258] B. M. Byrne, *Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming*. Routledge, 2013, ISBN: 978-0-8058-6373-4.

[259] R. H. Myers, *Classical and Modern Regression With Applications*. Duxbury press Belmont, CA, 1990, vol. 2, ISBN: 0534380166.

[260] R. B. Kline, *Principles and Practice of Structural Equation Modeling*. Guilford publications, 2015, ISBN: 9781462523344.

[261] ITU-T Recommendation G.107, *The E-model: A Computational Model for Use in Transmission Planning*. Geneva: International Telecommunication Union, 2015.

[262] G. Puri, "Impact of Delay on QoE when Playing with Different Input Devices", unpublished, M.S. thesis, Technische Universität Berlin, Berlin, Jun. 2018.

[263] N. Barman, S. Schmidt, S. Zadtootaghaj, M. G. Martini, and S. Möller, "An Evaluation of Video Quality Assessment Metrics for Passive Gaming Video Streaming", in *Proceedings of the 23rd packet video workshop*, 2018, pp. 7–12. DOI: 10.1145/3210424.3210434.

[264] M. Utke, S. Zadtootaghaj, S. Schmidt, S. Bosse, and S. Möller, "NDNetgaming-Development of a No-Reference Deep CNN for Gaming Video Quality Prediction", *Multimedia Tools and Applications*, pp. 1–23, 2020. DOI: 10.1007/s11042-020-09144-6.

[265] S. Göring, R. Steger, R. R. R. Rao, and A. Raake, "Automated Genre Classification for Gaming Videos", in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2020, pp. 1–6. DOI: 10.1109/mmsp48831.2020.9287122.

[266] S. Schmidt, S. Zadtootaghaj, S. S. Sabet, and S. Möller, "Report about Opinion Model for Mobile Online Gaming Applications (G.OMMOG)", ITU-T Study Group 12, Geneva, ITU-T Contribution C.484, 2020.

[267] S. Zadtootaghaj, S. Schmidt, S. Möller, S. S. Sabet, C. Griwodz, N. Barman, M. G. Martini, R. R. Ramachandra Rao, S. Göring, and A. Raake, "Proposal for new Work Item P.BBQCG: Parametric bitstream-based Quality Assessment of Cloud Gaming Services", ITU-T Study Group 12, Geneva, ITU-T Contribution C.489, 2020.

Appendix A

Additional Material Related to Empirical Studies or Related Work

Table A.1: Rotated pattern matrix for iGEQ items extracted from Law et al. [154]

Item	Factor	F1	F2	F3	F4	F5	F6	F7
I felt skillful	Competence	**0.70**	0.06	-0.02	0.06	0.24	-0.04	-0.05
I felt successful	Competence	**0.60**	0.04	0.10	0.10	0.01	-0.07	-0.09
I was interested in the game's story	Immersion	-0.04	**0.72**	-0.05	0.05	-0.05	0.00	-0.02
I found it impressive	Immersion	0.03	**0.61**	0.02	0.18	0.11	0.02	0.01
I forgot everything around me	Flow	0.11	0.07	**0.72**	-0.13	0.09	-0.09	-0.04
I was fully occupied with the game	Flow	0.15	0.10	**0.45**	0.07	0.14	0.02	0.32
I felt good	Positive Affect	0.21	0.02	0.11	**0.61**	0.05	-0.09	-0.10
I felt content	Positive Affect	0.22	0.00	0.09	**0.58**	-0.02	-0.09	-0.12
I felt challenged	Challenge	0.08	0.05	0.06	0.13	**0.66**	-0.03	0.09
I had to put a lot of effort in to it	Challenge	0.11	0.14	0.03	-0.07	**0.56**	0.11	0.09
I felt irritable	Tension	-0.01	-0.01	0.05	0.01	-0.01	**0.81**	-0.01
I felt frustrated	Tension	-0.13	-0.13	0.04	0.09	0.28	**0.66**	-0.01
I found it tiresome	Negative Affect	0.06	0.11	-0.04	-0.24	0.01	**0.51**	-0.11
I felt bored	Negative Affect	0.09	0.02	-0.04	-0.21	-0.21	**0.46**	-0.21

Table A.2: Statistics of variance and consistency issues per sum of wrongly answered trapping questions (WTQ). For each sum of WTQ, the number and percentage of affected ratings is given.

Wrong Trapping Questions			**Variance Issues**		**Consistency Issues**		**Remaining**	
∑WTQ	ratings	pct	ratings	pct	ratings	pct	ratings	pct
0	1218	71.1	15	1.2	60	4.9	1143	93.8
1	183	10.7	9	4.9	12	6.6	162	88.5
2	57	3.3	3	5.3	9	15.8	45	78.9
3	36	2.1	15	41.7	0	0.0	21	58.3
4	36	2.1	27	75.0	0	0.0	9	25.0
5	129	7.5	111	86.0	0	0.0	18	14.0
6	21	1.2	21	100.0	0	0.0	0	0.0
7	21	1.2	15	71.4	3	14.3	6	28.6
8	12	0.7	9	75.0	0	0.0	3	25.0

Table A.3: Descriptive statistics of subjective ratings in Dataset 4.3 (CS) and Dataset 4.3 (Lab) for test environment comparison between lab studies and CS tests.

		Quality of Experience					Input Quality				
Dataset 4.3 (comparison)	CID	M (lab)	M (CS)	M of CS (mapped)	SD (lab)	SD (CS)	M (lab)	M (CS)	M of CS (mapped)	SD (lab)	SD (CS)
CS1 vs. Lab1	1	3.81	4.17	4.17	0.69	0.55	4.01	4.15	4.15	0.58	0.37
CS1 vs. Lab1	2	2.84	3.04	2.77	0.83	1.17	2.79	3.17	2.87	0.77	0.89
CS2 vs. Lab2	3	4.13	3.74	3.74	0.45	0.57	4.03	4.01	4.01	0.45	0.47
CS2 vs. Lab2	4	2.28	2.56	2.39	0.75	0.89	2.55	2.95	2.67	0.69	0.94
CS2 vs. Lab2	5	1.79	1.70	1.71	0.57	0.60	2.02	2.28	2.07	0.55	0.62
CS2 vs. Lab2	6	1.70	1.79	1.78	0.42	0.64	2.00	2.34	2.13	0.54	0.79
CS2 vs. Lab2	7	1.51	1.35	1.43	0.27	0.25	1.69	1.91	1.73	0.41	0.57
CS2 vs. Lab2	8	1.51	1.35	1.43	0.27	0.21	1.84	2.01	1.83	0.48	0.51
CS2 vs. Lab2	9	4.20	4.15	4.15	0.50	0.57	4.10	4.18	4.18	0.48	0.51
CS2 vs. Lab2	10	3.75	4.14	3.63	0.77	0.50	3.58	4.00	3.62	0.72	0.63
CS2 vs. Lab2	11	2.67	2.87	2.63	0.70	0.90	2.60	2.82	2.56	0.50	0.65
CS2 vs. Lab2	12	2.37	2.91	2.66	0.63	0.82	2.35	2.78	2.52	0.43	0.73
CS2 vs. Lab2	13	2.16	2.42	2.27	0.74	0.92	2.16	2.51	2.28	0.51	0.82
CS2 vs. Lab2	14	2.13	2.24	2.13	0.72	1.07	2.03	2.42	2.20	0.52	0.93
CS3 vs Lab3	15	4.13	3.88	3.88	0.45	0.68	4.03	3.91	3.91	0.45	0.55
CS3 vs Lab3	16	3.10	3.43	3.08	0.99	0.67	3.52	3.74	3.38	0.60	0.52
CS3 vs Lab3	17	2.13	1.78	1.77	0.68	0.69	2.78	2.23	2.03	0.79	0.85
CS3 vs Lab3	18	4.20	4.30	4.30	0.50	0.50	4.10	4.31	4.31	0.48	0.40
CS3 vs Lab3	19	3.48	3.80	3.37	0.89	0.71	3.68	4.11	3.72	0.57	0.46
CS3 vs Lab3	20	1.93	2.64	2.45	0.54	0.73	2.72	3.26	2.95	0.83	0.65
-	Total	**2.79**	**2.91**	**2.79**	**0.62**	**0.68**	**2.93**	**3.15**	**2.96**	**0.57**	**0.64**

Table A.4: Pair-comparison statistics of gaming QoE per condition between Dataset 4.3 (CS) and Dataset 4.3 (Lab) for test environment comparison between lab studies and CS tests. The p-values highlighted in boldly indicate different findings for the corresponding pair-comparison.

Game	Conditions			Laboratory			Crowdsourcing			Crowdsourcing		
	CID (A - B)	A	B	QoE			QoE			QoE (mapped)		
				t	df	p	t	df	p	t	df	p
RE	1 - 2	0ms	300ms	5.98	431.8	<.001	5.78	382.6	<.001	5.54	382.5	<.001
SR	3 - 4	0ms	100ms	11.09	432.5	<.001	6.72	391.1	<.001	9.39	391.2	<.001
SR	3 - 5	0ms	200ms	13.85	433.2	<.001	11.73	391.1	<.001	14.20	391.2	<.001
SR	3 - 6	0ms	200+50ms	14.40	433.2	<.001	11.34	389.7	<.001	13.86	389.8	<.001
SR	3 - 7	0ms	200+100ms	15.14	434.0	<.001	13.21	395.6	<.001	15.54	395.8	<.001
SR	3 - 8	0ms	300ms	15.30	433.6	<.001	13.69	392.5	<.001	16.06	392.7	<.001
SR	4 - 5	100ms	200ms	2.94	432.4	<.001	5.30	384.8	<.001	5.08	384.7	<.001
SR	4 - 6	100ms	200+50ms	3.50	432.4	<.001	4.82	383.7	<.001	4.63	383.6	<.001
SR	4 - 7	100ms	200+100ms	4.50	433.1	<.001	7.05	388.4	<.001	6.78	388.5	<.001
SR	4 - 8	100ms	300ms	4.54	432.8	<.001	7.43	386.0	<.001	7.14	385.9	<.001
SR	5 - 6	200ms	200+50ms	0.55	431.8	0.58	-0.52	383.7	0.60	-0.50	383.6	0.62
SR	5 - 7	200ms	200+100ms	1.61	432.5	0.11	1.92	388.4	0.06	1.85	388.5	0.06
SR	5 - 8	200ms	300ms	1.62	432.2	**0.11**	2.19	386.0	0.03	2.10	385.9	0.04
SR	6 - 7	200+50ms	200+100ms	1.07	432.5	**0.29**	2.45	387.2	0.01	2.35	387.2	0.02
SR	6 - 8	200+50ms	300ms	1.07	432.2	**0.29**	2.73	384.8	0.01	2.62	384.7	0.01
SR	7 - 8	200+100ms	300ms	-0.01	432.1	0.99	0.22	389.7	0.83	0.20	389.8	0.84
TR	9 - 10	0ms	100ms	2.75	431.8	0.01	0.17	388.7	**0.86**	3.98	388.7	<.001
TR	9 - 11	0ms	200ms	9.28	431.8	<.001	7.30	391.4	<.001	10.71	391.5	<.001
TR	9 - 12	0ms	200+50ms	10.99	432.4	<.001	7.35	392.3	<.001	10.82	392.4	<.001
TR	9 - 13	0ms	200+100ms	12.25	432.4	<.001	10.41	388.4	<.001	13.76	388.4	<.001
TR	9 - 14	0ms	300ms	12.59	431.8	<.001	12.21	384.4	<.001	15.62	384.3	<.001
TR	10 - 11	100ms	200ms	6.53	431.8	<.001	7.00	394.5	<.001	6.72	394.6	<.001
TR	10 - 12	100ms	200+50ms	8.27	432.4	<.001	7.06	391.2	<.001	6.78	391.2	<.001
TR	10 - 13	100ms	200+100ms	9.53	432.4	<.001	10.01	395.2	<.001	9.60	395.3	<.001
TR	10 - 14	100ms	300ms	9.84	431.8	<.001	11.80	386.5	<.001	11.31	386.5	<.001
TR	11 - 12	200ms	200+50ms	1.81	432.4	0.07	-0.10	385.9	0.92	-0.09	385.8	0.93
TR	11 - 13	200ms	200+100ms	3.07	432.4	<.001	2.82	385.8	0.01	2.69	385.8	0.01
TR	11 - 14	200ms	300ms	3.31	431.8	<.001	4.22	393.0	<.001	4.04	393.1	<.001
TR	12 - 13	200+50ms	200+100ms	1.24	431.8	**0.21**	2.97	386.8	<.001	2.84	386.7	<.001
TR	12 - 14	200+50ms	300ms	1.46	432.4	**0.14**	4.42	390.1	<.001	4.23	390.1	<.001
TR	13 - 14	200+100ms	300ms	0.21	432.4	0.84	1.37	390.1	0.17	1.31	390.1	0.19
SR	15 - 16	60fps	30fps	6.19	432.5	<.001	2.82	402.9	0.01	6.05	403.0	<.001
SR	15 - 17	60fps	10fps	11.44	434.0	<.001	13.09	396.2	<.001	15.95	396.3	<.001
SR	16 - 17	30fps	10fps	5.55	433.1	<.001	10.55	387.9	<.001	10.11	387.8	<.001
TR	18 - 19	60fps	30fps	4.35	431.8	<.001	2.18	412.6	0.03	5.54	412.8	<.001
TR	18 - 20	60fps	10fps	13.26	433.1	<.001	7.46	444.7	<.001	10.34	445.0	<.001
TR	19 - 20	30fps	10fps	9.05	433.1	<.001	6.07	404.4	<.001	5.80	404.5	<.001

Table A.5: Pair-comparison statistics of input quality per condition between Dataset 4.3 (CS) and Dataset 4.3 (Lab) for test environment comparison between lab studies and CS tests. The p-values highlighted in bold indicate different findings for the corresponding pair-comparison.

Game	CID	Conditions		Laboratory IPQ			Crowdsourcing IPQ			Crowdsourcing IPQ (mapped)		
	(A - B)	A	B	t	df	p	t	df	p	t	df	p
RE	1 - 2	0ms	300ms	8.57	438.1	<.001	5.51	394.7	<.001	5.44	394.4	<.001
SR	3 - 4	0ms	100ms	10.03	438.7	<.001	6.38	401.8	<.001	8.96	401.6	<.001
SR	3 - 5	0ms	200ms	13.41	439.3	<.001	10.68	401.8	<.001	13.20	401.6	<.001
SR	3 - 6	0ms	200+50ms	13.52	439.3	<.001	10.36	400.7	<.001	12.91	400.4	<.001
SR	3 - 7	0ms	200+100ms	15.26	440.0	<.001	12.54	405.6	<.001	14.96	405.4	<.001
SR	3 - 8	0ms	300ms	14.44	439.7	<.001	12.26	403.0	<.001	14.74	402.8	<.001
SR	4 - 5	100ms	200ms	3.56	438.7	<.001	4.54	396.5	<.001	4.48	396.3	<.001
SR	4 - 6	100ms	200+50ms	3.67	438.7	<.001	4.14	395.6	<.001	4.09	395.3	<.001
SR	4 - 7	100ms	200+100ms	5.67	439.3	<.001	6.70	399.6	<.001	6.62	399.4	<.001
SR	4 - 8	100ms	300ms	4.71	439.0	<.001	6.28	397.5	<.001	6.20	397.3	<.001
SR	5 - 6	200ms	200+50ms	0.11	438.1	0.91	-0.44	395.6	0.66	-0.43	395.3	0.66
SR	5 - 7	200ms	200+100ms	2.16	438.7	0.03	2.31	399.6	0.02	2.28	399.4	0.02
SR	5 - 8	200ms	300ms	1.18	438.4	0.24	1.79	397.5	0.08	1.76	397.3	0.08
SR	6 - 7	200+50ms	200+100ms	2.06	438.7	0.04	2.76	398.5	0.01	2.72	398.3	0.01
SR	6 - 8	200+50ms	300ms	1.07	438.4	**0.28**	2.24	396.5	0.03	2.21	396.3	0.03
SR	7 - 8	200+100ms	300ms	-0.99	438.4	0.32	-0.56	400.7	0.58	-0.55	400.5	0.58
TR	9 - 10	0ms	100ms	3.55	438.1	<.001	1.35	399.9	**0.18**	4.32	399.7	<.001
TR	9 - 11	0ms	200ms	10.34	438.1	<.001	8.61	402.3	<.001	11.40	402.1	<.001
TR	9 - 12	0ms	200+50ms	11.87	438.7	<.001	9.26	403.1	<.001	12.09	402.9	<.001
TR	9 - 13	0ms	200+100ms	13.17	438.7	<.001	11.10	399.8	<.001	13.92	399.5	<.001
TR	9 - 14	0ms	300ms	14.21	438.1	<.001	12.44	396.3	<.001	15.35	396.0	<.001
TR	10 - 11	100ms	200ms	6.79	438.1	<.001	7.17	404.9	<.001	7.08	404.7	<.001
TR	10 - 12	100ms	200+50ms	8.35	438.7	<.001	7.81	402.1	<.001	7.70	401.9	<.001
TR	10 - 13	100ms	200+100ms	9.65	438.7	<.001	9.56	405.6	<.001	9.42	405.4	<.001
TR	10 - 14	100ms	300ms	10.66	438.1	<.001	10.85	398.1	<.001	10.69	397.8	<.001
TR	11 - 12	200ms	200+50ms	1.64	438.7	0.10	0.46	397.5	0.64	0.46	397.2	0.65
TR	11 - 13	200ms	200+100ms	2.94	438.7	<.001	2.20	397.5	0.03	2.17	397.2	0.03
TR	11 - 14	200ms	300ms	3.87	438.1	<.001	3.13	403.6	<.001	3.08	403.4	<.001
TR	12 - 13	200+50ms	200+100ms	1.29	438.1	0.20	1.77	398.3	0.08	1.74	398.1	0.08
TR	12 - 14	200+50ms	300ms	2.20	438.7	0.03	2.71	401.2	0.01	2.67	401.0	0.01
TR	13 - 14	200+100ms	300ms	0.89	438.7	0.37	0.89	401.2	0.37	0.88	401.0	0.38
SR	15 - 16	60fps	30fps	3.49	438.7	<.001	1.25	412.9	**0.21**	3.98	412.7	<.001
SR	15 - 17	60fps	10fps	7.98	440.0	<.001	11.45	406.9	<.001	14.08	406.7	<.001
SR	16 - 17	30fps	10fps	4.70	439.3	<.001	10.51	399.4	<.001	10.36	399.2	<.001
TR	18 - 19	60fps	30fps	2.86	438.1	<.001	1.07	421.6	**0.29**	3.67	421.5	<.001
TR	18 - 20	60fps	10fps	9.10	439.2	<.001	5.36	450.2	<.001	7.70	450.2	<.001
TR	19 - 20	30fps	10fps	6.33	439.2	<.001	4.89	414.3	<.001	4.81	414.2	<.001

Appendix B

Measurement Instruments Used to Assess Gaming QoE

In the following, the items included in the complete measurement method presented in Section 3.3.2 are summarized.

Pre-test Questionnaire:

1. What is your Year of Birth?

2. What is your gender?

a. Female
b. Male
c. Transgender
d. Prefer not to say

3. Roughly how many hours per week do you spend on playing video games?

a. Between 0 to 1 hours
b. Between 1 to 5 hours
c. Between 5 to 10 hours
d. Between 10 to 20 hours
e. More than 20 hours

4. Roughly how often do you play video games in a week?

a. Never
b. Between 1 to 3 times a week
c. Between 3 to 7 times a week
d. Between 7 to 14 times a week
e. More than 14 times week

5. How would you describe your gaming experience (expertise)?

a. 1 – Beginner
b. 2
c. 3 – Intermediate
d. 4
e. 5 – Expert

6. I like playing video games.

a. 1 - Strongly Disagree
b. 2 – Disagree
c. 3 – Undecided
d. 4 – Agree
e. 5 - Strongly Agree

7. On which kind of device do you usually play games?

a. PC (Desktop)
b. Smartphone / Tablet
c. Console (PlayStation, XBox, ...)
d. Others

8. What kind of monitor are you typically using when playing?

a. Television (> 30”)
b. Desktop Monitor (> 20”)
c. Laptop (> 12”)
d. Tablet (> 8”)
e. Large Smartphone (> 5”)
f. Small Smartphone (< 5”)
g. Other

9. How experienced are you in playing the game "[game name]"?

a. 1 – Unexperienced
b. 2
c. 3 – Intermediate
d. 4
e. 5 - Expert

Post-game Questionnaire:

The post-game questionnaire covers the following aspects: performance indication (PI), learnability (LE), appeal (AP), and intuitive controls (IC). Component scores are computed as the average value of its items. The used items are summarized in Table B.1 whereas an example of the used 7-point EC-ACR scale for all items is shown in Fig. B.1.

Table B.1: Overview of items used in the post-game questionnaire

Order	Item Text	Item ID
1	I could easily assess how I was performing in the game.	PI1
2	Learning to operate the game is easy for me.	LE1
3	I liked the graphics and images used in the game.	AP1
4	Learning the game controls was easy.	IC1
5	It was clear to me how my performance was going.	PI2
6	It is easy for me to become skillful at using the game.	LE2
7	The game appealed to my visual senses.	AP2
8	The game controls are intuitive.	IC2
9	I was informed about my progress in the game.	PI3
10	I find the game easy to use.	LE3
11	The game was aesthetically appealing.	AP3
12	It was easy to remember the corresponding control.	IC3

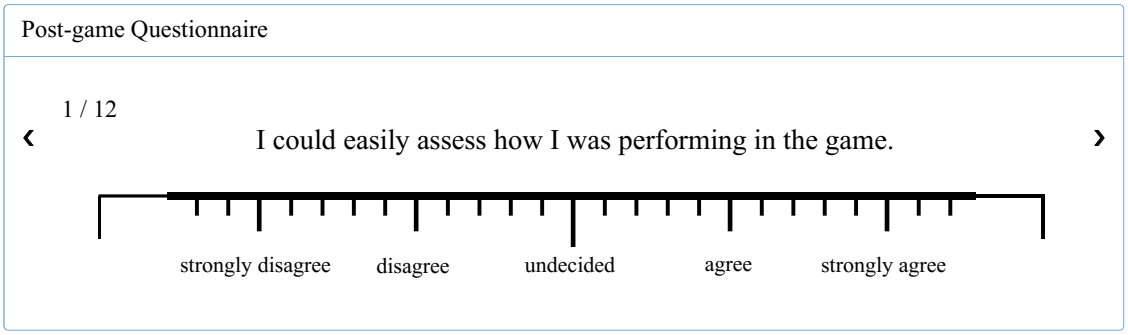


Figure B.1: Example of item and scales used in the post-game questionnaire.

Post-test Questionnaire:

For the post-test questionnaire, the following instructions are given to participants: “In the following, we would like you to tell us about your judgement criteria. Please indicate on the scales below, how important in general (not just for this study) the listed aspects are for your rating of the overall quality of your gaming experience.”

An example of the used 7-point EC-ACR scale for all items is shown in Fig. B.2. The bold written aspect (in the example controllability) was exchanged with: video quality, audio quality, controllability, responsiveness, immediate feedback, video fragmentation, video unclearness, video discontinuity, suboptimal video luminosity, and playing performance. The order of items was randomized and an open question about potential other aspects was added.

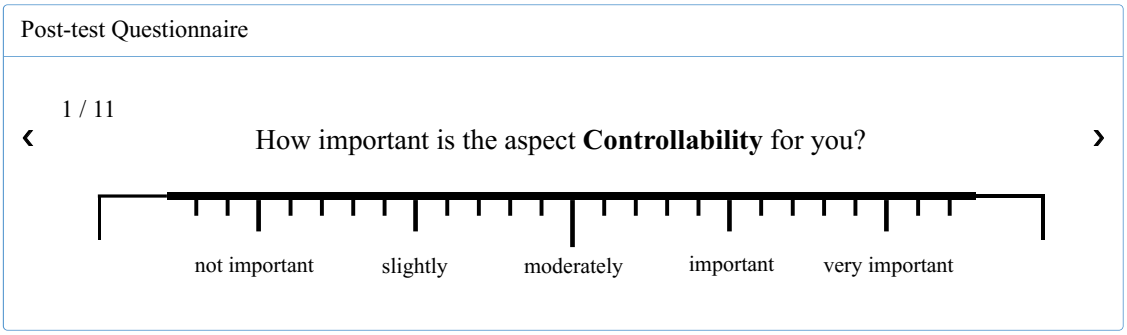


Figure B.2: Example of item and scales used in the post-test questionnaire.

Post-condition Questionnaire:

Participants are asked to indicate how they felt while playing the game for each of the following items by clicking on the 7-point scale. The questionnaire covers the following aspects:

- Input quality: controllability (CN), responsiveness (RE), immediate feedback (IF)
- Output quality: audio quality (AQ), video quality (VQ), video fragmentation (VF), video unclearness (VU), video discontinuity (VD), suboptimal video luminosity (VL)
- Player Experience: immersion (IM), competency (CO), negative affect (NA), flow (FL), tension(TE), positive affect (PA), challenge (CH)
- Self-judgement of playing performance (PR), and service acceptability (AC)

The component scores of each aspect is computed as the average value of its items. The full list of items is summarized in Table B.2 whereas the corresponding rating scales are shown in Fig. B.3, Fig. B.4, and Fig. B.5.

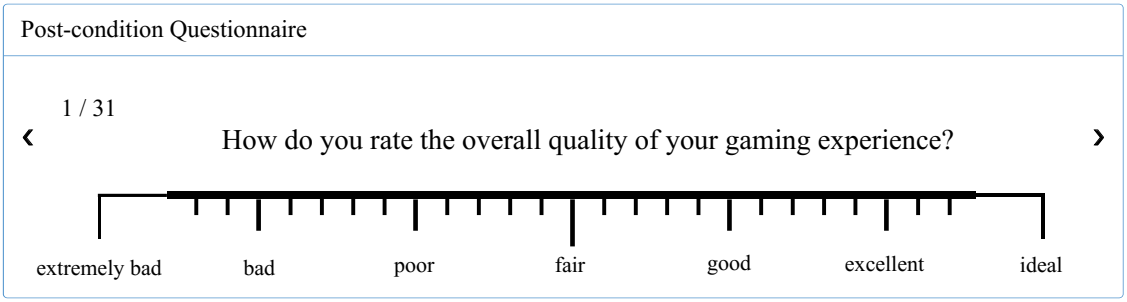


Figure B.3: Example of the first rating scale type used in the post-condition questionnaire.

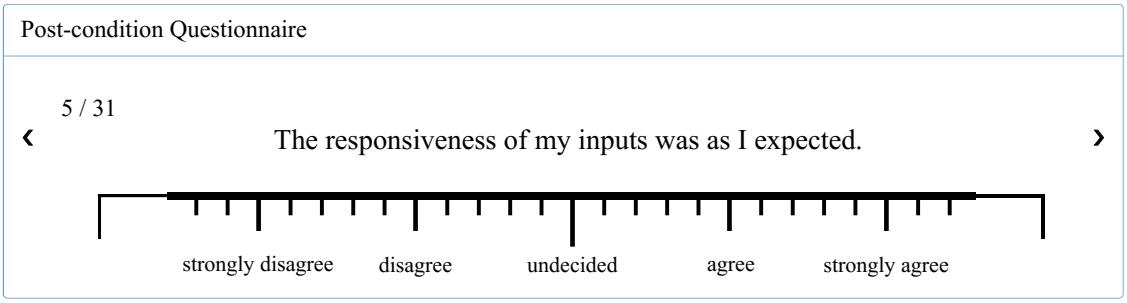


Figure B.4: Example of the second rating scale type used in the post-condition questionnaire.

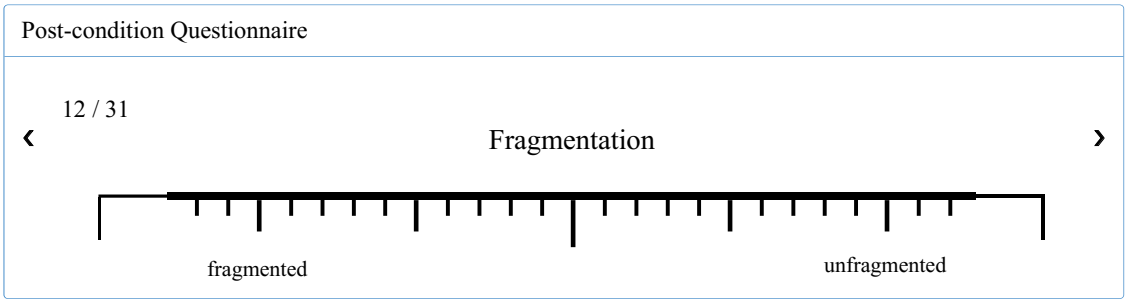


Figure B.5: Example of the third rating scale type used in the post-condition questionnaire. The 7-point continuous bipolar scale was used which was attached with the following antonym pairs: fragmented and unfragmented (VF), unclear and clear (VU), discontinuous and continuous (VD), suboptimal and optimal (VL), no and yes (AC).

Table B.2: Overview of items used in the post-condition questionnaire

Order	Item Text	Item ID	Scale
1	How do you rate the overall quality of your gaming experience?	QOE	1
2	I felt that I had control over my interaction with the system.	CN1	2
3	I noticed a delay between my actions and the outcomes.	RE1	2
4	I felt a sense of control over the game interface and input devices.	CN2	2
5	The responsiveness of my inputs was as I expected.	RE2	2
6	I felt in control of my game actions.	CN3	2
7	I received immediate feedback on my actions.	IF1	2
8	My inputs were applied smoothly.	RE3	2
9	I was notified about my actions immediately.	IF2	2
10	How do you rate the overall audio quality?	AQ	1
11	How do you rate the overall video quality?	VQ	1
12	Fragmentation	VF	3
13	Unclearness	VU	3
14	Discontinuity	VD	3
15	Suboptimal Luminosity	VL	3
16	I found it impressive.	IM1	2
17	I felt successful.	CO1	2
18	I felt bored.	NE1	2
19	It felt like a rich experience.	IM2	2
20	I forgot everything around me.	FL1	2
21	I felt frustrated.	TE1	2
22	I found it tiresome.	NE2	2
23	I felt irritable.	TE2	2
24	I felt skillful.	CO2	2
25	I felt completely absorbed.	FL2	2
26	I felt content.	PO1	2
27	I felt challenged.	CH1	2
28	I had to put a lot of effort into it.	CH2	2
29	I felt good.	PO2	2
30	How do you rate your own playing performance?	PR	1
31	Would you accept using a service under these conditions?	AC	3

Appendix C

Additional Material Related to the GIPS

Table C.1: List of initial items for GIPS. The final items are highlighted in bold. The index *i* indicates an inverted item. In addition, the position of the item in the used questionnaire (No.) as well as the initial source of the items (Ref) is provided. An asterisk attached to the source indicates an adaptation of the item by experts.

Code	Label	No.	Ref
Controllability			
CN1	I felt that I had control over my interaction with the system.	3	[115]
CN2	I felt a sense of control over the game interface and input devices.	6	[165],[242]*
CN3	I felt in control of my game actions.	9	[163]*
CN4	I thought controlling the game was easy.	12	self
CN5	I was able to control the course of events in the game.	15	self
CN6	The physical controls (keys, buttons, pads, etc.) were appropriate for the game.	18	[243]*
CN7	I was able to perform my intended interactions.	21	self
CN8	I was able to control the game as desired.	24	self
CN9	The input device (mouse, touch, gamepad, etc.) was precise enough to control the game.	26	[243]*
Responsiveness (later also Immediate Feedback)			
RE1	Overall, I was satisfied with the responsiveness of the game.	1	self
RE2$_i$	I noticed delay between my actions and the outcomes.	4	[243]*
RE3	The responsiveness of my inputs was as I expected.	7	[163]*
RE4	I received immediate feedback on my actions.	10	[244],[165]
RE5	My inputs were applied smoothly.	13	self
RE6	I was notified about my actions immediately.	16	[165]*
RE7	The game responded as expected to my inputs.	19	self
RE8	My inputs were notified by the system.	22	self
Performance Indication			
PI1	I could easily assess how I was performing in the game.	2	[149]
PI2	I had a good idea about the status of the game.	5	[149]*
PI3	I was aware of how well I was performing in the game.	8	[244]
PI4	It was clear to me how my performance was going.	11	[168]*
PI5	I was informed about my progress in the game.	14	[149]*
PI6	I was notified once reaching a goal of the game.	17	[165]*
PI7	All necessary feedback information was visible.	20	self
PI8	I was able to see the results of my interaction with the game.	23	self
PI9	The feedback from the game was as I expected.	25	self

Table C.2: Inter-item correlation matrix for the latent factor controllability.

	CN1	CN2	CN3	CN4	CN5	CN6	CN7	CN8	CN9
CN1	1.00	0.93	0.92	0.81	0.82	0.49	0.82	0.87	0.80
CN2	0.93	1.00	0.94	0.82	0.83	0.52	0.84	0.89	0.82
CN3	0.92	0.94	1.00	0.84	0.83	0.53	0.85	0.91	0.85
CN4	0.81	0.82	0.84	1.00	0.71	0.57	0.78	0.83	0.78
CN5	0.82	0.83	0.83	0.71	1.00	0.45	0.78	0.84	0.75
CN6	0.49	0.52	0.53	0.57	0.45	1.00	0.57	0.51	0.65
CN7	0.82	0.84	0.85	0.78	0.78	0.57	1.00	0.85	0.81
CN8	0.87	0.89	0.91	0.83	0.84	0.51	0.85	1.00	0.83
CN9	0.80	0.82	0.85	0.78	0.75	0.65	0.81	0.83	1.00

Table C.3: Inter-item correlation matrix for the latent factor responsivness.

	RE1	RE2	RE3	RE4	RE5	RE6	RE7	RE8
RE1	1.00	0.76	0.90	0.56	0.89	0.42	0.90	0.54
RE2	0.76	1.00	0.77	0.44	0.79	0.36	0.78	0.43
RE3	0.90	0.77	1.00	0.55	0.91	0.47	0.94	0.59
RE4	0.56	0.44	0.55	1.00	0.52	0.81	0.56	0.53
RE5	0.89	0.79	0.91	0.52	1.00	0.43	0.93	0.60
RE6	0.42	0.36	0.47	0.81	0.43	1.00	0.45	0.46
RE7	0.90	0.78	0.94	0.56	0.93	0.45	1.00	0.63
RE8	0.54	0.43	0.59	0.53	0.60	0.46	0.63	1.00

Table C.4: Inter-item correlation matrix for the latent factor performance indication.

	PI1	PI2	PI3	PI4	PI5	PI6	PI7	PI8	PI9
PI1	1.00	0.85	0.87	0.88	0.77	0.51	0.76	0.63	0.63
PI2	0.85	1.00	0.86	0.86	0.75	0.51	0.75	0.56	0.60
PI3	0.87	0.86	1.00	0.93	0.81	0.48	0.78	0.65	0.61
PI4	0.88	0.86	0.93	1.00	0.84	0.52	0.81	0.67	0.63
PI5	0.77	0.75	0.81	0.84	1.00	0.57	0.85	0.71	0.66
PI6	0.51	0.51	0.48	0.52	0.57	1.00	0.58	0.52	0.44
PI7	0.76	0.75	0.78	0.81	0.85	0.58	1.00	0.70	0.73
PI8	0.63	0.56	0.65	0.67	0.71	0.52	0.70	1.00	0.60
PI9	0.63	0.60	0.61	0.63	0.66	0.44	0.73	0.60	1.00

Table C.5: Item-total statistics for controllability, responsiveness (one-factor and two-factor solution), and performance indication during GIPS development process.

Controllability		Responsiveness (one-factor)		Responsiveness (two-factor)		Performance Indication	
Item	Cronbach's α if item deleted	Item	Cronbach's α if item deleted	Item	Cronbach's α if item deleted	Item	Cronbach's α if item deleted
CN1	0.96	RE1	0.92	RE1	0.96	PI1	0.94
CN2	0.96	RE2	0.93	RE2	0.98	PI2	0.94
CN3	0.96	RE3	0.92	RE3	0.96	PI3	0.94
CN4	0.97	RE4	0.94	RE5	0.95	PI4	0.94
CN5	0.97	RE5	0.92	RE7	0.95	PI5	0.94
CN6	0.98	RE6	0.94			PI6	0.96
CN7	0.96	RE7	0.92			PI7	0.94
CN8	0.96	RE8	0.94			PI8	0.95
CN9	0.96					PI9	0.95

Appendix D

Information about Used Games in Research

Table D.1: Games used in various studies and datasets including their genres and content classification according to Section 6.2.

Game Name	Genre or Type	Study or Dataset	Delay Sensitivity	Frameloss Sensitivity	Encoding Complexity
Dodge	dexterity	Study 4.1, 4.2	High	-	-
GTA (web-version)	action/shooting	Study 4.1, 4.2	High	-	-
Shooting Range	action/shooting	Study 4.1-4.6	High	-	-
Flappy Bird	arcade	Study 4.1, 4.2	High	-	-
Rocket Escape	racing	Study 4.1, 4.2	Low	-	-
T-Rex	jump and run	Study 4.1-4.6	High	-	-
Project Cars	racing	Study 3.1, Study 6.1, G.1072 (interactive)	Low	High	Medium
GTA 5	action/shooting	Study 3.1, Study 6.1, G.1072 (interactive)	High	High	High
Rayman Legends	jump and run	Study 6.1, G.1072 (interactive)	High	High	Low
Bejeweled 3	puzzle	G.1072 (interactive)	Low	Low	Low
Counter Strike: GO	action/shooting	G.1072 (interactive)	High	High	High
Dota2	MOBA	G.1072 (interactive)	High	High	Medium
Hearthstone	card playing	G.1072 (interactive)	Low	Low	Low
Overwatch	action/shooting	G.1072 (interactive)	High	High	High
Tekken 7	fighting	G.1072 (interactive)	Low	High	High
Worms W.M.D.	turn-based strategy	G.1072 (interactive)	Low	Low	Low
League of Legends	MOBA	G.1072 (passive)	High	High	High
Nier: Automata	action role-playing	G.1072 (passive)	High	High	High
Stick Fight	arcade fighting	G.1072 (passive)	Low	Low	Low
Apex Legends	action/shooting	G.1072 (passive)	High	High	High
Black Desert	MMORPG	G.1072 (passive)	High	High	High
Dauntless	action role-playing	G.1072 (passive)	High	High	High
Fifa 2020	sport	G.1072 (passive)	Low	Low	Medium
Final Fantasy XV	action role-playing	G.1072 (passive)	High	High	High
Fortnite	action/shooting	G.1072 (passive)	High	High	High
MapleStory 2	fighting RPG	G.1072 (passive)	High	High	Medium
Minecraft	open-world sandbox	G.1072 (passive)	High	High	High

Table D.2: Games used in various studies and datasets including a description of the player's task.

Game Name	Description
Dodge	The player has to dodge colored boxes that spawn at different positions and try to hit the character
GTA (web-version)	The player has to shoot precisely on round targets as often as possible.
Shooting Range	The player has to shoot on moving targets (from right to left) three times to avoid that the targets reach the left border of the screen.
Flappy Bird	The player has to balance a flying bird (up and down movement) through a world of differently positioned pillars.
Rocket Escape	The player has to navigate an abstract car through a randomly changing road.
T-Rex	The player has to jump over boxes without touching them.
Project Cars	The player has to navigate a car on a highly realistic racing track.
GTA 5	The player has to shoot precisely on round targets as often as possible.
Rayman Legends	The player has to jump over different obstacles until reaching the end of an obstacle course.
Bejeweled 3	The player has to select a move (without time limit) gems to create groups of at least 3 in a row (similar to the game candy crush).
Counter Strike: GO	The player has to move through a 3D world and shoot at other virtual players.
Dota2	The player has to hit monsters by clicking (causing an attack) in a timely precise manner and use skills to kill opponents.
Hearthstone	The player has to select cards to set up a defense and offense strategy in a turn-based match.
Overwatch	The player has to move through a 3D world and shoot at other virtual players.
Tekken 7	The player has to move and perform combinations of punches and hits to beat an opponent.
Worms W.M.D.	The player has to shoot with a weapon on targets in a mostly static world (direction and power of shot must be controlled)
League of Legends	The player has to hit monsters by clicking (causing an attack) in a timely precise manner and use skills to kill opponents.
Nier: Automata	The player has to combine attacks like sword strokes to beat opponents while dodging their attacks.
Stick Fight	The player has to shoot at enemies while jumping on various platforms.
Apex Legends	The player has to move through a 3D world and shoot at other virtual players.
Black Desert	The player has to combine attacks like sword strokes to beat opponents while dodging their attacks.
Dauntless	The player has to combine attacks like sword strokes to beat opponents while dodging their attacks.
Fifa 2020	The player has to control a team using movements, dodges, passes, and shots to score a goal (typical soccer game).
Final Fantasy XV	The player has to combine attacks like sword strokes to beat opponents while dodging their attacks.
Fortnite	The player has to move through a 3D world and shoot at other virtual players.
MapleStory 2	The player has to combine attacks like sword strokes to beat opponents while dodging their attacks.
Minecraft	The player has to move in an abstract world a build builds or objects by combining collected materials.