I n t e r n a t i o n a l   T e l e c o m m u n i c a t i o n   U n i o n

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

P.809
(06/2018)

SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Methods for objective and subjective assessment of speech and video quality

---

Subjective evaluation methods for gaming quality

Recommendation ITU-T P.809

ITU-T

ITU-T P-SERIES RECOMMENDATIONS

TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS

Vocabulary and effects of transmission parameters on customer opinion of transmission quality	Series	P.10
Voice terminal characteristics	Series	P.30
		P.300
Reference systems	Series	P.40
Objective measuring apparatus	Series	P.50
		P.500
Objective electro-acoustical measurements	Series	P.60
Measurements related to speech loudness	Series	P.70
Methods for objective and subjective assessment of speech quality	Series	P.80
Methods for objective and subjective assessment of speech and video quality	**Series**	**P.800**
Audiovisual quality in multimedia services	Series	P.900
Transmission performance and QoS aspects of IP end-points	Series	P.1000
Communications involving vehicles	Series	P.1100
Models and tools for quality assessment of streamed media	Series	P.1200
Telemeeting assessment	Series	P.1300
Statistical analysis, evaluation and reporting guidelines of quality measurements	Series	P.1400
Methods for objective and subjective assessment of quality of services other than speech and video	Series	P.1500

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T P.809

Subjective evaluation methods for gaming quality

Summary

Recommendation ITU-T P.809 presents guidelines for conducting subjective experiments for the quality of experience (QoE) assessment of gaming services. First, an overview about gaming QoE features covering hedonic and pragmatic quality as well as player experience is given. Additionally, methods for two test paradigms, passive viewing-and-listening tests and interactive tests, are described including information about the test environment and test set-up, participant instructions, selection of game materials and a list of available questionnaires for the assessment of several gaming QoE features such as flow and immersion.

History

Edition	Recommendation	Approval	Study Group	Unique ID*
1.0	ITU-T P.809	2018-06-13	12	11.1002/1000/13626

Keywords

Evaluation, gaming, quality, questionnaire, test design.

---

* To access the Recommendation, type the URL http://handle.itu.int/ in the address field of your web browser, followed by the Recommendation's unique ID. For example, http://handle.itu.int/11.1002/1000/11830-en.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure, e.g., interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at http://www.itu.int/ITU-T/ipr/.

Table of Contents

Page

Introduction

In addition to voice, video and web browsing, increasingly popular services running on top of IP-based networks are online computer games. However, in comparison to the aforementioned services, less information exists about subjective evaluation methods for assessing the quality of experience (QoE) of gaming services. To ensure the validity, reliability, and objectivity of results, standardized methods for the assessment of subjective ratings are highly important. This allows the comparison of results of different studies and forms a basis for building instrumental quality prediction models for gaming services.

Recommendation ITU-T P.809

Subjective evaluation methods for gaming quality

1 Scope

This Recommendation describes subjective evaluation methods providing information about the quality of gaming services, as experienced by users of such services. Since there is currently no standardized method available for the evaluation of gaming quality, this Recommendation should be seen as state of the art in order to help choose suitable methods to conduct subjective user tests. The assessment of gaming quality of experience (QoE) is important when planning and implementing online gaming services as well as for the development of instrumental quality prediction models.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

[ITU-T P.851] Recommendation ITU-T P.851 (2003), *Subjective quality evaluation of telephone services based on spoken dialogue systems*.

[ITU-T P.880] Recommendation ITU-T P.880 (2004), *Continuous evaluation of time-varying speech quality*.

[ITU-T P.910] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.

[ITU-T P.911] Recommendation ITU-T P.911 (1998), *Subjective audiovisual quality assessment methods for multimedia applications*.

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 cloud gaming [b-ITU-T G.1032]: 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 are performed at the server, while the client is responsible for video decoding and capturing of client input.

3.1.2 eSport [b-ITU-T G.1032]: A form of sports where the primary aspects of the sport are facilitated by electronic systems; the input of players and teams as well as the output of the eSports system are mediated by human-computer interfaces.

3.1.3 game bricks [b-ITU-T G.1032]: A rule-based game classification splitting the games into several fundamental elements such as moving and shooting. In total, ten "gameplay bricks" in two categories are proposed by [b-Djaouti], rules stating goals, including avoid, match, and destroy, and rules defining the means and constraints to reach the goals consisting of create, manage, move, random, select, shoot, and write.

3.1.4 genre [b-ITU-T G.1032]: A classification of games where games are grouped according to their gameplay characteristics.

3.1.5 online gaming [b-ITU-T G.1032]: A service that enables a video game to be either partially or primarily played over a broadband network. The service renders the game at the client device while the updated states of game are transferred over a broadband network.

3.2 Terms defined in this Recommendation

This Recommendation defines the following term:

3.2.1 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 [b-Juul].

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR	Absolute Category Rating
EEG	Electroencephalography
GEQ	Game Experience Questionnaire
GPU	Graphics Processing Unit
HD	High Definition
MMORPG	Massively Multiplayer Online Role-Playing Game
QoE	Quality of Experience
SI	Spatial Perceptual Information
TI	Temporal Perceptual Information
UX	User Experience
VE	Virtual Environment

5 Conventions

In this Recommendation, online gaming is referred to as a service that renders the game on the client device while the updated states of a game are transferred over a broadband network. In contrast, in cloud gaming there is no processing (execution of the game logic, rendering of the 3D virtual scene) on the client. Within the scope of this Recommendation, the person interacting with a game is referred to as the player, whereas the software, which is used in cloud gaming set-ups to display a remotely rendered game video stream, is referred to as the client.

6 Quality of experience aspects for gaming

As with previous studies about the quality experience in video games, gaming QoE is a multi-dimensional construct involving the quality features illustrated in Figure 1 (see also [b-Möller]), which are explained in the following clauses.

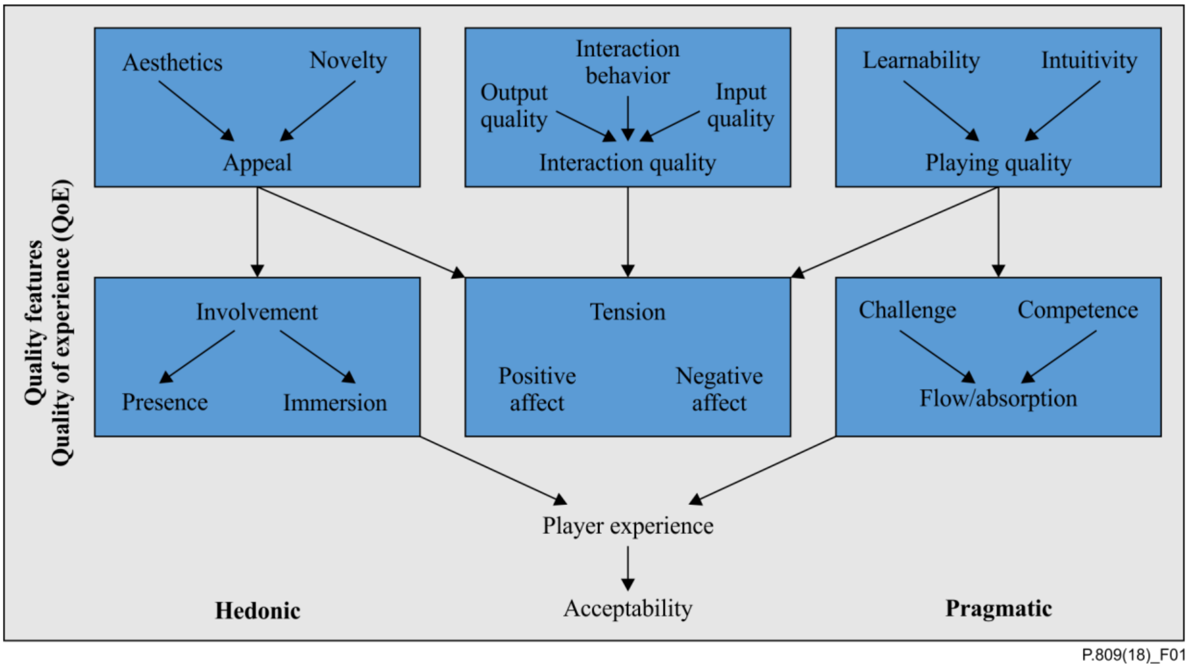

Figure 1 – Quality features of gaming QoE

6.1 Aesthetics and appeal

Aesthetics are the sensory experience that the system elicits, and the extent to which this experience fits an individual's goals and spirit, see [b-Vilnai]. The system's personality refers to 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, see [b-Stelmaszewska], [b-Hassenzahl].

6.2 Interaction quality

Interaction quality refers to the playability of the game in terms of the degree to which all functional and structural elements of the game (hardware and software) provide a positive player experience. This definition considers playability as a prerequisite of positive player experience (similar to usability being considered a prerequisite of user satisfaction), or as a technical and structural basis for this, but not the player experience itself. This interpretation of interaction quality agrees with the general definition of this term for multimodal systems and includes input quality (player to system, output quality (system to player, e.g., in terms of graphics quality, video quality, sound quality), as well as the interactive behavior (in task-oriented interaction this is called "cooperativity", but as a game storyline is not designed to be cooperative to the user, the general term interactive behavior is more appropriate here).

6.3 Playing quality

Playing quality can be considered as a kind of game usability. This is defined by Pinelle et al. [b-Pinelle] as "The degree to which a player is able to learn, intuitively 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)".

6.4 Engagement

The following clauses present an overview of several concepts describing engaging experiences while playing games. These concepts are: involvement, immersion, presence, flow and absorption.

6.4.1 Involvement

Involvement is a psychological state experienced as a consequence of focusing one's mental energy and attention on a coherent set of stimuli or meaningfully related activities or events. Involvement is increased by performing tasks and participating in activities that stimulate, challenge, and engage the user either cognitively, physically, or emotionally [b-Witmer].

6.4.2 Immersion

According to the current state of the art, there are multiple, commonly used definitions for *Immersion.* Witmer and Singer defined immersion as "A psychological state characterized by perceiving oneself to be enveloped by, included in, and interacting with an environment that provides a continuous stream of stimuli and experiences." Next to the natural modes of interaction and control, and perception of self-movement, the authors state that the degree of immersion can be influence by the isolation of a user from their physical environment. Head-mounted displays can provide this isolation to a high level whereas an arcade-style video game may be able to lead to a high level of involvement but not immersion [b-Singer]. In contrast to the previous viewpoint, defining immersion to be an individual experience, Slater et al. propose that "The degree of immersion can be objectively assessed as the characteristics of a technology, and has dimensions such as the extent to which a display system can deliver an inclusive, extensive, surrounding and vivid illusion of virtual environment to a participant." [b-Slater]. The authors state, that immersion occurs when the sense of physical reality is shut out, a high resolution and quality of the displays is available and the virtual reality is panoramic rather than limited to a narrow field [b-Slater].

Specially with a focus on games, immersion is used to describe the degree of involvement with a computer game and has been classified into three phases as: "engagement", "engrossment", and "total immersion" [b-Brown], [b-Jennett]. Brown and Cairns state that several barriers can limit the degree of involvement. To enter the level of *engagement*, the gamer has to overcome the barrier of gamer preference, 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 gamers described a sense of presence and of being cut off from reality to such an extent that the game was all that mattered [b-Brown], [b-Jennett].

Björk and Holopainen categorized immersion into four categories: tactical immersion, strategic immersion, narrative immersion and spatial immersion. While the first three are related to controlling the game and becoming interested in the story, spatial immersion describes the same concept as total immersion or presence [b-Bjork]. While indeed the used technology, especially the screen size as part of the definition by Slater, can have an impact on immersion, which is considered as an experience as defined by Witmer and Singer.

6.4.3 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. It also depends on the user's previous experiences and current state [b-Witmer].

6.4.4 Flow

Flow is considered [b-Csik] to be an equilibrium between boredom and fear, between requirements and abilities; it is a dynamic experience of complete dissolution of an acting person in his/her activity. 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 [b-Chen]. Hassenzahl relates flow to user experience (UX): "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 user experience derived from fulfilling the need for competence (i.e., mastery); it is a particular experience stemming from the fulfilment of a particular be-goal.", [b-Hassenzahl-1]. In general, everybody can experience flow, but there seem to be factors, which reduce flow in games like age, reaction time, abilities, and exposure to computers (digital natives vs. newbies) [b-Hugentobler].

6.4.5 Absorption

Cognitive *absorption* 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 [b-Agarwal]. "Being absorbed refers to being in a state of deep attention with the event experienced" [b-Weniger]. The notion of cognitive engagement can be described by the three dimensions 'attention focus', 'curiosity', and 'interest' [b-Webster]. Webster and Ho argue that absorption is "identical to flow, just without the dimension of control." They say that “individual control is not necessary for cognitive engagement, because 'passive engagement' (e.g., watching TV) might exist while 'passive flow' is impossible" [b-Weniger], [b-Webster].

6.4.6 Relationships between engagement concepts

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 [b-Jennett-1]. Jennett further argues that "Immersion is 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.", [b-Jennett-1]. 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" [b-Agarwal], [b-Jennett-1]. Brockmyer et al. summarize that a "continuum of deepening engagement from presence to flow to absorption" may exist for some individual experiences.

6.5 Positive and negative effect

Positive effect can come in many different forms and it is usually the goal of all gaming activity. Fun has been defined as "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, triumphalism, control, and mastery of material." [b-Murphy]. Negative effects might be frustration and boredom. Applied to computer games, Lazzaro [b-Lazzaro] 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 triumphalism), 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 [b-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 [b-Schaffer].

NOTE – The "achiever", "explorer", socializer" and "killer" user types are defined in [b-Bartle] in the context of the multi-user dungeon game genre.

6.6 Player experience

Based on the definition of quality of experience, player experience describes the degree of delight or annoyance perceived by the player after the gaming experience. It contains the aspects of tension, immersion, positive and negative effects, challenge, competence and flow.

6.7 Acceptability

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 quantity of the target group. Acceptability is influenced by player experience, but also by other factors such as costs, accessibility, service conditions, etc.

7 Test paradigms

As is the case with all evaluation methods, the chosen method should reflect the later use scenario as closely as possible to reach ecological validity, i.e., that the test method measures what it is expected to measure. 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 with 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 which 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 in 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 viewing-and-listening paradigm.

In the following clauses, two test paradigms will be described in more detail:

1) passive viewing-and-listening tests with audio-visual stimuli (clause 9);

2) interactive tests with game scenes (clause 10).

Before describing the specifics of each test paradigm, a general introduction about the experimental set-up which is valid for both test paradigms will be given in clause 8.

These two test paradigms are not the only ones which are conceivable. For example, a test set-up with long-gaming stimuli which are to be evaluated continuously in a passive viewing-and-listening paradigm, similar to what is described in [ITU-T P.880] for speech stimuli. As experience with other test paradigms is still lacking, ITU-T currently only provides Recommendations for the two mentioned test paradigms.

In Table 1, an overview of use cases for each test paradigm is given. While it is not recommended to perform certain tests, e.g., a short passive test with the aim to measure flow, there currently is no proof that this would not be possible. It is recommended to use short passive viewing-and-listening tests if the aim of the study is to assess the output quality (video and audio quality). An exactly comparable content as well as a reduction of user influencing factors are the advantage. The content must be representative for the game. For example, a scene of a racing game should cover also accidents which may lead to stronger degradations. When the aim of the study is to assess the interaction quality, e.g., the influence of delay on the control the player has over the game, a short interactive test is recommended. A stimulus duration of only 1.5 minutes might not be enough to investigate concepts such as flow and immersion. For such engagement concepts, a long interactive test is recommended. A more detailed description about the different test paradigms can be found in clause 9 regarding passive viewing-and-listening tests, and in clause 10 regarding interactive tests.

Table 1 – Overview of test paradigms

Aim of study / test paradigm	Short passive	Long passive	Short interact	Long interact
Output quality	x			
Interaction quality			x	
Engagement concepts				x

8 Experimental set up

Previous research has shown that a wide range of parameters influence the quality perception of a game. To repeatedly obtain comparable results, a very carefully designed test environment and procedures are essential.

8.1 Test environment

In general, the assessment methods of [ITU-T P.910] and [ITU-T P.911] should be regarded. However, under certain circumstances such as mobile gaming, it may be more appropriate to test in an environment which resembles typical playing conditions.

Classical multimedia lab tests are commonly carried out in somewhat "neutral" environments, such as sound-shielded rooms with daylight imitation. Whereas these environments create controlled conditions for each participant, it is obviously not representative for real-life gaming situations. In particular, in the case of mobile gaming on portable devices (e.g., smartphone, tablets) such an environment may generate misleading results with respect to the impact of device and display size on aspects of QoE. Experiments described in [b-Beyer] showed a significant impact of display size on several QoE dimensions, whereas the impact of the usage environment, neutral lab room vs. simulated metro environment, showed no significant influence. It was concluded that the pure physical simulation, consisting of background noise and space restrictions for the participants, might have missed a social impact of co-travelers in a real underground train.

A survey of 104 players performed by the University of Zagreb showed that participants can enjoy a provided game scenario (without degradations) regardless of the input and output equipment used (i.e., monitors and keyboard/mouse) which might differ from the equipment that participants are used to (under the condition that the equipment functions properly and that all performance parameters are acceptable). It is hereby concluded that the equipment used for subjective test does not underlie strict requirements, as long as the interaction with the game is not influenced. It was also reported that supervisors in a test room might influence the immersion (feeling of being watched, reduced concentration and autonomy).

8.1.1 Display specification

For video quality assessment, it was shown in [b-Winkler] that the perceived quality is strongly influenced by viewing distance, display size, brightness, contrast, sharpness, and colour. Not only the size of the display, but also the refresh rate of the display can bring higher quality if the frame rate is high. If high definition (HD) video games are targeted for the test, a full HD resolution display needs to be used.

The display size is an important aspect whenever control elements are placed directly on the screen. The usability of a game is influenced when a device is too small. As an outcome of the study in [b-Beyer] a screen size for mobile games should always be larger than five inches. However, the screen size should also not be too large or a negative impact on reaction times or less awareness over the whole scenario could result. For this reason, in the eSport domain a screen size of 24 inches is commonly used.

Depending on the display size used for subjective tests, the viewing distance, D, should be equal to three times the picture height, H (the video window size, not the physical display size), D = 3*H.

Regarding the monitor refresh rate, technology is evolving quickly. Additionally, new methods to synchronize graphics processing unit (GPU) rendering and monitor refresh rates are available on the market. When using such new technologies, there most likely will be a strong difference from older technologies, especially when coding parameters or very fast paced games are considered.

8.2 Test platform

A key issue in conducting gaming QoE studies is the design of the game platform itself. The main tasks attributed to digital game platforms include the following: gathering the player's input, calculating the virtual world state (based on the defined logic of the virtual world), rendering the virtual scene, and displaying the resulting real-time video to the user. Depending on the distribution of these tasks on the entities of the gaming platform, the following games can be differentiated:

– local games: all tasks are performed on a local computer or game console;

– online games: the calculation of the virtual world state is done on a remote server while other tasks (input, rendering of the virtual scene, and displaying) are performed locally. The player's inputs and information regarding the next virtual world state are transferred to and from a remote server, respectively;

– cloud games: the calculation and the rendering of the virtual scene are done remotely, while gathering the player's input and displaying are done locally. The player's inputs are transferred from the client to the remote server. Video or 3D graphical information is streamed from the remote server to the client. In this case, the server also performs encoding of the video stream, while the decoding is done locally.

For each of these scenarios, a standardized game platform should be defined when conducting QoE tests, due to significant differences in processing and network requirements.

Another important aspect to address is the differentiation of user devices, which can be grouped into: handheld consoles, mobile phones, tablets, consoles, and personal computers. As some elements of the technical platform (such as computational and graphical processing capacity of involved devices) undergo rapid technological development, a standardization of particular devices is not deemed worthwhile.

Instead, technical parameters must be selected and standardized which describe user-perceivable aspects of the end-to-end gaming platform including end-user devices. Parameters which may be of interest in QoE studies include (some parameters may apply only for some types of task distributions):

– input characteristics and devices (touchscreens, controllers, keyboard and mouse);

– (minimum) frame rate (usually an average number of frames per second);

– video codec parameters;

– network bandwidth;

– network delay;

– network packet loss;

– server processing delay;

– device delay (input device and display refresh rate);

– game type;

– game perspective / camera (e.g., first person linear perspective, third person linear perspective, and third person isometric perspective);

– spatial perceptual information (SI) and temporal perceptual information (TI) of the game's visual output [ITU-T P.910];

– effect of dynamically changing values for these parameters (e.g., jitter).

8.3 Participants

As games, by nature, are an interactive endeavor, the varying skill levels of participants needs to be encompassed by the experimental set up. While it might be desirable to have a mixed group of participants to obtain comparable results, it may also be germane to choose persons representing a service's target user group. However, in the latter case the test results are not comparable to those of a test conducted with a mixed group.

The taxonomy of [b-Möller] lists four user factors which potentially influence gaming QoE: experience, playing style, intrinsic motivation, other static or dynamic user factors such as age, gender, native language, current emotional status, boredom, distraction, or curiosity. User experience and skill are of considerable impact on the perception of a game in general and on impairments in particular. Experience can be divided into general game experience (i.e., how much time the test participant spends weekly playing digital games), experience related to game type (i.e., how much experience does the test participant have related to the type of the game under test) and experience playing the exact game under test. For example, a player who predominantly plays games of one genre may not be experienced and skilled at playing games from another genre, and should hence be considered a player with low experience for that particular genre. Additionally, players that have experience playing a particular game are in the best position to rate the game performance, as they know exactly how the virtual world should behave.

In both scientific and popular literature, a distinction between pro/advanced gamers and "newbies"/beginners is common. The deciding criterion is the time spent playing in a certain time frame. A person who spends more than three hours per week is considered an advanced player.

Prior to a test, participants should be screened for normal visual acuity or corrected-to-normal acuity and for normal colour vision. In case audible stimuli will be presented, appropriate pre-screening procedures like audiometric tests should be selected.

The impact of user experience on quality rating has been experimentally confirmed in the case of massively multiplayer online role-playing games (MMORPGs) [b-Suznjevic]. It might be possible to use objective data of specific games, e.g., ranks/points the gaming/streaming platform, to classify users based on their gaming experience. Whereas playing experience can be expected to significantly vary with age, playing style might be more related to a player's personality, which could be measured e.g., with the "Big 5" inventory, using standard screening questionnaires such as NEO-FFI [b-Costa] or alike. Player experience is commonly classified with regard to the number of weekly hours spent playing activities, and a typical threshold to differentiate between casual players and experts seems to be around 10 hours per week. Gender dynamics has been shown to affect playing experience of women in [b-Vermeulen]; it can be expected that this factor is moderated by playing style or personality.

Playing motivation is commonly steered in laboratory tests with the help of scenarios, by giving precise playing tasks to participants. In addition, self-regulation questionnaires for assessing motivation based on the self-determination theory might be helpful for this purpose, such as [b-Ryan]. In addition, frameworks for motivation assessment have been developed for certain game types such as MMORPG's [b-Yee].

So far, the emotional state has been considered as a user state which can influence the gaming experience. The emotional state may also be important as an input variable, i.e., as a human influence factor. In a natural gaming situation, a game will be initiated when a certain emotional state is present. Therefore, a valid result can only be obtained by considering the incoming emotional state of a player (i.e., emotional state at the time players arrive for the test session). On one hand, it can be evaluated by using subjective methods, e.g., such as the self-assessment manikin defined in clause 11.2; on the other hand by using physiological indices such as the alpha asynchrony index which is a neural indicator of liking/disliking [b-Arndt].

Workload could be assessed by questionnaires or physiological methods. As a subjective measure, an adapted version of the National Aeronautics and Space Administration task load index (NASA-TLX) can be used, but also other measures, such as performance indices can be employed. Boredom and/or fatigue could be assessed using electroencephalography (EEG)-derived physiological measures, such as activities on the alpha or theta bands [b-Antons]. Experiments clarifying the applicability of such measures in a gaming context are still missing.

8.4 Game material

The selection of game scenarios is an important issue. While video parameters such as the spatial and temporal perceptual information of the scenes are only of interest in some set-ups (e.g., cloud gaming), the actual game material is of importance in all test set-ups. When coding parameters are under investigation, it is recommended that game scenarios are chosen in a way that cover both complex and simple video game sequences. The complexity of video game sequences can be estimated by spatial and temporal indexes proposed in [ITU-T P.910].

Due to the variety of game types and players' individual preferences thereof, the quality perception of a game will be influenced by its content. For a quality assessment experiment, this means that the selection of games influences the obtained ratings and needs to be appropriate for the tested usage scenario.

To classify games, a genre classification is not accurate enough. A game of a certain genre can easily contain multiple interaction types such as shooting, moving, selecting or constructing. These interactions, in different ways, can be sensitive to degradations such as delay. From this viewpoint, what is evaluated is not a complete game but a specific scenario. A classification of games based on characteristics that gives more information about their different sensitivity towards degradations is necessary to overcome this issue. The game bricks, a classification based on game rules [b-Djaouti], as well as game characteristics described in [b-Aarseth], [b-Claypool] could turn out to be valuable in this respect.

For information concerning specific test paradigms, refer to clauses 9 and 10.

8.5 Experimental design

As denoted in [ITU-T P.911], different experimental designs, such as complete randomized design, Latin, Graeco-Latin and Youden square designs, replicated block designs, etc. (see [b-Kirk]) can be used, the selection of which should be driven by the purpose of the experiment.

However, the effect of repetitions of the same or comparable game scenarios on motivation, challenge, and skill has to be considered in the process of devising a test plan.

8.6 Questionnaires

Questionnaires are one of the most popular self-assessment methods available. Unlike methods such as thinking-aloud, they do not influence the gameplay directly while playing. They provide a standardized method for quantifying aspects of the player experience. However, the temporal resolution is limited since the information is often based on experiences and emotions from the past. For this reason, the number of items should be kept as low as possible. Generally, all kinds of comparative or repeated measurements should be avoided, since the inherent training effect for the test participant in itself may influence the obtained results. Since this is not always possible, randomizing the test conditions is mandatory. There is a variety of questionnaires available that are frequently used. However, the choice of the questionnaire depends strongly on the study objective. Typically, three kinds of questionnaires are used: a pre-test questionnaire asking for demographic information to classify the player, an in-game-questionnaire (also post-test) filled out at the end of each experimental condition aiming for the player experience, and a post-test questionnaire to summarize the test and get insights about choices the player made or similar. In-game questionnaires focus either on a variety of different dimensions to cover the whole player experience or focus on single dimensions such as immersion or flow. Clause 11 gives an overview of available questionnaires and describes their use case.

8.6.1 Pre-test questionnaire

The pre-test questionnaire may contain the following information:

– Demographic information
  - age, gender, profession;
  - gaming experience (average hours a week, typical session length, sessions a week, number of played games, time started playing);
  - game preferences (current games played, favorite genre / type, hardware usage);
  - knowledge about games used in the study;
  - experiences with degradation.

– Gaming related personality
  - the 5-domains-of-play [b-VandenBerghe];
  - gamification user types hexad scale [b-Tondello];
  - immersive tendencies questionnaire (ITQ) [b-Singer].

8.6.2 Overall quality

To assess the overall quality of a gaming experience, the absolute category rating (ACR) method can be used. The ACR method is a category judgment, where test sequences are presented one at a time are rated independently on a category scale (see [ITU-T P.911]). However, the stimulus duration and wording of the item should be adapted to gaming in comparison with [ITU-T P.911]. It is important to assess the gaming experience and not the quality of service or video quality. As shown in the taxonomy in clause 6, gaming QoE is a multidimensional construct which includes many additional aspects compared to the QoE of video (streaming) services. While discrete category rating scales are very common, it is recommended to use the following item and 7-point continuous scale [ITU-T P.851] to assess the overall quality of a gaming experience:

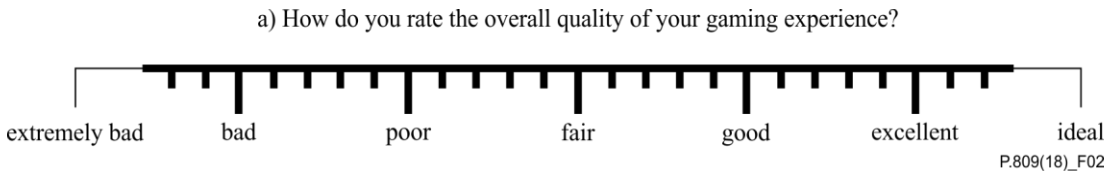

Figure 2 – Item and scale for overall gaming experience

9 Passive viewing-and-listening tests with audiovisual stimuli

A passive viewing-and-listening test offers two major advantages: it is possible to ensure that every participant is rating exactly the same content in an experiment that is easy to conduct, and the playing abilities of the participants will not influence the outcome of the study. However, since gaming is an interactive service, not every aspect can be assessed in a passive viewing-and-listening test. The focus of a study hereby should be on the output quality and not on the interaction quality or input quality. Since, to the best of our current knowledge, interactivity is necessary for a feeling of flow, this concept cannot be assessed in a passive test. The same applies for judgements of playing quality (game usability) or for the influence of parameters such as delay.

9.1 Instructions for participants

For each subjective test, information about the test design as well as the rules and goals of the game should be given to the participants. Although game rules and goals are not directly relevant in a passive viewing-and-listening test paradigm, this ensures that the participants have similar knowledge of the game, and avoids the participants to ask questions about it. The assessment methods (e.g., questionnaires) should be explained before starting the test. Especially for the assessment of video quality it must be ensured, that participant do not rate the graphic quality (graphical details, abstract or realistic graphics) but rather video compression artifacts such as blockiness and blurriness. Even in a passive viewing-and-listening test it is possible to assess the impact of video degradations on the player experience to some extent. Participants in that case must be instructed to put themselves in the position of a player who played the game scenario under the giving conditions.

The instructions could be a printed version or an introductory video and should include the following information:

– information about the goal of the game and basic description of objects in the game such as enemies, and obstacles;

– information about experimental details, such as session duration or questionnaires;

– there are no "correct" ratings. The instructions should not suggest that there is a correct rating or provide any feedback as to the "correctness" of any response. The instructions should emphasize that the test is being conducted to learn subjects' judgments of the quality of the samples, and that it is the subject's opinion that determines the appropriate rating.

9.2 Test stimuli

The duration of a stimulus (video scene of game play) partially depends on the game content. The scenario should be chosen in a way that is representative for the game. It should cover different game mechanics that would be visible in an interactive scenario. A duration of 10 seconds used for traditional video quality test [ITU-T P.910], might not be enough to reach this requirement. Based on [b-Schmidt], [b-Claypool-1], where significant differences between video quality ratings for different stimulus durations was shown, even though the video complexity of the selected scenes and the content itself was very similar, it is recommendable in a passive viewing-and-listening test to use a stimulus duration of 30 seconds. In such a short duration, it is unlikely to be able to measure concepts such as flow and immersion, but it is possible to assess the output quality under many conditions.

When aiming for complex constructs such as player engagement (e.g., flow or immersion) a stimulus duration of only 30 seconds is not sufficient. For this case, so far there is no recommendation for an ideal stimuli length available but it seems reasonable to use a duration of 10-15 minutes. It should be noted, that participants in a passive viewing-and-listening test will suffer from fatigue faster than in an interactive test.

9.3 Game material

In general, there are two methods available to create audiovisual stimuli for passive viewing-and-listening tests: encoding a reference video using different encoding parameters or recording actual game play using different streaming parameters. However, encoding videos might confuse test participants who would judge the content as unrealistic since they would expect a change in the playing behavior in the case of strong degradations, such as a very-low frame rate. If the aim of a study is to have a link between a passive viewing-and-listening and the interactive test, then using recorded scenes from the interactive test should be considered. If this is not the case, it is recommended to convert the videos, since recording actual game play will always lead to changes in the content, resulting in unpredictable influences on the player experience.

10 Interactive tests with game scenes

10.1 Test stimuli

The duration of a stimulus (interactive game play), in an interactive test highly depends on the aim of the subjective experiment. Two approaches are conceivable in this respect:

Short interactive: In a short interactive test, in which a typical stimulus (interactive game play) length is between 90-120 seconds, it is possible to assess the interaction quality (e.g., the impact of delay on the control), but the assessment of more complex player experience features, highly depends on the player and the game content. Games that meet the interest of the player which are intuitive without being bored, can already immerse players within a short time of a few minutes.

Long interactive: When aiming for all QoE aspects mentioned in clause 6, so far there is no recommendation for an ideal stimuli length available, but it is reasonable to use a duration of 10-15 minutes to ensure that players get emotionally attached to a game scenario while aiming to measure emotions and other QoE aspects such as flow.

"The preliminary results of research conducted by Stanney indicate that presence is not enhanced by the prolonged exposure. This study indicated that the virtual environment itself seems to promote a high level of presence (or not) within the very first 15 minutes of exposure" [b-Immersion]. Beyer et al. could show a similar effect. The physiological EEG data comparing the influence of quality variations showed a significant effect for the wakefulness state: Playing in a low-quality condition caused significantly higher spectral power in the alpha frequency band during the first half (10 minutes) of that session compared to the high-quality condition. "While this effect was also observable in the second half of the sessions, it was less pronounced and did not reach significance level. This might imply that the longer a player played the game the less influence is exerted on the wakefulness state by the video quality. As a game is an interactive endeavor as opposed to mere passive video consumption, the player may over time adapt to the degraded visual quality, and the game's interactive content might dominate the perception" [b-Beyer-1].

Concerning the total duration of the subjective experiment, a QoE study evaluating the impact of frame rate and video codec bitrate on the QoE of cloud games showed that there is no significant variation of QoE based on prolonged exposure to the gaming stimuli [b-Brühlmann]. Scores of particular scenarios (i.e., combination of fixed system parameter values, bitrate and framerate) were compared for tests performed at the start of experiment and on the end of experiment (which lasted three hours) and no significant differences between distribution of scores for any tested scenario was determined.

Irritating auditory notifications as well as turning off the display instantaneously should be avoided. These methods can annoy or startle the participants of a study. Fading over two seconds to a gray screen (transparent transition) as well as turning down the sound volume accordingly has proven to be a more comfortable method to end a condition [b-Schmidt-1]. In case of a game with a fixed ending, e.g., the score menu in a soccer game, this might not be necessary.

10.2 Instructions for participants

For each subjective test, information about the test design as well as the rules and controls of the game should be given to the participants. This ensures that the participants have similar knowledge of the game, and it reduces learning effects. The assessment methods (e.g., questionnaires) should be explained before starting the test. Especially for the assessment of video quality it must be ensured, that participant do not rate the graphic quality (graphical details, abstract or realistic graphics) but rather artefacts such as blockiness and blurriness. The instruction could be a printed version or an introductory video. It should include the following information:

– information about how to control the game via input device such as main control button;

– information about the goal of the game and basic description of objects in the game such as enemies, and obstacles;
– information about experimental details, such as session duration or questionnaire;
– there are no "correct" ratings. The instructions should not suggest that there is a correct rating or provide any feedback as to the "correctness" of any response. The instructions should emphasize that the test is being conducted to learn subjects' judgments of the quality of the samples, and that it is the subject's opinion that determines the appropriate rating.

For all games a training session should be used. The training session may use a (non-degraded) reference condition, and it may additionally show degraded conditions, in order to anchor the use of the rating scale(s).

10.3 Game material

The interactivity of the game content was shown to influence game experience, e.g., in conjunction with network bandwidth or delay [b-Möller-1]. Hence, a classifier of the degree of interactivity for each utilized game is necessary and should be reported along the results of the QoE evaluation. In turn, it was also shown that the game type may have an impact on the required bitrate, and in this way also on gaming QoE [b-Suznjevic-1]. Such indirect relationships must be investigated further to come up with a meaningful categorization of games.

The chosen scenes from the game must be representative for the game 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.

Claypool states that the spatial and temporal accuracy in that a player must perform actions to successfully progress in the game, as well as the perspective of the game, are crucial when investigating the impact of delay. Other attributes such as the number of actions per minute, the number of surprising events or whether a game is turn-based or not should be considered as well [b-Sackl]. Not only should those aspects be considered for choosing the game scenarios, but it is important that they are also reported whenever publishing results of studies in order to later on compare the results. Finally, they can be used to explain outcomes of research projects.

For a subjective study the session duration should be representative of the selected game, and it strongly depends on the actual game. Since the total time of an experiment is limited and often many conditions are investigated, the use of long, round-based games should be avoided, unless these games are of specific interest.

The use of horror games or overly violate game should be avoided if possible, not only for ethical reasons, but especially when physiological measurements are used to capture the user's state.

10.4 Test system set-up

The test system should be able to execute the game without any unwanted system-introduced impairments.

10.5 Socializing aspect

Many players like to socialize while playing. Playing together with or competing against other human players is a driving motivation for them. The influences of this co-experience are yet not well investigated. Vermeulen et al. were able to demonstrate, that female players perform and rate differently when they assume to play against a male opponent [b-Vermeulen]. For a study investigating the impact of network parameters or alike it is recommended to disregard social aspects in the assessment unless this is the aim of the study. A set-up in that multiple users play against or with each other is very complex and difficult to analyze, since even more factors will influence quality ratings.

11 Questionnaires

11.1 Game experience questionnaire

The game experience questionnaire (GEQ) "comprehensively and reliably characterizes the multifaceted experience of playing digital games [b-Poels]. The GEQ, which was developed, implemented and validated during the "Fun of Gaming" (FUGA) project, has so far been one of the most popular tools to assess game experience. More information about the GEQ can be found in [b-IJsselsteijn].

The GEQ has a modular structure and consists of:

1) the Core module (concerning the actual player experience during a scenario);

2) the Social Presence module (concerning involvement with other social entities);

3) the Post-game module (concerning experiences once stopped playing).

The core part of the GEQ is a 33-item questionnaire using 5-point ACR scales designed to assess seven dimensions of player experience: sensory and imaginative immersion, tension, competence, flow, negative effect, positive effect, and challenge [b-IJsselsteijn]. More information about those dimensions are available in clause 6. By using these many dimensions, the GEQ is a tool to assess the player experience in a very detailed manner. However, it does not offer information available about the cause of those experiences that might be the result of an influenced interaction quality. In addition to this module, a concise in-game version is available (iGEQ). Here only two items are used for each dimension. To avoid confusion, the items of the in-game version for the immersion dimension should be adjusted in case of a game without any story.

The GEQ was used in a variety of studies and was proven to be a valuable tool. However, some concerns have recently been raised regarding the reliability of the GEQ core module [b-DeGrove] and about the inadequate length of the questionnaire when using within-subjects experimental designs. Norman states, that the GEQ seems reasonable and applicable in studying player experiences with video games. However, 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 noncompetitive games (e.g., simulations) [b-Norman].

11.2 Self-assessment manikin

The self-assessment manikin (SAM) scales developed by Bradley & Lang in the 1980's are a pictorial rating system to obtain self-assessments of experienced emotions on three dimensions: *pleasure* (*affective valence)*, *dominance* and *arousal*. The questionnaire consists of three scales depicting a horizontal array of sketched 'manikins' showing visible emotional signs related to the respective dimensions. The first of these scales, measuring the dimension *pleasure*, is related to attributes like happiness, satisfaction, and relaxation. The second dimension, *arousal*, refers to aspects such as stimulation, excitement, or feeling wide awake. It describes the perceived vigilance as a physiological and psychological condition of a person. The range reaches from excitation to doziness or boredom. *Dominance*, the last dimension, concerns feeling in control vs. being controlled, or feeling influential vs. being influenced. This describes how much a person feels in control of a situation. The respective scale can be seen in Figure 3, [b-Irtel].

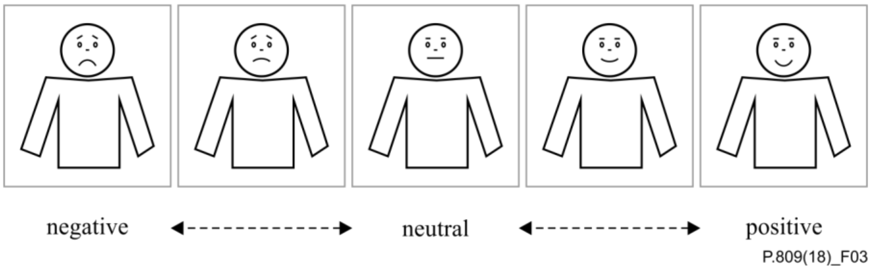

Figure 3 – 5-step SAM scale for pleasure

11.3 Diagnostic degradation characterization

Unlike for audiovisual quality assessments, degradation category rating and pair comparison methods are not appropriate for the assessment of gaming quality due to their inherent repetition of game sequences. Custom scales can be employed to assess the subjective impact of system degradations such as delay. One example is given in Figure 4.

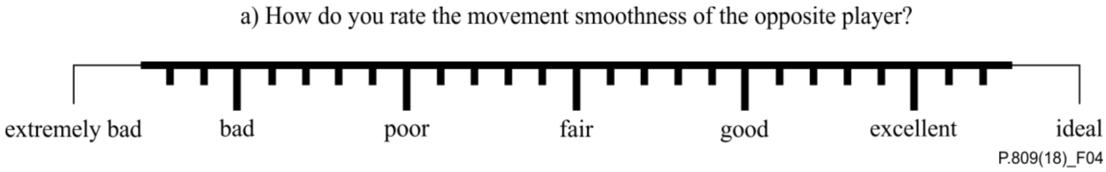

Figure 4 – Exemplary scale to assess the smoothness of an opponent's movements within a game

11.4 Flow questionnaires

11.4.1 Flow-short-scale

The flow-short-scale (was published in 2003 by Rheinberg, Vollmeyer and Engeser as flow-kurzskala (FKS, German version). The first 10 items of the scale measure the components of *flow-experience* as first described by Csikszentmihalyi [b-Csik]. The scale collapses characteristics of flow into two dimensions labeled as *fluency of performance* (e.g., concentration and focus, control, clarity) and *absorption by activity* (e.g., involvement, distorted sense of time, optimal challenge, absent-mindedness). Three additional items can be used to measure *worry* someone may have in the situation the measurement is made. The flow-short-scale uses 7-point ACR scales, ranging from 1 (not at all) to 7 (very much), to assess a flow experience immediately after or while conducting the according activity. The flow-short-scale has been translated into several languages: English, French, Italian, Danish, Czech, Turkish, Dutch [b-Rheinberg]. Even though the scale was not developed specifically for the gaming domain, it was used in numerous gaming studies [b-Weibel], [b-Engeser].

11.4.2 Flow state scale

The flow state scale developed by Jackson and Marsh assesses participants' level of flow experience [b-Jackson]. The questionnaire consists of 36-items on 9 subscales. A 5-point Likert-type scale ranging from 1 (strongly disagree) to 5 (strongly agree) is used. The dimensions of flow include [b-Barry]:

Challenge-skill balance: skills match the task and will be successful

Action-awareness merging: automatic response to task

Clear goals: experience of having a pre-set goal to achieve

Unambiguous feedback: feedback on performance

Concentration on task: focused on task

Paradox of control: performs task with ease

Loss of self-consciousness: immersed in task

Transformation of time: time speeds up or slows down during activity

Autotelic experience: activity intrinsically rewarding

Jackson and Eklund improved this method and created the flow state scale-2, which assesses flow at two levels:

1) dispositional level (DFS): Frequency of flow experience in particular domains (e.g., sport, work, school) - these are known as the dispositional versions of the flow scales;

2) state level (FSS-2): Extent of flow experienced in a particular event or activity (e.g., a race, a work project, or a test) - these are known as the state versions of the flow scales [b-FlowScales].

11.5 Engagement questionnaires

11.5.1 Game engagement questionnaire

The game engagement questionnaire (GEngQ) was initially developed to determine the impact of playing violent video games by measuring tendency to become engaged in video game-playing. The questionnaire consists of 19 questions answered on a 7-point Likert scale (strongly disagree to strongly agree). The questionnaire measures engagement by including the four dimensions: immersion, presence, flow and psychological absorption [b-Brockmyer].

11.5.2 Immersive experience questionnaire

The immersive experience questionnaire (IEQ) uses a mixture of questions combining aspects of flow, cognitive absorption and presence. Five factors underlying immersion were identified: person factors (*cognitive involvement*, *real world dissociation* and *emotional involvement*) and game factors (*challenge* and *control*). Sixteen pairs of related questions were created using negative and positive wording in order to control for wording effects. Answers for each question are marked on 7-point Likert scales. A final question was also included that asked how immersed the participant felt overall on a scale of 1 to 10, where 1 was 'not at all' and 10 was 'very much so'. Therefore, there were thirty-three questions overall. As a result of a study it appears that the immersion questionnaire developed was a successful indicator of immersion, and people can reliably reflect on their own immersion in a single question [b-Jennett-1]. The IEQ has been tested empirically across a far-reaching array of different scenarios and game types [b-Nordin].

11.6 Igroup presence questionnaire

The igroup presence questionnaire (IPQ) is a scale for measuring the sense of presence experienced in a virtual environment (VE). It has been constructed using a large pool of items and two survey waves with approximately 500 participants. It was originally constructed in German, but is now also available in English, Dutch, French and Japanese. The current version of the IPQ has three subscales and one additional general item not belonging to a subscale. The three subscales emerged from principal component analyses and can be regarded as fairly independent factors [b-igroup]:

1) spatial presence – the sense of being physically present in the VE;

2) involvement – measuring the attention devoted to the VE and the involvement experienced;

3) experienced realism – measuring the subjective experience of realism in the VE.

The questionnaire consists of 14 items. Answers for each question are marked on 7-point Likert scales. The authors understand presence as the subjective sense of being in a virtual environment and argue further, that it should be separated from the ability of a technology to immerse a user. They state that immersion is a variable of the technology and can be described objectively, presence is a variable of a user's experience [b-igroup]. Hartmann et al. [b-Hartmann] noted that the IPQ scales yielded acceptable internal consistency and the results of experimental tests support the validity of the IPQ, but only one of the three IPQ subscales actually measures spatial presence, whereas *involvement* and *realness* may address closely related constructs or determinants rather than actual sub-dimensions of spatial presence.

11.7 Presence questionnaire

The Presence questionnaire (PQ) was developed by Witmer and Singer in 1998 to measure spatial presence in immersive virtual environments [b-Singer]. In addition to assess presence itself it also includes contributing factors. Witmer and Singer identified involvement and immersion as conditions for presence [b-Singer]. Presence requires the ability to focus on a meaningfully coherent set of stimuli in the VE to the exclusion of unrelated stimuli in the physical location. This increase of attention is also referred to the term involvement [b-Singer]. In 2005, the authors improved the questionnaire by performing a principal-components analysis of the PQ data from 325 participants following exposure to immersive virtual environments. The analyses suggest that a 4-factor model provides the best fit to our data. The factors are *involvement, adaptation/immersion, sensory fidelity,* and *interface quality*. Within these relationships, sensory fidelity items (previously named realism factors) was reported to be more closely related to involvement, whereas interface quality items (previously named control factors) appear to be more closely related to adaptation/immersion, even though there is a moderately strong relationship between the involvement and adaptation/immersion factors [b-Witmer]. The initial version of the questionnaire containing 32 items was reduced to 29 items (by removing 26, 27, and 28). As scales 7-points Likert-scales are used.

11.8 Player experience of need satisfaction

Another questionnaire frequently used to quantify the experience of playing digital games is the player experience of need satisfaction (PENS). The PENS survey is designed to explain the game play factors that lead to enjoyable and meaningful player experiences. The questionnaire contains 21 items, where it reviews the experience in terms of 5 components, such as *competence, autonomy, presence, relatedness,* and *intuitive controls*. All but one are measured using 3-item scales (apart from presence, which is a 9-item scale), ranked on a 7-point Likert scale. It has been statistically validated, however the questionnaire is copyrighted [b-Denisova].

11.9 Conclusions on questionnaires

All of the above questionnaires were used in a variety of studies. However, many of them measure overlapping constructs and are not suited for every use case. Little research is available comparing the results of these questionnaires [b-Norman], [b-Denisova], [b-Brühlmann].

More work is necessary to provide a list of questionnaires that is validated for concrete use cases. This list tried to summarize a variety of frequently used questionnaires, but does not claim to be complete. Depending on the aim of the study, tradition questionnaires such as the positive and negative affect schedule (PANAS) or the user experience questionnaire (UEQ) can also be used.

12 Player performance measurement

Since a link between the outcome of a game and its perceived quality exists, the QoE can, in a very limited way, be assessed using objective performance metrics. The selection of these depends strongly on the type of game and the objective of the experiment.

Frequently, scores (e.g., points, number of kills, goals, units built/destroyed) are used but also other parameters such as game session length or frequency may be employed.

Since they are non-intrusive in nature, performance metrics are well suited to monitor games in field tests without interruptions.

Although performance metrics are a seemingly easy way to assess quality, they offer only limited insight into a player's subjective quality perception.

Bibliography

[b-ITU-T G.1032] Recommendation ITU-T G.1032 (2017), *Influence factors on gaming quality of experience*.

[b-Aarseth] Aarseth, E., Smedstad, S.M., and Sunnan, L. (2003), *A MULTI-DIMENSIONAL TYPOLOGY OF GAMES*, in Proceeding from conference Digra.

[b-Agarwal] Agarwal, R., and Karahana, E. (2000), *Time flies when you're having fun: Cognitive Absorption and beliefs about information technology usage,* MIS Quarterly, Vol. 24 (4), pp. 665-694.

[b-Antons] Antons, J.N., Schleicher, R., Arndt, S., Möller, S., and Curio, G. (2012), *Too tired for calling? A physiological measure of fatigue caused by bandwidth limitations*, in Quality of Multimedia Experience (QoMEX), Fourth International Workshop, pp. 63-67.

[b-Arndt] Arndt, S., *et al.* (2013), *The effects of text-to-speech system quality on emotional states and frontal alpha band power*, in Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference, pp. 489-492.

[b-Barry] Barry, G., Van Schaik, P., MacSween, A., Dixon, J., and Martin, D. (2016), *Exergaming (XBOX Kinect^{TM}) 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, No. 1, p. 25.

[b-Bartle] Bartle, R. (1996), *Hearts, clubs, diamonds, spades: Players who suit MUDs*, J. MUD Res., Vol. 1, No. 1, p. 19.

[b-Beyer] Beyer, J., and Möller, S. (2014), *Assessing the Impact of Game Type, Display Size and Network Delay on Mobile Gaming QoE*, PIK-Prax. Informationsverarbeitung Kommun., Vol. 37, No. 4, pp. 287-295.

[b-Beyer-1] Beyer, J., Varbelow, R., J Antons, J.N., and Möller, S. (2015), *Using electroencephalography and subjective self-assessment to measure the influence of quality variations in cloud gaming*, in Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop, pp. 1-6.

[b-Bjork] Bjork, S., and Holopainen, J. (2004), *Patterns in game design* (game development series).

[b-Brockmyer] Brockmyer, J.H., Fox, C.M., Curtiss, K.A., McBroom, E., Burkhart, K.M., and Pidruzny, J.N. (2009), *The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing*, J. Exp. Soc. Psychol., Vol. 45, No. 4, pp. 624-634.

[b-Brown] Brown, E., and Cairns, P. (2004), *A grounded investigation of game immersion*, in CHI'04 extended abstracts on Human factors in computing systems, pp. 1297-1300.

[b-Brühlmann] Brühlmann, F., and Schmid, G.M. (2015), *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, pp. 1181-1186.

[b-Chen] Chen, J. (2007), *Flow in games (and everything else)*, Commun. ACM, Vol. 50, No. 4, pp. 31-34.

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

[b-Claypool-1] Claypool, M. (2009), Motion and Scene Complexity for Streaming Video Games, in *Proceedings of the 4th International Conference on Foundations of Digital Games*, New York, NY, USA, pp. 34-41.

[b-Costa] Costa, P.T., and MacCrae, R.R. (1992), *Revised NEO personality inventory (NEO PI-R) and NEO five-factor inventory (NEO FFI): Professional manual*, Psychological Assessment Resources.

[b-Csik] Csikszentmihalyi, M. (1987), *Das flow-Erlebnis: jenseits von Angst und Langeweile*, im Tun aufgehen. Klett-Cotta.

[b-DeGrove] De Grove, F., Van Looy, J., and Courtois, C. (2010), *Towards a Serious Game Experience Model: Validation, Extension and Adaptation of the GEQ for Use in an Educational Context*, in Playability and Player Experience, Breda University of Applied Sciences, Vol. 10, pp. 47-61.

[b-Denisova] Denisova, A., Nordin, A.I., and Cairns, P. (2016), *The Convergence of Player Experience Questionnaires*, in Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play, pp. 33-37.

[b-Djaouti] Djaouti, D., Alvarez, J., Jessel, J.P., Methel, G., and Molinier, P. (2008), *A Gameplay Definition Through Videogame Classification*, Int. J. Comput. Games Technol., Vol. 2008, pp. 4:1-4:7.

[b-Engeser] Engeser S., and Rheinberg, F. (2008), *Flow, performance and moderators of challenge-skill balance*, Motiv. Emot., Vol. 32, No. 3, pp. 158-172.

[b-FlowScales] The Flow Scales and Positive Psychology (2016), *Bodyandmindflow.com.au*. [Online]. Available: <http://www.bodyandmindflow.com.au/the-flow-scales/>. [Accessed: 13-Dec-2016].

[b-Hartmann] Hartmann T., *etal.*, (2015), *The Spatial Presence Experience Scale (SPES)*, Journal of Media Psychology, Vol. 28, No. 1, pp. 1-15.

[b-Hassenzahl] Hassenzahl, M., Platz, A., Burmester, M., and Lehner, K. (2000), *Hedonic and ergonomic quality aspects determine a software's appeal*, in Proceedings of the SIGCHI conference on Human Factors in Computing Systems, pp. 201-208.

[b-Hassenzahl-1] Hassenzahl, M. (2008), *User experience (UX): towards an experiential perspective on product quality*, in Proceedings of the 20th Conference on l'Interaction Homme-Machine, pp. 11-15.

[b-Hugentobler] Hugentobler, U. (2011), *Messen von Flow mit EEG in Computerspielen*, PhD thesis, University of Zürich.

[b-igroup] Igroup Presence Questionnaire (IPQ) overview (2016), *igroup.org – project consortium*. [Online]. Available: http://www.igroup.org/pq/ipq/index.php. [Accessed: 13-Dec-2016].

[b-IJsselsteijn] IJsselsteijn, W.A., De Kort, Y. A. W., and Poels, K. (2013), *The Game Experience Questionnaire: Development of a self-report measure to assess the psychological impact of digital games*, Manuscript in Preparation.

[b-Immersion] Definition for Immersion and sens of presence (2016), <http://w3.uqo.ca/cyberpsy/en/pres_en.htm - def> [Accessed: 13-Dec-2016].

[b-Irtel] Irtel, H. (2007), *PXLab: The psychological experiments laboratory*, Univ. Mannh. Mannh.

[b-Jackson] Jackson, S.A., Marsh, H.W., etal. (1996), *Development and validation of a scale to measure optimal experience: The Flow State Scale*, Journal of Sport and Exercise Psychology, Vol. 18, pp. 17-35.

[b-Jennett] Jennett, C.I. (2010), *Is game immersion just another form of selective attention? An empirical investigation of real world dissociation in computer game immersion*, UCL (University College London).

[b-Jennett-1] Jennett, C., et al. (2008), *Measuring and defining the experience of immersion in games*, Int. J. Human-Computer Studies., Vol. 66, No. 9, pp. 641-661.

[b-Juul] Juul, J. (2005), *Video games between real rules and fictional worlds*, Video Games Real Rules Fict. Worlds.

[b-Kirk] Kirk, R.E. (1969), *Experimental Design: Procedures for the Behavioural Sciences*, BrooksCole Belmont.

[b-Lazzaro] Lazzaro, N., and Keeker, K. (2004), *What's my method?: a game show on games*, in CHI'04 Extended Abstracts on Human Factors in Computing Systems, pp. 1093-1094.

[b-Möller] Möller, S., Schmidt, S., and Beyer, J. (2013), *Gaming taxonomy: An overview of concepts and evaluation methods for computer gaming QOE*, in Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop, pp. 236-241.

[b-Möller-1] Möller, S., Pommer, D., Beyer, J., and Rake-Revelant J. (2013), *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), pp. 1-5.

[b-Murphy] Murphy, C. (2011), *Why games work and the science of learning*, in Interservice, Interagency Training, Simulations, and Education Conference, pp. 260-272.

[b-Nordin] Nordin, A.I., Denisova, A., and Cairns, P. (2014), *Too Many Questionnaires: Measuring Player Experience Whilst Playing Digital Games*, in Seventh York Doctoral Symposium on Computer Science and Electronics, Vol. 69.

[b-Norman] Norman, K.L. (2013), *Geq (game engagement/experience questionnaire): a review of two papers*, Interact. Comput., Vol. 25, No. 4, pp. 278-283.

[b-Pinelle] Pinelle, D., Wong, N., and Stach, T. (2008), *Heuristic evaluation for games: usability principles for video game design*, in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2008, pp. 1453-1462.

[b-Poels] Poels, K., De Kort, Y. A. W., and Ijsselsteijn, W.A. (2008), *FUGA-The fun of gaming: Measuring the human experience of media enjoyment. Deliverable D3. 3: Game Experience Questionnaire*, FUGA Proj.

[b-Rheinberg] Rheinberg, F. (2015), *Die Flow-Kurzskala (FKS) übersetzt in verschiedene Sprachen The Flow-Short-Scale (FSS)*, translated into various languages.

[b-Ryan] Ryan, R.M., and Connell, J.P. (1989), *Perceived locus of causality and internalization: examining reasons for acting in two domains*, Journal of Personality and Social Psychology, Vol. 57, No. 5, p. 749.

[b-Sackl] Sackl, A., Schatz, R., Hossfeld, T., Metzger, F., Lister, D., and Irmer, R. (2016), *QoE Management made uneasy: The case of Cloud Gaming*, in Communications Workshops (ICC), 2016 IEEE International Conference, pp. 492-497.

[b-Schaffer] Schaffer, N. (2009), *Verifying an integrated model of usability in games*, Rensselaer Polytechnic Institute.

[b-Schmidt] Schmidt, S., Zadtootaghaj, S., Möller, S. (2017), *A Comparison of Interactive and Passive Quality Assessment for Gaming Research*, Quality of Multimedia Experience (QoMEX), Ninth International Conference on IEEE.

[b-Schmidt-1] Schmidt, S., Zadtootaghaj, S., and Möller, S. (2017), *Towards the delay sensitivity of games: There is more than genres*, in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), pp. 1-6.

[b-Singer] Witmer, B.G., and Singer, M.J. (1998), *Measuring presence in virtual environments: A presence questionnaire*, Presence Teleoperators Virtual Environ., Vol. 7, No. 3, pp. 225-240.

[b-Slater] Slater, M., and Wilbur, S. (1997), *A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments*, Presence Teleoperators Virtual Environ., Vol. 6, No. 6, pp. 603-616.

[b-Stelmaszewska] Stelmaszewska, H., Fields, B., and Blandford, A. (2004), *Conceptualising user hedonic experience*, in Proc. ECCE, Vol. 12, pp. 12-15.

[b-Suznjevic] Suznjevic, M., Skorin-Kapov, L., and Matijasevic, M. (2013), *The impact of user, system, and context factors on gaming qoe: a case study involving mmorpgs*, in Proceedings of Annual Workshop on Network and Systems Support for Games, pp. 1-6.

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

[b-Tondello] Tondello, G.F., Wehbe, R.R., Diamond, L., Busch, M., Marczewski, A., and Nacke, L.E. (2016), *The Gamification User Types Hexad Scale*, in Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play, New York, NY, USA, pp. 229-243.

[b-VandenBerghe] VandenBerghe, J. (2016), *GDC Vault - The 5 Domains of Play: Applying Psychology's Big 5 Motivation Domains to Games*. http://www.gdcvault.com/play/1015364/The-5-Domains-of-Play. [Accessed: 13-Dec-2016].

[b-Vermeulen] Vermeulen, L., Núñez-Castellar, E., and Van Looy, J. (2014), *Challenging the other: Exploring the role of opponent gender in digital game competition for female players*, Cyberpsychology Behav. Soc. Network., Vol. 17, No. 5, pp. 303-309.

[b-Vilnai] Vilnai-Yavetz, I., Rafaeli, A., and Yaacov, C.S. (2005), Instrumentality, aesthetics, and symbolism of office design, *Environment and Behavior*, vol. 37, no. 4, pp. 533-551.

[b-Webster] Webster, J., and Ho, H. (1997), *Audience Engagement in Multi-Media Presentations*, Data Base for the Advances in Information Systems.

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

[b-Weniger] Weniger, S., and Loebbecke, C. (2011), *Cognitive absorption: Literature review and suitability in the context of hedonic IS usage*, Department of Business, Media and Technology Management, University of Cologne, Germany.

[b-Winkler] Winkler, S. (1999), Issues in vision modeling for perceptual video quality assessment, Signal Process, Vol. 78, No. 2, pp. 231-252.

[b-Witmer] Witmer, B.G., Jerome, C.J., and Singer, M.J. (2005), *The factor structure of the presence questionnaire*, Presence, Vol. 14, No. 3, pp. 298-312.

[b-Yee] Yee, N. (2006), *Motivations for play in online games*, Cyberpsychology Behavior, Vol. 9, No. 6, pp. 772-775.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series D	Tariff and accounting principles and international telecommunication/ICT economic and policy issues
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	Cable networks and transmission of television, sound programme and other multimedia signals
Series K	Protection against interference
Series L	Environment and ICTs, climate change, e-waste, energy efficiency; construction, installation and protection of cables and other elements of outside plant
Series M	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	**Telephone transmission quality, telephone installations, local line networks**
Series Q	Switching and signalling, and associated measurements and tests
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of Things and smart cities
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2018